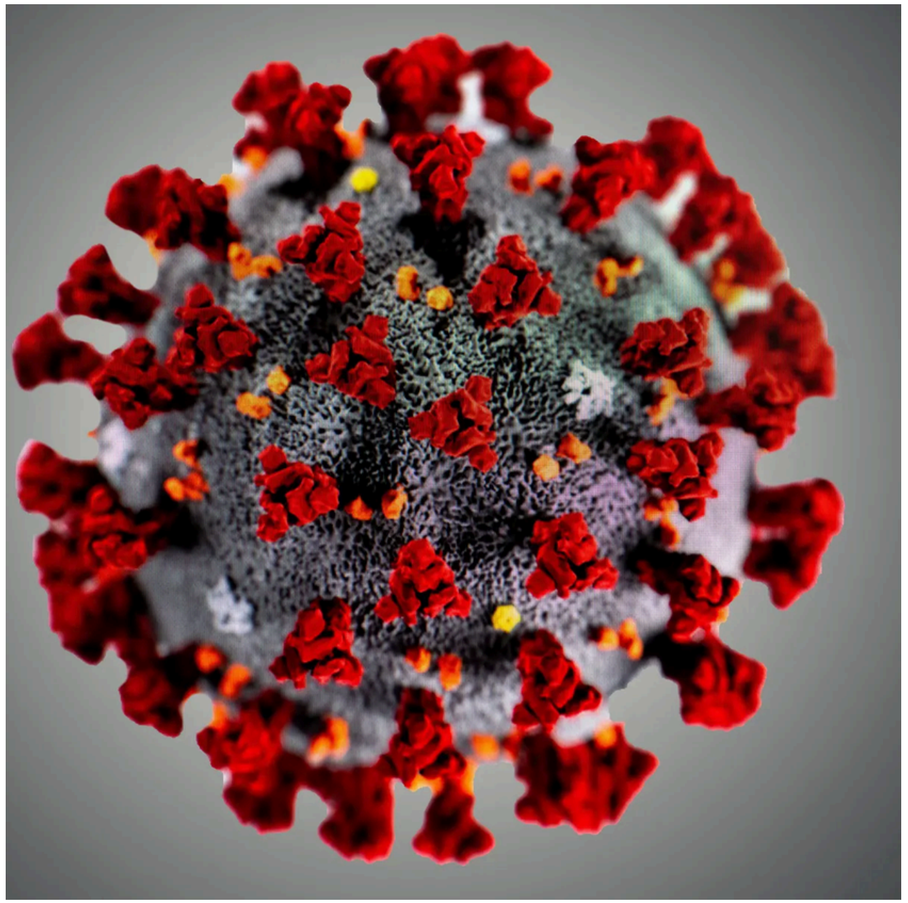
Misinformation Detection: Progress, Pitfalls, and the Path to Societal Impact

Tommaso Caselli - GroNLP
t.caselli@rug.nl

6th ROMCIR WORKSHOP @ ECIR
April 02, 2026

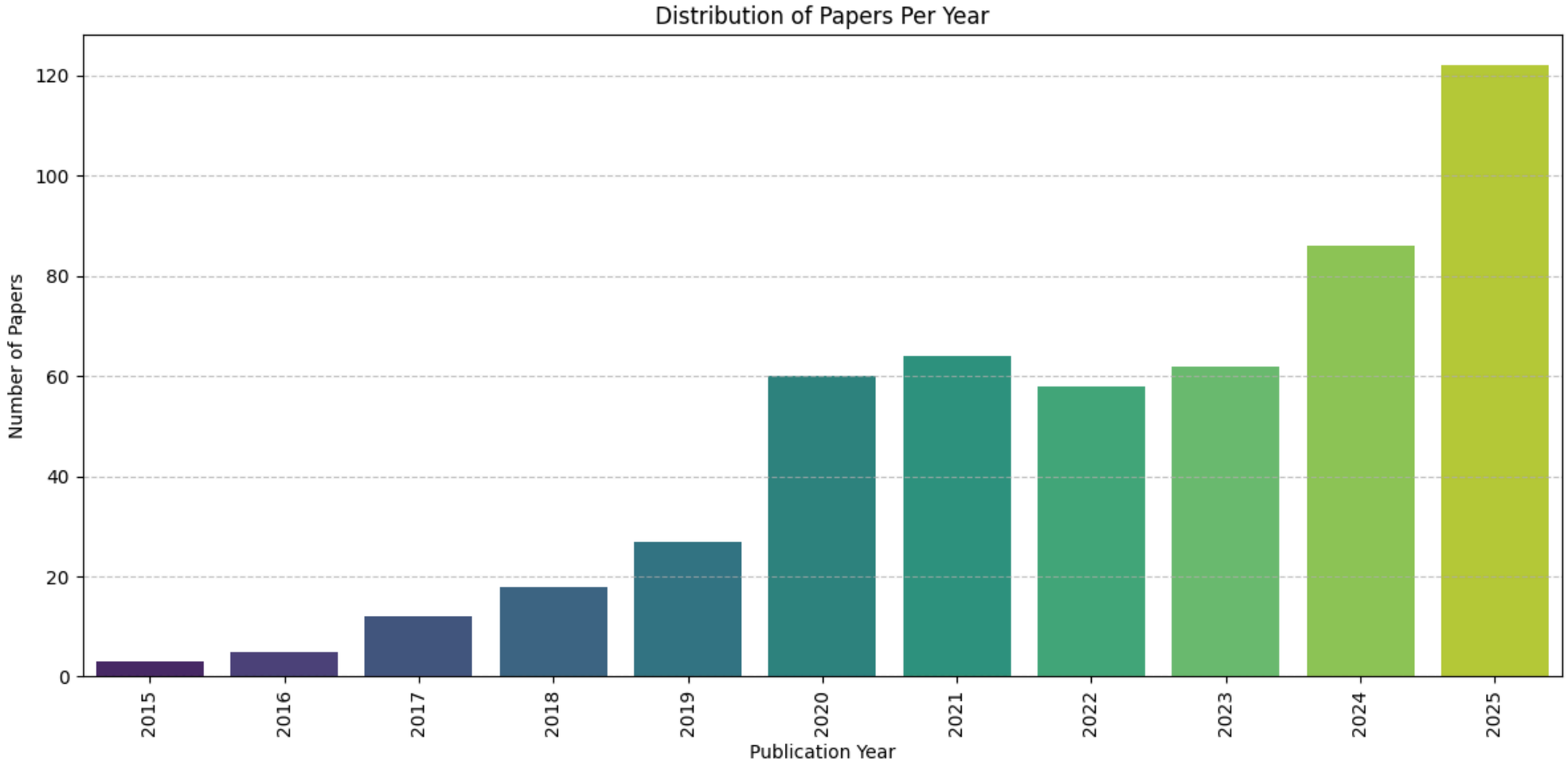
university of groningen

DO YOU REMEMBER?



Photograph: Alamy

AND SUDDENLY ...



Source: ACL Anthology

... AND SO DID I

VeryfIT-benchmark of fact-checked claims for Italian: A CALAMITA challenge J Gili, V Patti, L Passaro, T Caselli Proceedings of the 10th Italian Conference on Computational Linguistics …	4	2024
Overview of the CLEF-2024 CheckThat! lab task 1 on check-worthiness estimation of multigenre content M Hasanain, R Suwaileh, S Weering, C Li, T Caselli, W Zaghouani, ... 25th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF …	12	2024
FC_RUG at CheckThat! 2024: few-shot learning using GEITje for check-worthiness detection in Dutch S Weering, T Caselli 25th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF …	2	2024
Check-IT!: A corpus of expert fact-checked claims for Italian J Gili, L Passaro, T Caselli Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC …	5	2023
Thesis Titan at CheckThat!-2023: Language-Specific Fine-tuning of mDeBERTaV3 for Subjectivity Detection. FA Leistra, T Caselli CLEF (Working Notes), 351-359	1	2023
Overview of the clef-2023 checkthat! lab: Task 2 on subjectivity in news articles A Galassi, F Ruggeri, BC Alberto, A Firoj, T Caselli, K Muchaid, F Antici, ... CEUR Workshop Proceedings 3497, 236-249	22	2023
Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets P Nakov, A Barrón-Cedeño, G Da San Martino, F Alam, R Míguez, ... CEUR Workshop Proceedings 3180, 368-392	79	2022
Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society F Alam, S Shaar, F Dalvi, H Sajjad, A Nikolov, H Mubarak, ... Findings of the association for computational linguistics: EMNLP 2021, 611-649	225	2021
Fighting the COVID-19 infodemic with a holistic BERT ensemble G Tziafas, K Kogkalidis, T Caselli Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship …	11	2021

Source: Google Scholar

FOCUS: LANGUAGE SPECIFIC RESOURCES

- NL corpus for misinformation detection (Alam et al., 2021)
  - Given a claim, determine its veracity (and more)
- NL corpus for subjectivity detection in news (Galassi et al., 2023)
  - Subjective claims should not be fact-checked
- IT corpus for claim verification (Gili et al., 2023; Gili et al., 2024)
  - Professional fact-checked claims, with evidence, vedict, and justification
- NL corpus on checkworthyness detection (Hasanain et al., 2024)
  - Only public interest information is eligible to be fact-checked

FOCUS: LANGUAGE SPECIFIC RESOURCES

- Misinformation can be culturally specific
  - *Wappies* were only in the Netherlands during COVID-19

- Each country has a different set of *characters*
  - *NL: Hugo de Jonge, Femke Luise, Gezond Verstand*
  - *IT: Roberto Speranza, Matteo Salvini, Roberto Burioni*
  - *US: Donald Trump, Anthony Fauci*

- Multilingual approaches do not transfer equally well to any language

NL MISINFORMATION DETECTION - COVID-19 (Alam et al. 2021)

		English				Arabic				Bulgarian				Dutch			
Q.	**Cls.**	**Maj.**	**FT**	**BT**	**RT**	**Maj.**	**FT**	**ArBT**	**XLM-r**	**Maj.**	**FT**	**mBT**	**XLM-r**	**Maj.**	**FT**	**BTje**	**XLM-r**
Binary (Coarse-grained)																	
Q1	2	48.7	**77.7**	**76.5**	**78.6**	56.8	**63.1**	**83.8**	**84.2**	58.3	**75.5**	**84.0**	**87.6**	36.5	**61.9**	**75.4**	**80.0**
Q2	2	91.6	89.0	**92.1**	**92.7**	68.3	**81.7**	**84.0**	**83.1**	95.0	85.2	94.7	**95.0**	64.9	**87.9**	**75.1**	**83.1**
Q3	2	96.3	69.3	**96.4**	**96.9**	96.3	82.0	96.0	**96.3**	96.5	79.3	96.0	**96.5**	62.3	**69.9**	**76.9**	**78.3**
Q4	2	66.7	**96.3**	**85.6**	**89.0**	67.2	**96.2**	**90.3**	**89.0**	86.8	**96.5**	**87.7**	**88.4**	63.9	**72.7**	**77.1**	**83.9**
Q5	2	67.7	**83.8**	**80.6**	**84.4**	46.8	**74.0**	**65.9**	**66.7**	70.5	**81.5**	**80.5**	**82.9**	44.4	**75.3**	**66.8**	**70.9**
Q6	2	86.7	**92.1**	**88.9**	**90.5**	72.5	**79.3**	**88.9**	**89.8**	83.2	**95.0**	**84.5**	**85.1**	84.7	74.9	**86.9**	**88.1**
Q7	2	78.3	**80.6**	**85.5**	**86.1**	57.7	**81.6**	**77.4**	**77.4**	80.1	**87.2**	**81.6**	**81.7**	65.6	**74.1**	**78.3**	**79.6**
Avg.		76.6	**84.1**	**86.5**	**88.3**	66.5	**79.7**	**83.8**	**83.7**	81.5	**85.8**	**87.0**	**88.2**	60.3	**73.8**	**76.6**	**80.5**
Multiclass (Fine-grained)																	
Q2	5	67.9	44.7	**69.2**	**70.6**	62.9	53.3	**75.6**	**76.2**	77.3	**78.8**	**77.8**	**79.3**	36.5	**39.7**	**45.7**	**51.1**
Q3	5	78.9	57.4	**82.5**	**82.8**	44.4	**75.6**	**53.7**	**59.5**	64.2	**78.2**	**68.1**	**68.8**	32.0	**77.7**	**50.9**	**53.9**
Q4	5	19.9	**69.2**	**56.0**	**58.0**	28.1	**54.2**	**46.9**	**50.6**	58.8	**69.0**	**65.6**	**67.1**	21.0	**42.9**	**46.3**	**53.1**
Q5	5	46.8	**84.9**	**62.0**	**70.0**	41.2	**52.6**	**52.6**	**52.4**	36.0	**81.5**	**58.0**	**61.6**	18.4	**69.6**	**40.7**	**46.4**
Q6	8	84.0	71.7	**86.5**	**87.7**	68.7	**71.5**	**82.2**	**84.8**	76.6	**79.6**	**77.2**	**78.8**	74.4	46.0	**76.7**	**76.3**
Q7	10	78.1	**82.4**	**83.4**	**85.3**	13.8	**40.8**	**57.5**	**61.6**	80.1	**66.8**	**81.7**	**81.8**	65.4	45.3	**72.2**	**74.1**
Avg.		62.6	**68.4**	**73.3**	**75.8**	43.2	**58.0**	**61.4**	**64.2**	65.5	**75.6**	**71.4**	**72.9**	41.3	**53.5**	**55.4**	**59.1**

Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society (Alam et al., Findings 2021)

NL MISINFORMATION DETECTION - COVID-19 (Alam et al. 2021)

		English				Arabic				Bulgarian				Dutch			
Q.	**Cls.**	**Maj.**	**FT**	**BT**	**RT**	**Maj.**	**FT**	**ArBT**	**XLM-r**	**Maj.**	**FT**	**mBT**	**XLM-r**	**Maj.**	**FT**	**BTje**	**XLM-r**
								Binary (Coarse-grained)									
Q1	2	48.7														75.4	**80.0**
Q2	2	91														1	**83.1**
Q3	2	9															**78.3**
Q4	2																**83.9**
Q5	2																**70.9**
Q6	2																**88.1**
Q7	2																**79.6**
Avg.																	**80.5**
Q2	5																**51.1**
Q3	5																**53.9**
Q4	5																**53.1**
Q5	5	4														7	**46.4**
Q6	8	84.														76.7	**76.3**
Q7	10	78.1	82													72.2	**74.1**
Avg.		62.6	**68.4**	73.3	**75.8**	43.2	**58.0**	**61.4**	**64.2**	65.5	**75.6**	**71.4**	**72.9**	41.3	**53.5**	**55.4**	**59.1**

- **Language specific models underperform**
- **Multlingual setting is inconsistent**
- **Context-free models (FastText) are very competitive**

Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society (Alam et al., Findings 2021)

NL SUBJECTIVITY DETECTION IN NEWS (Galassi et al. 2023)

	Team	F1
	Multilingual	
1	NN [40]	81.97
-	tarrekko *	81.16
2	Thesis Titan [41]	81.00
-	ES-VRAI [37]	77.96
3	*baseline*	73.56
4	TOBB ETU [42]	66.62
	Arabic	
1	NN [40]	78.75
-	tarrekko *	78.66
2	Thesis Titan [41]	77.53
3	Accenture [35]	72.53
4	*baseline*	65.75
5	TOBB ETU [42]	64.51
	Dutch	
† 1	Thesis Titan [41]	81.43
-	tarrekko *	77.74
2	NN [40]	75.57
3	TOBB ETU [42]	73.01
4	*baseline*	66.68
5	Accenture [35]	62.32

	Team	F1
	English	
-	tarrekko *	78.19
1	DWReCo [36]	78.18
2	Gpachov [39]	77.34
3	Thesis Titan [41]	76.78
4	KUCST *	73.07
5	NN [40]	72.84
6	Fraunhofer SIT [38]	72.72
7	*baseline*	71.98
8	Accenture [35]	68.90
9	TOBB ETU [42]	63.46
10	Awakened *	60.41
11	TUDublin [43]	40.32
	German	
1	Thesis Titan [41]	81.52
2	NN [40]	74.13
-	tarrekko *	73.08
3	TOBB ETU [42]	71.19
4	DWReCo [36]	69.82
5	Fraunhofer_SIT [38]	68.39
6	*baseline*	63.65
7	Accenture [35]	25.58

	Team	F1
	Italian	
1	Thesis Titan [41]	75.75
-	tarrekko *	71.61
2	NN [40]	71.01
3	Accenture [35]	65.52
4	*baseline*	63.70
5	TOBB ETU [42]	63.35
6	TUDublin [43]	45.92
	Turkish	
1	Thesis Titan [41]	89.94
-	tarrekko *	87.01
2	DWReCo [36]	84.11
3	NN [40]	81.21
4	Accenture [35]	78.11
5	*baseline*	77.40
6	TOBB ETU [42]	70.16

Overview of the CLEF-2023 CheckThat! Lab: Task 2 on Subjectivity in News Articles (Galassi et al., 2023, CLEF)

NL SUBJECTIVITY DETECTION IN NEWS (Galassi et al. 2023)

	Team	F1		Team	F1		Team	F1
	Multilingual			**English**			**Italian**	
1	NN [40]	81.97	-	tarrekko *	78.19	1	Thesis Titan [41]	75.75
-	tarrekko *	81.16	1	DWReCo [36]	78.18	-	tarrekko *	71.61
2	Thesis Titan [41]	81.00	2	Gpachov [39]	77.34	2	NN [40]	71.01
4	*baseline*	66.68	5	Fraunhofer_SIT [38]	68.39			
5	Accenture [35]	62.32	6	*baseline*	63.65			
			7	Accenture [35]	25.58			

- **Task can be successfully addressed**
- **Multilingual approaches work better (with some adaptations)**
- **Easier in some languages**

Overview of the CLEF-2023 CheckThat! Lab: Task 2 on Subjectivity in News Articles (Galassi et al., 2023, CLEF)

IT CLAIM VERIFICATION (Gili et al. 2023; Gili et al. 2024)

Model	Accuracy			Weig. Accuracy		
	Full	Small	Enr.	Full	Small	Enr.
LLaMA3.1-8B	**0.594**	**0.567**	**0.578**	0.463	0.472	0.484
LLaMA3.1-70B	0.593	0.556	0.556	0.491	0.483	0.473
Anita-8B	0.408	0.433	0.433	**0.590**	**0.565**	**0.565**
Minerva-7B	0.592	**0.567**	0.567	0.409	0.433	0.433

(Challenging the Abilities of Large Language Models in Italian: a Community Initiative (Nissim et al., 2025, ArXiv)

IT CLAIM VERIFICATION (Gili et al. 2023; Gili et al. 2024)

Model	uracy
	Enr.
LLaM	0.484
LLaM	0.473
ANIT	**0.565**
MINE	0.433

- LLM parametric knowldge is not sufficient
- Model's size has no impact
- Multilingual models are better than language specific ones

(Challenging the Abilities of Large Language Models in Italian: a Community Initiative (Nissim et al., 2025, ArXiv)

NL CHECKWORTHINESS DETECTION (Hasanain et al. 2024)

	Arabic			Dutch			English	
	Team	F1		Team	F1		Team	F1
1	IAI Group	0.569	1	TurQUaz	0.732	1	FactFinders	0.802
2	OpenFact	0.557	2	DSHacker	0.730	2	OpenFact	0.796
3	DSHacker	0.538	3	IAI Group	0.718	3	Fraunhofer SIT	0.780
4	TurQUaz	0.533	4	Mirela	0.650	4	mjmanas54	0.778
5	SemanticCuetSync	0.532	5	Zamoranesis	0.601	5	ZHAW_Students	0.771
6	mjmanas54	0.531	6	FC_RUG	0.594	6	SemanticCuetSync	0.763
7	Fired_from_NLP	0.530	7	OpenFact	0.590	7	SINAI	0.761
8	Madussree	0.530	8	HYBRINFOX	0.589	8	DSHacker	0.760
9	pandas	0.520	9	mjmanas54	0.577	9	IAI Group	0.753
10	HYBRINFOX	0.519	10	DataBees	0.563	10	Fired_from_NLP	0.745
11	Mirela	0.478	11	JUNLP	0.550	11	TurQUaz	0.718
12	DataBees	0.460	12	Fired_from_NLP	0.543	12	HYBRINFOX	0.711
13	Baseline	0.418	13	Madussree	0.482	13	SSN-NLP	0.706
14	JUNLP	0.212	14	Baseline	0.438	14	Checker Hacker	0.696
			15	pandas	0.308	15	NapierNLP	0.675
			16	SemanticCuetSync	0.218	16	Mirela	0.658
						18	DataBees	0.619
						19	Trio_Titans	0.600
						20	Madussree	0.583
						21	pandas	0.579
						22	JUNLP	0.541
						23	Sinai and UG	0.517
						24	grig95	0.497
						25	CLaC	0.494
						26	Aqua_Wave	0.339
						27	Baseline	0.307

Overview of the CLEF-2024 CheckThat! Lab Task 1 on Check-Worthiness Estimation of Multigenre Content (Haisanain et al., 2024, CLEF)

NL CHECKWORTHINESS DETECTION (Hasanain et al. 2024)

Arabic			Dutch			English		
	Team	F1		Team	F1		Team	F1
1	IAI Group	0.569	1	TurQUaz	0.732	1	FactFinders	0.802
2	OpenFact	0.557	2	DSHacker	0.730	2	OpenFact	0.796
23	Sinai and UG	0.517						
24	grig95	0.497						
25	CLaC	0.494						
26	Aqua_Wave	0.339						
27	Baseline	0.307						

- Data augmentation is needed
- Generative LLMs underperform
- Lack of robusteness to domain shifts between train and test

Overview of the CLEF-2024 CheckThat! Lab Task 1 on Check-Worthiness Estimation of Multigenre Content (Haisanain et al., 2024, CLEF)


THAT'S COOL, BUT ...
Made with Piñata Farms

WHAT IS MISINFORMATION?

(Caselli et al., *in progress*)

WHAT IS MISINFORMATION?

In this survey, we examine the relationship between automatically detecting false information online – including fact-checking, and detecting fake news, rumors, and hoaxes – and the core underlying Natural Language Processing (NLP) task needed to achieve this, namely *stance detection*. Therein, we consider mis- and disinformation, which both refer to false information, though disinformation has an additional intention to harm.

Hardalov et al., 2022

(Caselli et al., *in progress*)

WHAT IS MISINFORMATION?

- It's "complicated"

- **MISINFORMATION** is an umbrella term in NLP/AI

- The focus is on the ***veridicality / truthfulness*** of an instance of information
  - claim (written or spoken)
  - a document (e.g., news article or scientific paper)
  - an image
  - a combination of different modalities (e.g., a meme)

(Caselli et al., *in progress*)

WHAT IS MISINFORMATION?

- **MISINFORMATION is an umbrella term in NLP/AI**

Malinformation
Misinformation
Disinformation
Gossip
Rumors

Hoaxes
Conspiracy Theory
Fake News
Propaganda
Fallacies

(Caselli et al., *in progress*)

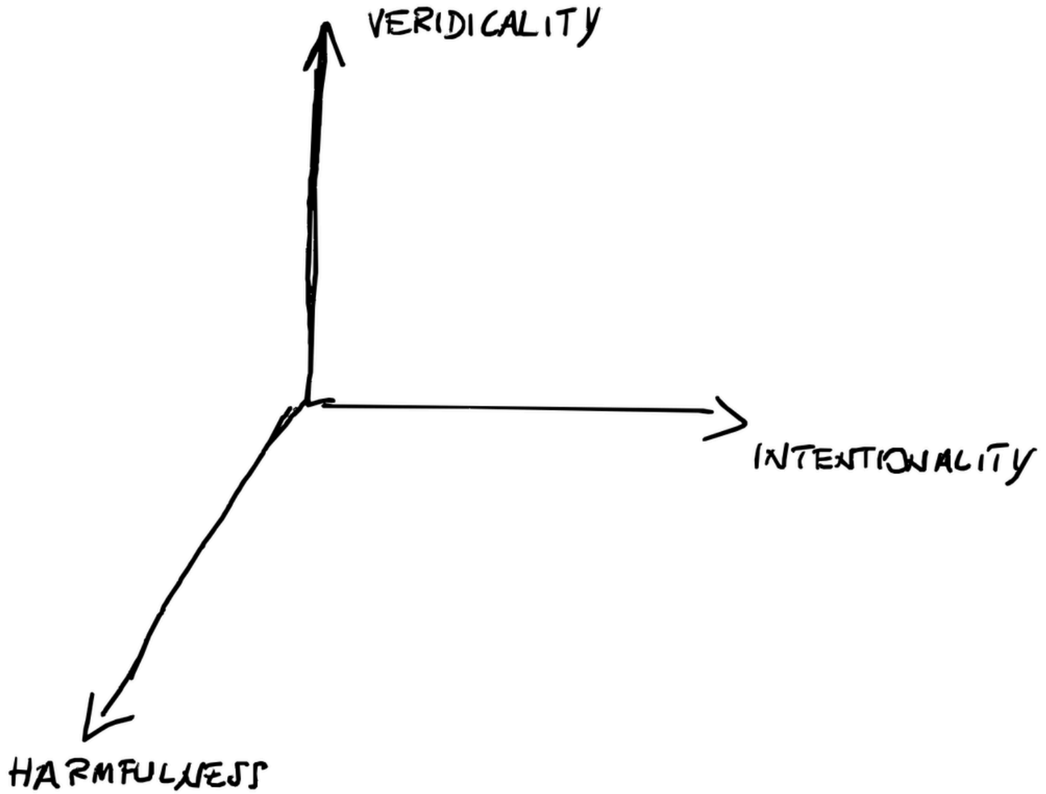
WHAT IS MISINFORMATION?

- Veridicality / Truthfulness is only *one* component
- There can be a specific ***intention*** to mislead someone
- There can be a willingness to ***harm*** someone

(Caselli et al., *in progress*)

WHAT IS MISINFORMATION?

- **Discussing / studying / modeling etc *misinformation* requires understanding the interactions across three axes:**
  - **VERIDICALITY**
  - **INTENTIONALITY**
  - **HARMFULNESS**

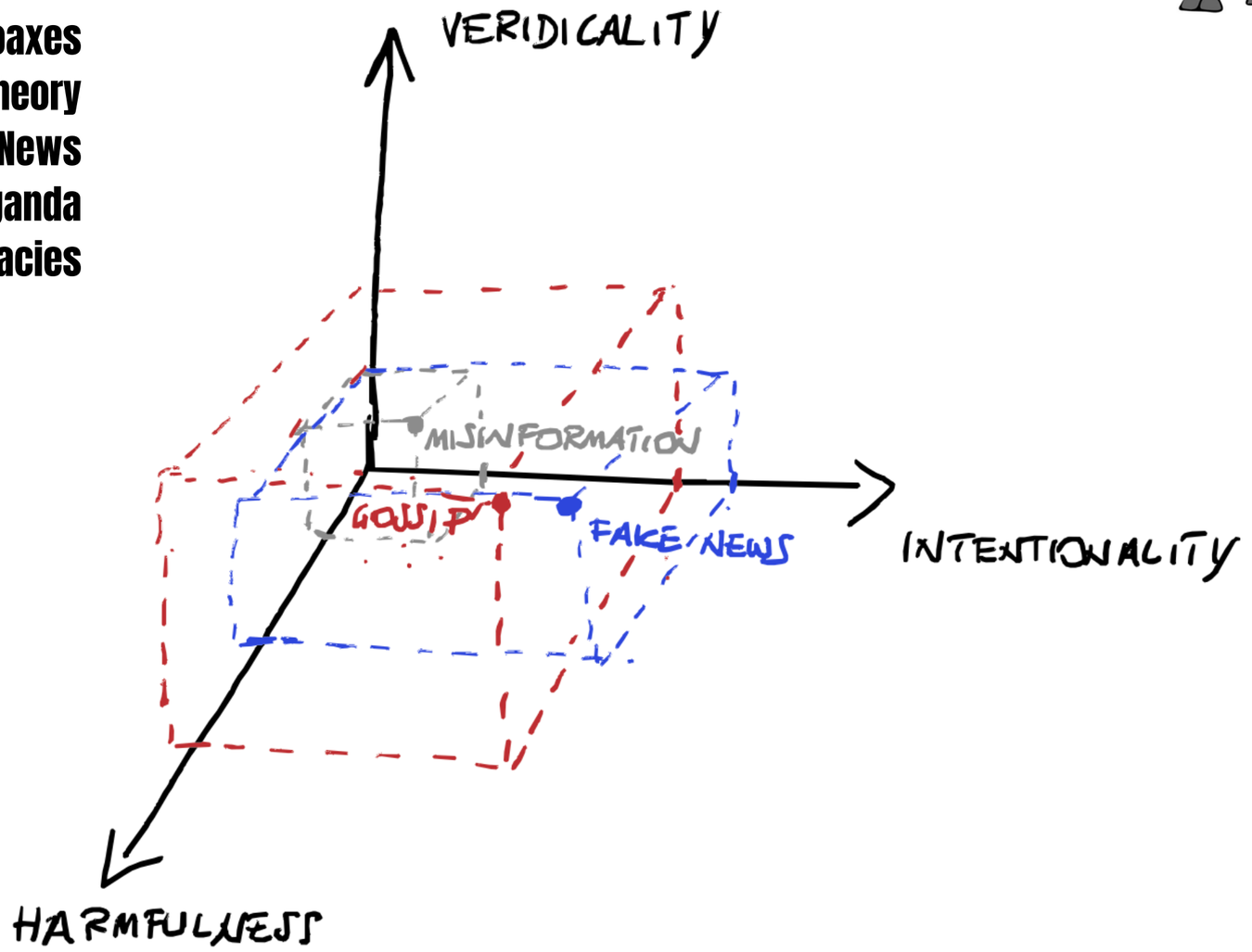


(Caselli et al., *in progress*)

WHAT IS MISINFORMATION?

Malinformation
Misinformation
Disinformation
Gossip
Rumors

Hoaxes
Conspiracy Theory
Fake News
Propaganda
Fallacies



(Caselli et al., *in progress*)

WHAT IS MISINFORMATION?

- **Modeling INTENTIONALITY and HARMFULNESS require access to the *context***

(Caselli et al., *in progress*)

https://en.wikipedia.org/wiki/Chimera_%28mythology%29#/media/File:Chimera_d'arezzo,_fi,_04.JPG

WHAT IS MISINFORMATION?

- Modeling INTENTIONALITY and HARMFULNESS requires access to the ***context***

- ***Context*** requires :

    - access to data* from the message producer(s): sociodemographics; network interactions; credibility
    - connection of specific messages to (broader) narratives
    - provenance of the message and features (bias, partisanship, etc) of the source

(Caselli et al., *in progress*)

WHAT IS MISINFORMATION?

- Most datasets for *misinformation detection* have partial CONTEXTUAL dimensions:
    - Derczynski et al. 2017, Gorrell et al. 2019 (RumorEval): conversation thread
    - Augenstein et al. 2019 (MultiFC): time of claim, time of verification, speaker
    - Alam et al. 2021: explicitly models HARMFULNESS
    - Gili et al. 2023 (CheckIT!): time of claim, time of verification, speaker, political affiliation, links to evidence sources
    - Schlichtkrull et al. 2024 (AVeriTeC): time of claim, speaker, *knowledge store* for evidence retrieval

(Caselli et al., *in progress*)

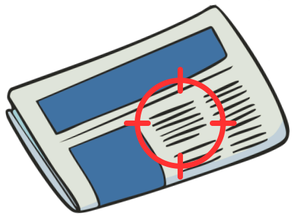
HOW DO WE DETECT MISINFORMATION?

(Caselli et al., *in progress*)

HOW DO WE DETECT MISINFORMATION?

Sources of information

Checkworthy verifiable claims



Evidence gathering

Analysis & evaluation

Verdict assignment

(Caselli et al., *in progress*)

HOW DO WE DETECT MISINFORMATION?

- Pre-LLMs:
  - train → test
  - fine-tune → test

- LLM era:
  - zero-shot / few-shot prompting → test
  - fine-tuning (LoRa) → test
  - further-pretrain + fine-tuning (LoRa) → test

(Caselli et al., *in progress*)

HOW DO WE DETECT MISINFORMATION?

- Datasets are primary sources for modelization and assessment

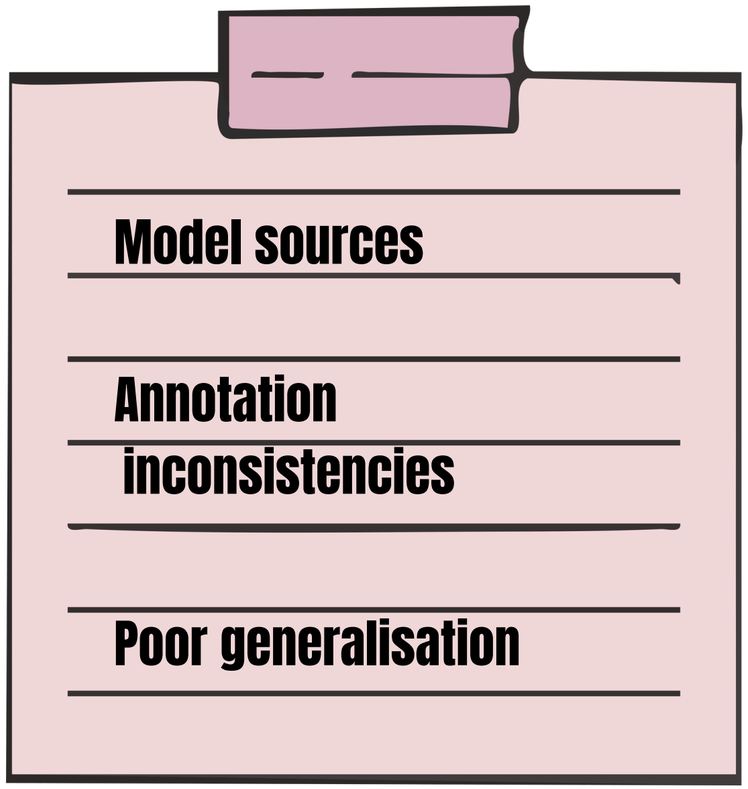
- We build datasets using three paradigms:
  - distant supervision
  - expert annotation
  - crowd annotation

(Caselli et al., *in progress*)

HOW DO WE DETECT MISINFORMATION?

- Datasets are primary sources for modelization and assessment

- We build datasets using three paradigms:
  - distant supervision
  - expert annotation
  - crowd annotation

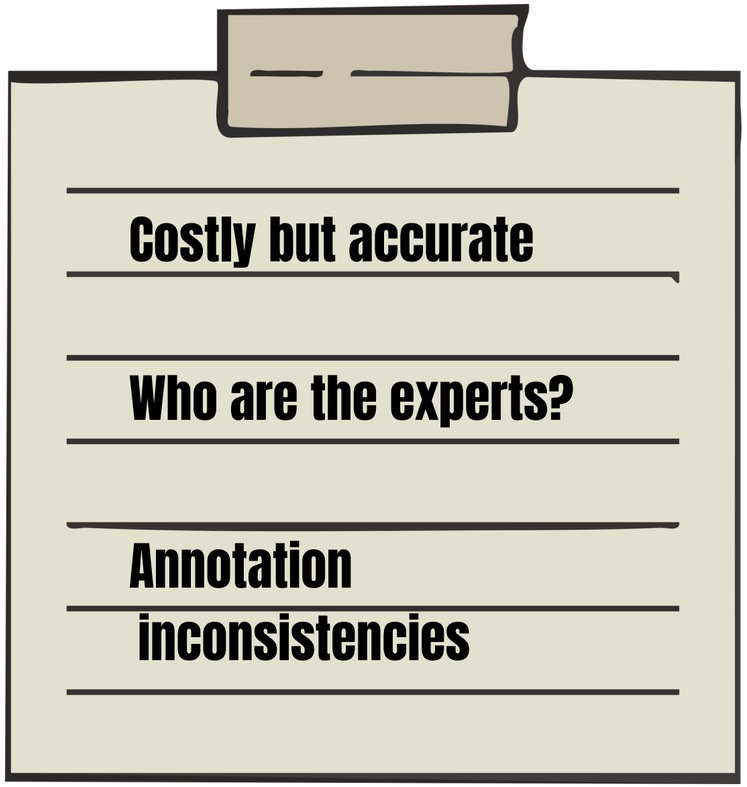


(Caselli et al., *in progress*)

HOW DO WE DETECT MISINFORMATION?

- Datasets are primary sources for modelization and assessment

- We build datasets using three paradigms:
  - distant supervision
  - expert annotation
  - crowd annotation



(Caselli et al., *in progress*)

HOW DO WE DETECT MISINFORMATION?

- Datasets are primary sources for modelization and assessment

- We build datasets using three paradigms:
  - distant supervision
  - expert annotation
  - crowd annotation

Disagreement is NOT noise

Subjective (descriptive annotation; Röttger et al. 2022)

Quality (low)

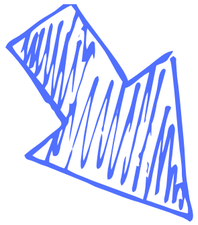
(Caselli et al., *in progress*)

HOW DO WE DETECT MISINFORMATION?

- **Core task: determine the veracity status of a piece of information**

> In March 2026, Spain's Prime Minister Pedro Sánchez announced that anyone who insulted the Prophet Muhammad or the Islamic faith would face a five-year prison sentence.

Directly predicting a veracity label



Veracity label prediction mediated via evidence

https://www.snopes.com/fact-check/spanish-prime-minister-islam/

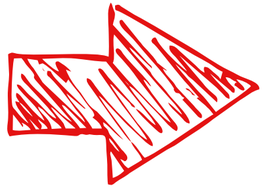
(Caselli et al., *in progress*)

HOW DO WE DETECT MISINFORMATION?

- **Core task: determine the veracity status of a piece of information**

> In March 2026, Spain's Prime Minister Pedro Sánchez announced that anyone who insulted the Prophet Muhammad or the Islamic faith would face a five-year prison sentence.

Directly predicting a veracity label



- **Strong dependance on the training data**
- **Reliance on surface-level features**
- **Reliance on parametric knowledge (LLMs)**

https://www.snopes.com/fact-check/spanish-prime-minister-islam/

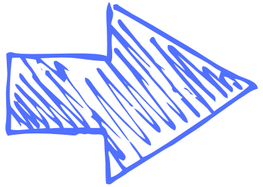
(Caselli et al., *in progress*)

HOW DO WE DETECT MISINFORMATION?

- **Core task: determine the veracity status of a piece of information**

> In March 2026, Spain's Prime Minister Pedro Sánchez announced that anyone who insulted the Prophet Muhammad or the Islamic faith would face a five-year prison sentence.

Veracity label predicted via evidence



- **Claim extraction & normalization**
- **Use of auxiliary tasks (stance detection)**
- **Need access to vetted knonwledge repositories**

https://www.snopes.com/fact-check/spanish-prime-minister-islam/

(Caselli et al., *in progress*)

HOW DO WE DETECT MISINFORMATION?

- Glockner et al. (2022) points out that automatic veracity determination does not take into account how humans do this

- NLP veracity determination relies on:
  - surface cues within the claim
  - claim metadata
  - evidence documents

- Access to the semantic content of a claim is not always sufficient to identify refuting sources

(Caselli et al., *in progress*)

Quentin Tarantino, *Pulp Fiction*, *1994*

WE HAPPY?

- Over time we have lost paying attention to the language of misinformation

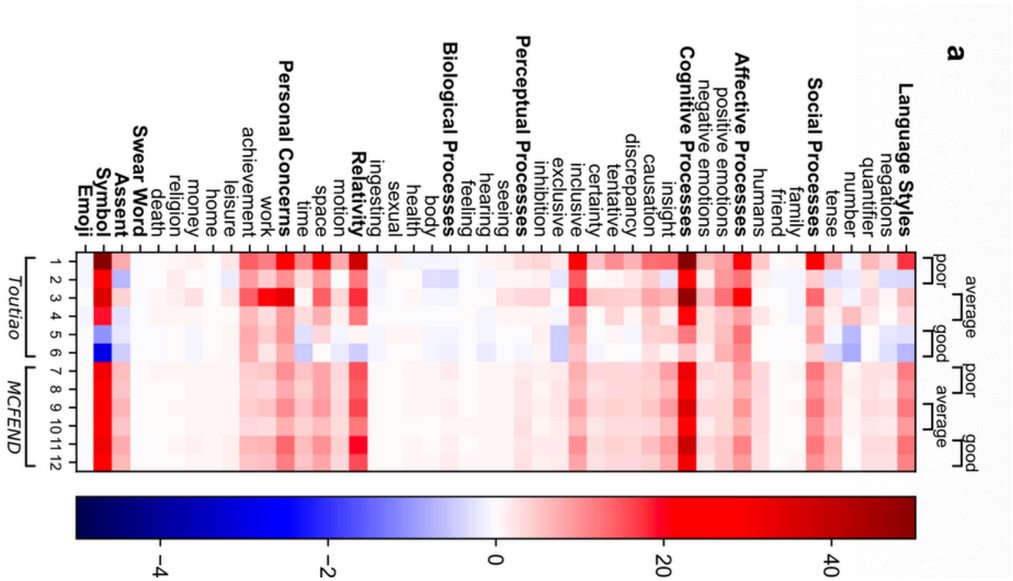
- General trends of “misinformative texts”:
  - emotionally loaded (Rosso et al. 2021)
  - uncertainty markers (Zhou et al. 2021)
  - more persuasive (Zhou et al. 2021)
  - easier to read (Choudhary & Arora 2021)
  - contain propaganda techniques (Huang et al. 2023)

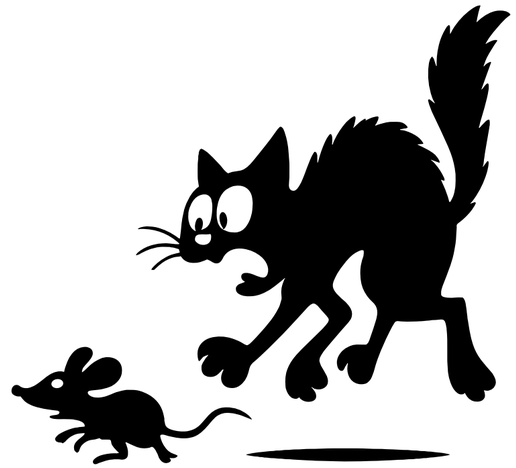
- Few work actively model or attempt to integrate these features in detection systems
  - we blindly rely on the “power” of LLMs and their embedding representations

(Caselli et al., *in progress*)

WE HAPPY?

- **Since 2023, Malevolent Agents have a strong ally: LLMs & Gen AI**
- **LLMs can mimic writing styles of organizations and sociodemgraphic groups (persona prompting)**
- **Machine generated texts can be detected but ...**





(Caselli et al., *in progress*)

Linguistic features of AI mis/disinformation and the detection limits of LLMs (Ma et al. 2025 - Nature Communications)

WE HAPPY?

- **Since 2023, Malevolent Agents have a strong ally: LLMs & Gen AI**
- **LLMs can mimic writing styles of organizations and sociodemographic groups (persona prompting)**
- **Real challenge: detect human generated misinformation (or use of LLMs to manipulate human texts)**
  - **mixture of accurate and inaccurate information**
  - **writing style mimic reliable sources of information**

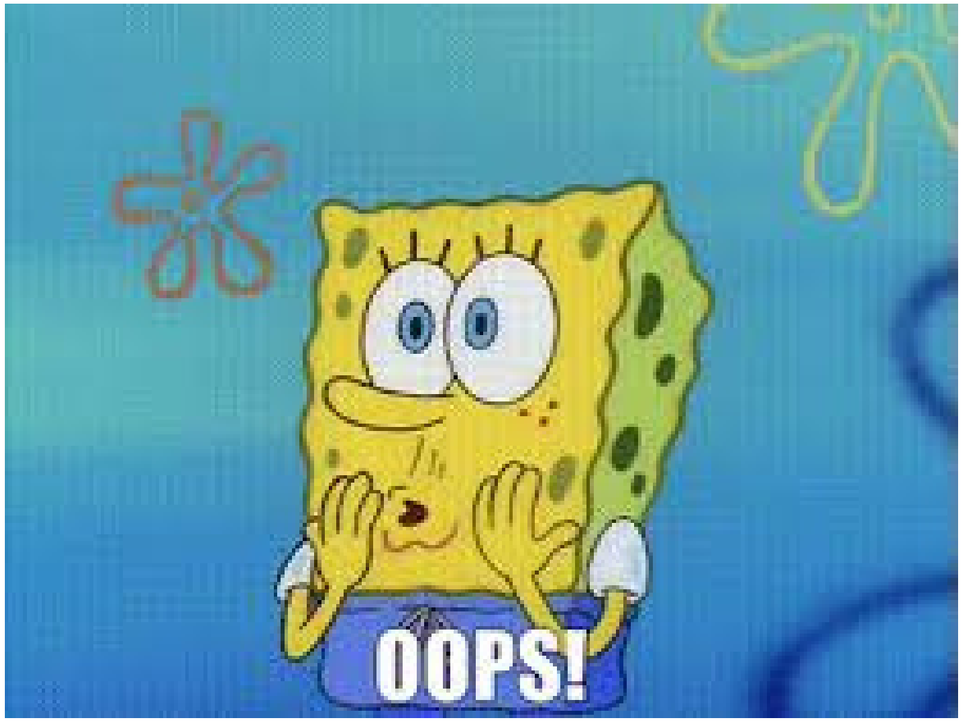
AJDABIYAH , Libya | Thu Apr 7 , 2011 6:34 pm EDT AJDABIYAH , Libya -LRB- Reuters -RRB- - Rebels fighting to overthrow Muammar Gaddafi said five of their fighters were killed ... "In rebel-held eastern Libya, wounded rebels being brought to a hospital Ajdabiyah said their trucks and tanks were hit on Thursday by a NATO air strike outside Brega. ~~NATO said it was investigating an attack by its aircraft on a tank column in the area along the Mediterranean coast on Thursday , saying the situation was " unclear and fluid . "~~ Rebels said at least five of their fighters were killed when NATO planes mistakenly bombed a rebel tank column near the contested port. "A number of vehicles were hit by a NATO strike ", officers from UN concluded. The fighting for Brega , the only active front , has dragged on for a week ...

Faking Fake News for Real Fake News Detection: Propaganda-Loaded Training Data Generation (Huang et al., 2023, ACL)

(Caselli et al., *in progress*)

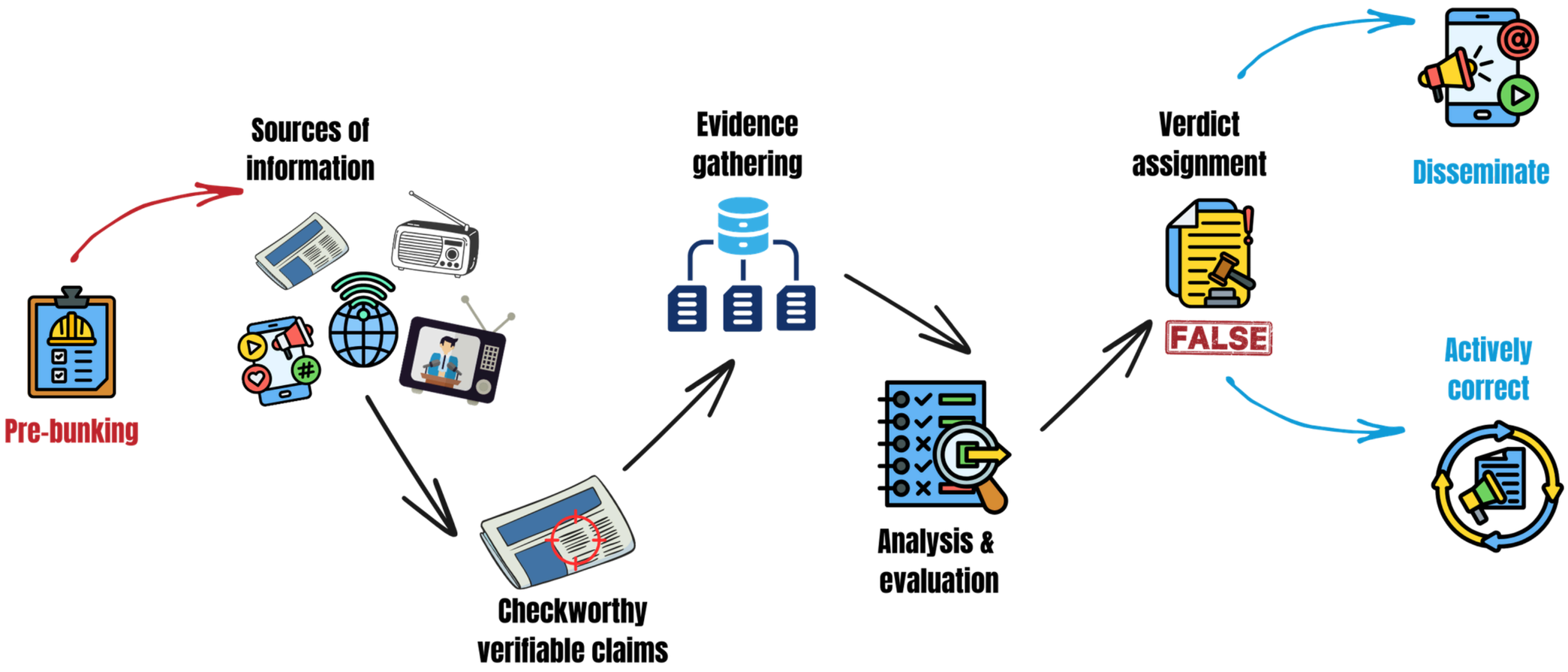
WE HAPPY?

- The missing party: professionals / stakeholders (Glockner et al., 2022)
  - many papers contain claims like “we want to help fact-checkers”
- WE NEVER REALLY ASKED THEM WHAT THEY NEED / WANT
- Our impact is only quantified in terms of F1 scores
- We have (had) zero impact on improving the work of professional fact-checkers / journalists



(Caselli et al., *in progress*)

WE HAPPY?



(Caselli et al., *in progress*)

WE HAPPY?

- **We need start adopting Participatory Design methods**

PD principles	1.	*PD is about consensus and conflict*	• PD entails a process of mutual learning between researchers and community • PD adopts a variety of research and design methods (workshops, participants observation, cards, ...)
	2.	*Design is an inherently disordered and unfinished process*	• *Use-before-use*: tool's use is envisioned *before* the tool is actually implemented • *Design-after-design*: tool's design isn't exhausted with delivery, but will be modified by the users' appropriation, use, and feedback
	3.	*Communities are often not completely determined* a priori	• Communities are not a unitary whole, but can get formed within and through the design process
NLP tools	4.	*Data and communities are not separate things*	• The shift from *language as data* to *language as people*: language data are produced by human speakers • Communities should be involved in the different stages of the NLP pipeline
	5.	*Community involvement is not scraping*	• Collaboration with a community should imply ethical engagement practices based on respect, equity and reciprocity • Researchers should communicate to the community the usage of the collected data in a transparent and appropriate way
	6.	*Never stop designing*	• Community adaptation should be treated as a feature of an NLP system at the design stage
Researchers' reflexivity	7.	*Text is a means rather than an end*	• The linguistic output of NLP systems should serve people's needs rather than imitate people's production of language.
	8.	*The thin red line between consent and intrusion*	• Do not assume that community members are technology experts nor technologically illiterate • A community's refusal to collaboration is a risk that must be accepted
	9.	*The need to combine research goals, funding, and concrete social political dynamics*	• Designers and researchers as intermediaries between the interests of the different actors involved (project beneficiaries, investors, funding agencies, and other stakeholders' goals)

Guiding Principles for Participatory Design-inspire Natural Language Processing (Caselli et al. 2021 - NLP4PI)

(Caselli et al., *in progress*)

WE HAPPY?

- **We need start adopting Participatory Design methods**

PD principles	1.	*PD is about consensus and conflict*	• PD entails a process of mutual learning between researchers and community • PD adopts a variety of research and design methods (workshops, participants observation, cards, ...)
	2.	*Design is an inherently disordered and unfinished process*	• *Use-before-use*: tool's use is envisioned *before* the tool is actually implemented • *Design-after-design*: tool's design isn't exhausted with delivery, but will be modified by the users' appropriation, use, and feedback
	3.	*Communities are often not completely determined a priori*	• Communities are not a unitary whole, but can get formed within and through the design process
	4.	*Data and communities are not separate things*	• The shift from *language as data* to *language as people*: language data are produced by human speakers • Communities should be involved in the different stages of the NLP pipeline
NL…		*scraping*	based on respect, equity and reciprocity
	5.	*Participatory Data Management*	• Researchers should communicate to the community the usage of the collected data in a transparent and appropriate way
	6.	*Never stop designing*	• Community adaptation should be treated as a feature of an NLP system at the design stage
Researchers' reflexivity	7.	*Text is a means rather than an end*	• The linguistic output of NLP systems should serve people's needs rather than imitate people's production of language.
	8.	*The thin red line between consent and intrusion*	• Do not assume that community members are technology experts nor technologically illiterate • A community's refusal to collaboration is a risk that must be accepted
	9.	*The need to combine research goals, funding, and concrete social political dynamics*	• Designers and researchers as intermediaries between the interests of the different actors involved (project beneficiaries, investors, funding agencies, and other stakeholders' goals)

Guiding Principles for Participatory Design-inspire Natural Language Processing (Caselli et al. 2021 - NLP4PI)

(Caselli et al., *in progress*)

WE HAPPY?

- **We need start adopting Participatory Design methods**

PD principles	1.	*PD is about consensus and conflict*	• PD entails a process of mutual learning between researchers and community • PD adopts a variety of research and design methods (workshops, participants observation, cards, ...)
	2.	*Design is an inherently disordered and unfinished process*	• *Use-before-use*: tool's use is envisioned *before* the tool is actually implemented • *Design-after-design*: tool's design isn't exhausted with delivery, but will be modified by the users' appropriation, use, and feedback
	3.	*Communities are often not completely determined* a priori	• Communities are not a unitary whole, but can get formed within and through the design process
NLP tool	5.	*Community involvement is not scraping*	• Collaboration with a community should imply ethical engagement practices based on respect, equity and reciprocity • Researchers should communicate to the community the usage of the collected data in a transparent and appropriate way
	6.	*Never stop designing*	• Community adaptation should be treated as a feature of an NLP system at the design stage
Researchers' reflexivity	7.	*Text is a means rather than an end*	• The linguistic output of NLP systems should serve people's needs rather than imitate people's production of language.
	8.	*The thin red line between consent and intrusion*	• Do not assume that community members are technology experts nor technologically illiterate • A community's refusal to collaboration is a risk that must be accepted
	9.	*The need to combine research goals, funding, and concrete social political dynamics*	• Designers and researchers as intermediaries between the interests of the different actors involved (project beneficiaries, investors, funding agencies, and other stakeholders' goals)

Guiding Principles for Participatory Design-inspire Natural Language Processing (Caselli et al. 2021 - NLP4PI)

(Caselli et al., *in progress*)

WE HAPPY?

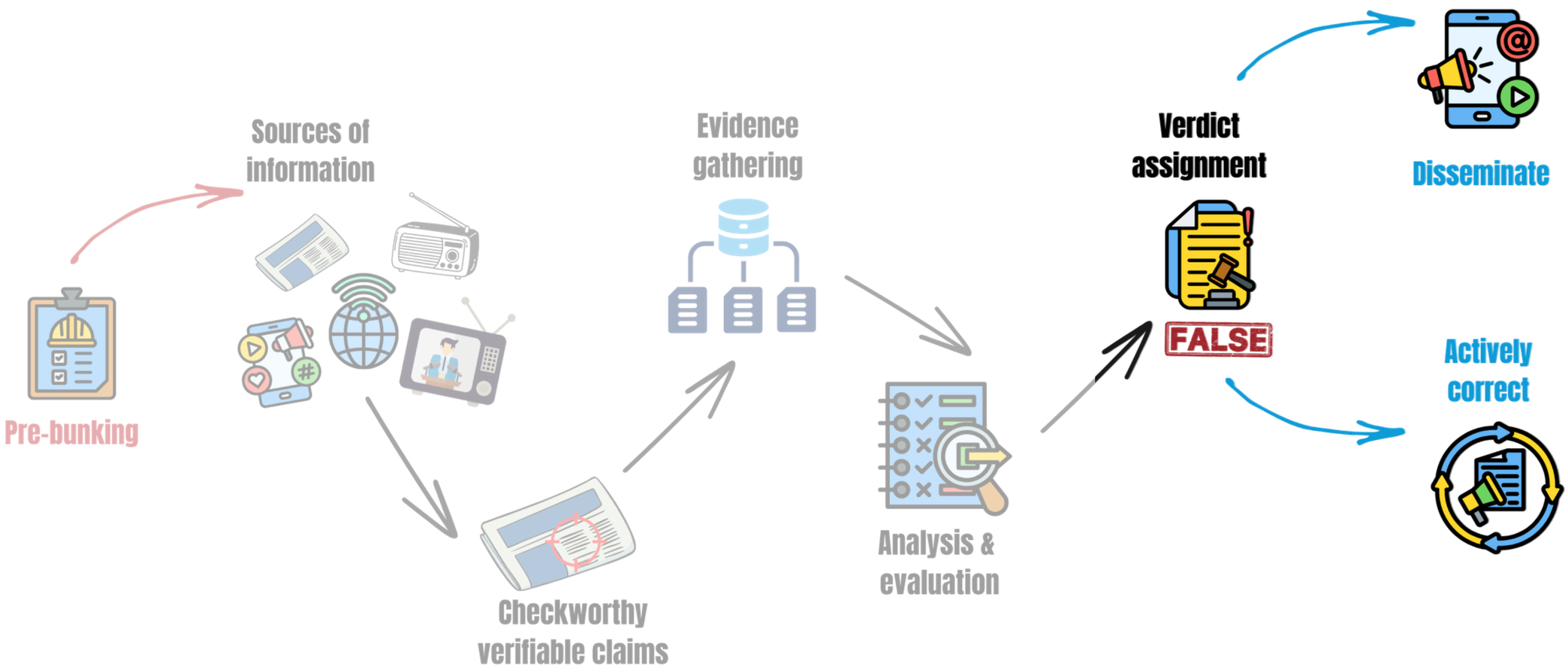
- **We need start adopting Participatory Design methods**

PD principles	1.	*PD is about consensus and conflict*	• PD entails a process of mutual learning between researchers and community • PD adopts a variety of research and design methods (workshops, participants observation, cards, ...)
	2.	*Design is an inherently disordered and unfinished process*	• *Use-before-use*: tool's use is envisioned *before* the tool is actually implemented • *Design-after-design*: tool's design isn't exhausted with delivery, but will be modified by the users' appropriation, use, and feedback
	3.	*Communities are often not completely determined* a priori	• Communities are not a unitary whole, but can get formed within and through the design process
P tools	4.	*Data and communities are not separate things*	• The shift from *language as data* to *language as people*: language data are produced by human speakers • Communities should be involved in the different stages of the NLP pipeline
	5.	*Community involvement is not scraping*	• Collaboration with a community should imply ethical engagement practices based on respect, equity and reciprocity
	6.	*Never stop designing*	• Community adaptation should be treated as a feature of an NLP system at the design stage
Researchers' reflexivity	7.	*Text is a means rather than an end*	• The linguistic output of NLP systems should serve people's needs rather than imitate people's production of language.
	8.	*The thin red line between consent and intrusion*	• Do not assume that community members are technology experts nor technologically illiterate • A community's refusal to collaboration is a risk that must be accepted
	9.	*The need to combine research goals, funding, and concrete social political dynamics*	• Designers and researchers as intermediaries between the interests of the different actors involved (project beneficiaries, investors, funding agencies, and other stakeholders' goals)

Guiding Principles for Participatory Design-inspire Natural Language Processing (Caselli et al. 2021 - NLP4PI)

(Caselli et al., *in progress*)

WE HAPPY?



(Caselli et al., *in progress*)

WE HAPPY?

- Cognitive factors faciliate the belief in misinformation (Lewandowsky et al 2012)
  - exposure to misinformation is like a bad stain: it leaves a trace

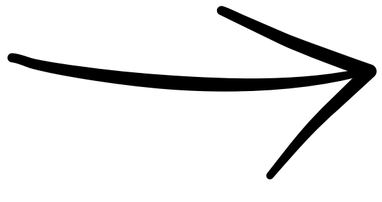
- Correction of misinformation improves belief accuracy (Porter & Wood 2024)
  - flagging misinformation is not enough
  - correction does not result in signficant backfire effects (Porter & Wood 2024)

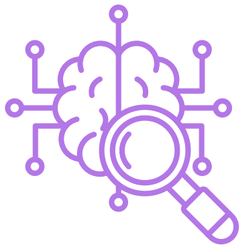
- Fact-checking is ONLY ONE possibile factual correction
  - alternative and effective correction strategies exist (Prike & Eckert 2023)

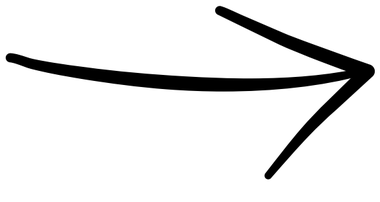
(Caselli et al., *in progress*)

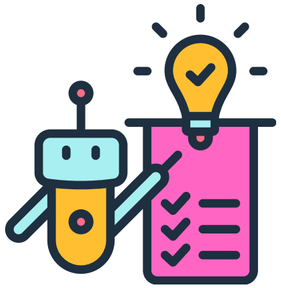
WE HAPPY?

We need to move from flagging to active and interactive corrections









Het klimaat verandert altijd, dus de huidige opwarming is niet bijzonder.

`Momenteel wordt een stijging van de gemiddelde temperatuur op aarde waargenomen. source: wikipedia.nl`

De uitspraak is misleidend. Het klopt dat het klimaat vaak veranderde, maar de huidige opwarming verloopt extreem snel [...]

DutCH-CRAG: Dutch Climate and Health misinformation Correction with Retrieval Augment Generation (Kinds et al. 2025 - CLIN)

(Caselli et al., *in progress*)

“THIS IS THE END, BEAUTIFUL FRIEND”

SUMMARY AND TAKE-HOME MESSAGES

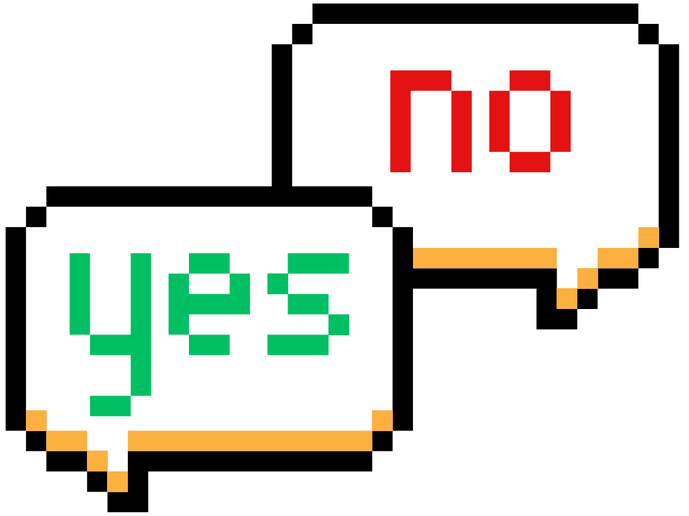
- We live in a world and age flooded by misinformation
  - does it make sense what we have done/are doing?

SUMMARY AND TAKE-HOME MESSAGES

- We live in a world and age flooded by misinformation
  - does it make sense what we have done/are doing?

- Higly relevant societal problem
- Vaired set of working tools
- Expanding to multiple languages (slowly) and domains

- Poor connections with communities outside NLP
- Shortcuts and simplifications
- Focus on the symptoms rather than the disease

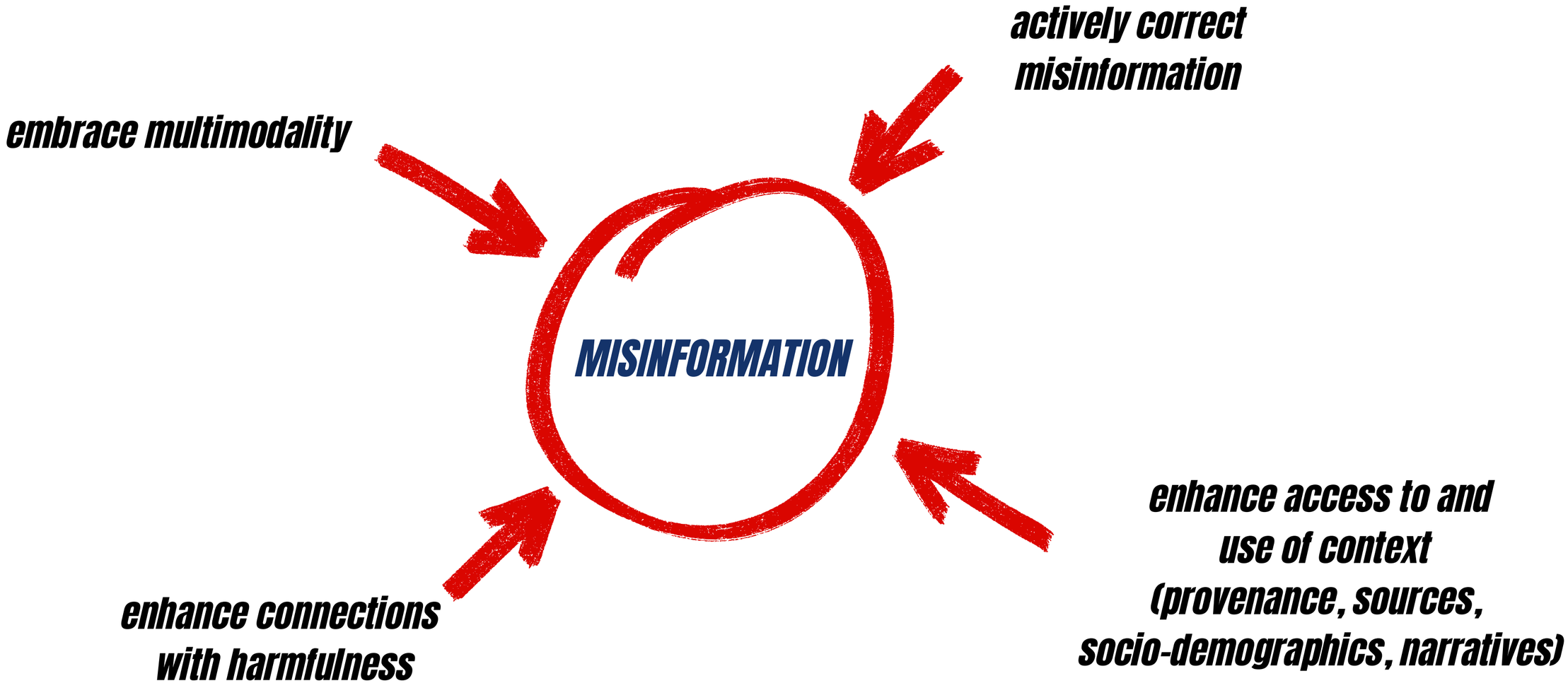


SUMMARY AND TAKE-HOME MESSAGES

- We live in a world and age flooded by misinformation
  - what is the ultimate goal?

- Mitigate presence of misinformation?
- Reduce exposure to misinformation?
- Eliminate misinformation?
- Resources to help professionals?
- Resources to help citizens?
- ....

SUMMARY AND TAKE-HOME MESSAGES



THANK YOU!!

QUESTIONS??