

Disinformation detection

Paolo Rosso

2023-24

DSIIIC

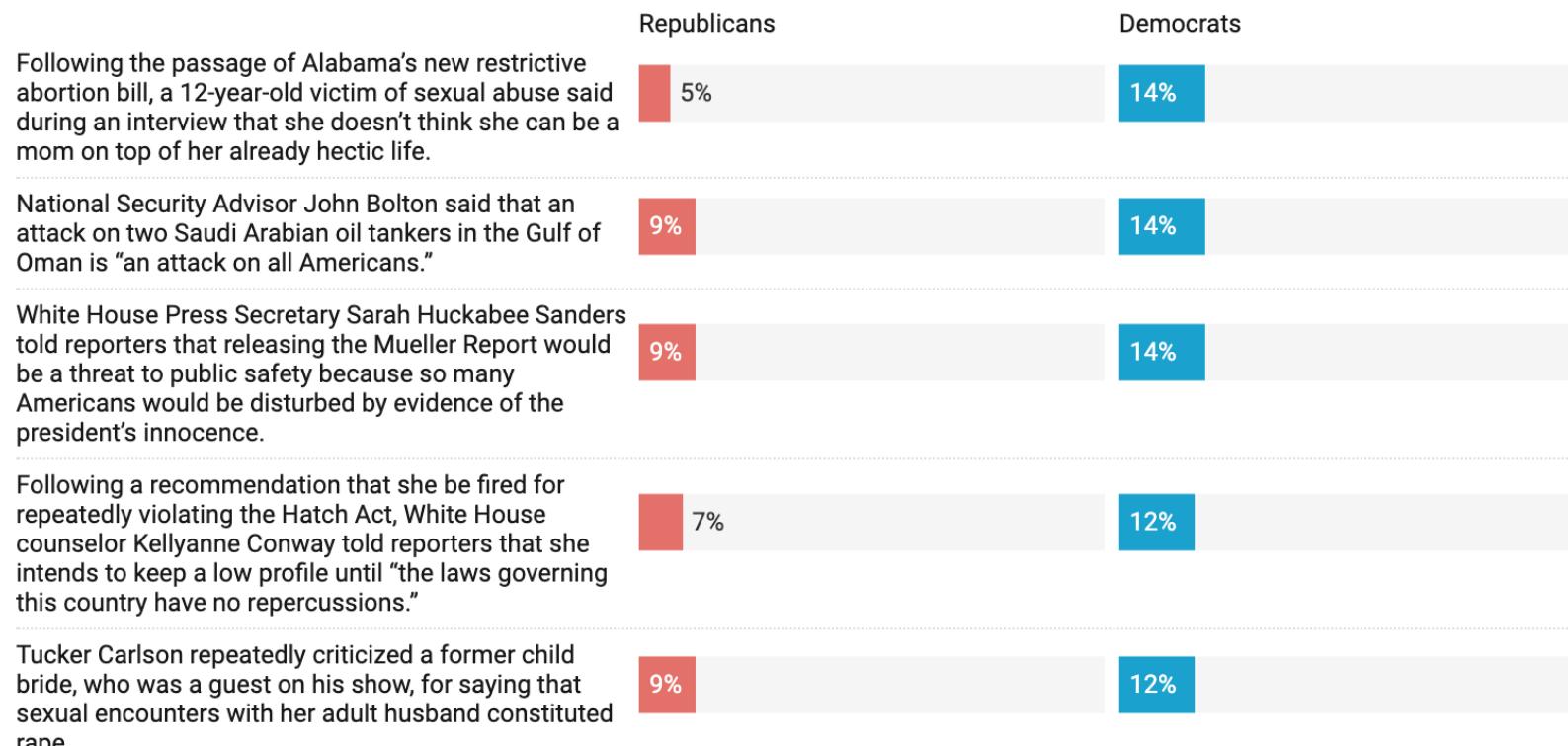


UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Misinformation (e.g. also satire) vs...

Top 5 most-believed satirical claims by The Onion...

...and the percentage of Republicans and Democrats who labeled the claims 'definitely true.'



Surveys conducted between Feb. 20, 2019 and July 31, 2019

Chart: The Conversation, CC-BY-ND • [Get the data](#)

...Disinformation

- 2017 French Presidential election
- Creation of a sophisticated **duplicate version of the Belgian newspaper Le Soir**, with a false article claiming that Macron was being funded by Saudi Arabia
- Circulation of documents online claiming falsely that Macron had opened an offshore bank account in the Bahamas

CrossCheck, Was Macron's Campaign for the French Presidency Funded by Saudi Arabia? <https://crosscheck.firstdraftnews.com/checked-french/macrons-campaign-french-presidencyfinanced-saudi-arabia/>

CrossCheck, Did Emmanuel Macron Open an Offshore Account? <https://crosscheck.firstdraftnews.com/checked-french/emmanuel-macron-open-offshore-account/>

LE SOIR  14° min 4°  -0.08% BEL 20 16/02 10:15

Actu Sports Culture Économie Débats Blogs Image

Actu Monde France

Penelopegate: pas de classement sans suite pour François Fillon Tensions à Paris, en marge d'une manifestation en soutien à Théo

 Recommander  Partager 2 420  15  Share 5  69 

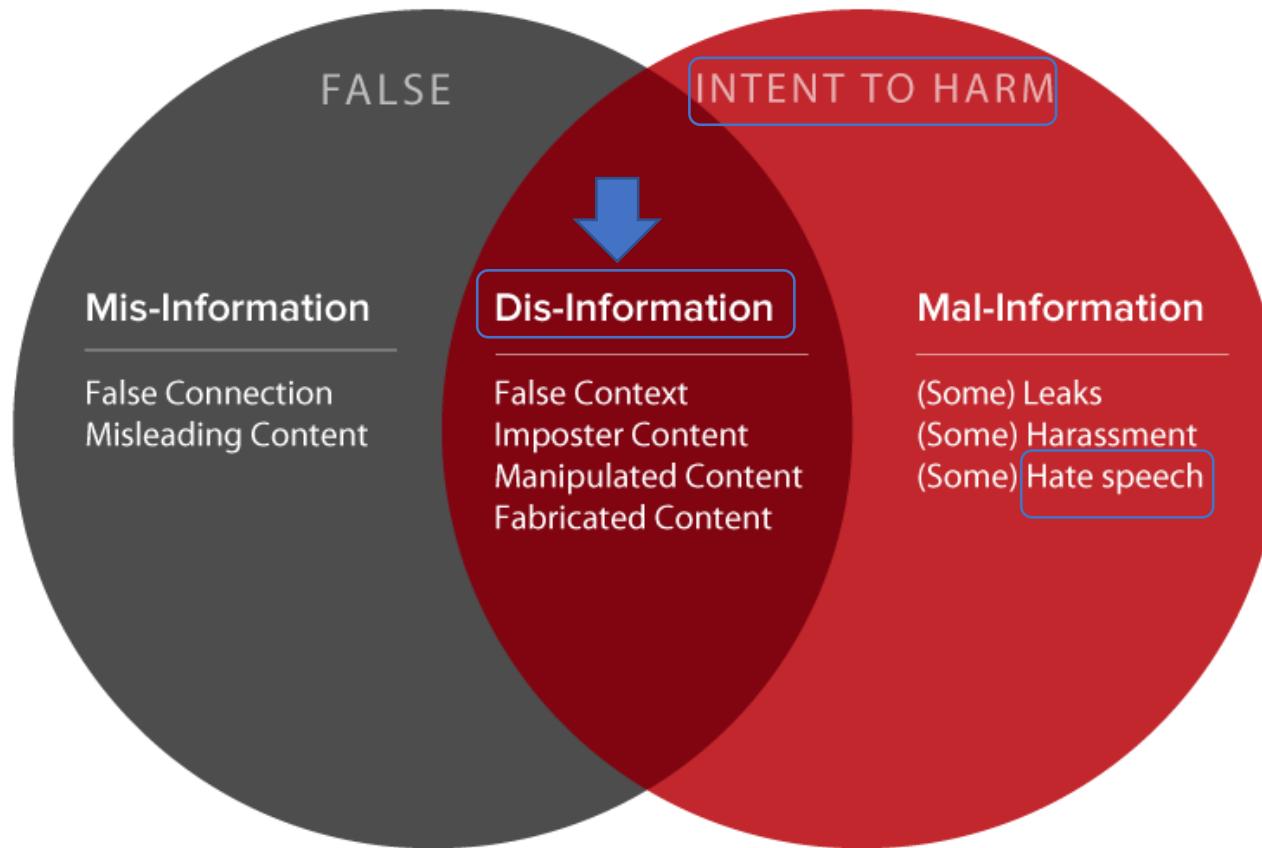
Emmanuel Macron, candidat préféré de l'Arabie Saoudite à l'élection présidentielle

Afp Mis en ligne mercredi 15 février 2017 Emmanuel Mac

FAKE NEWS



Harmful information



Wardle C., Derakhshan H. One year on, we are still not recognizing the complexity of **information disorder online**.

https://firstdraftnews.org/latest/coe_infodisorder/

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

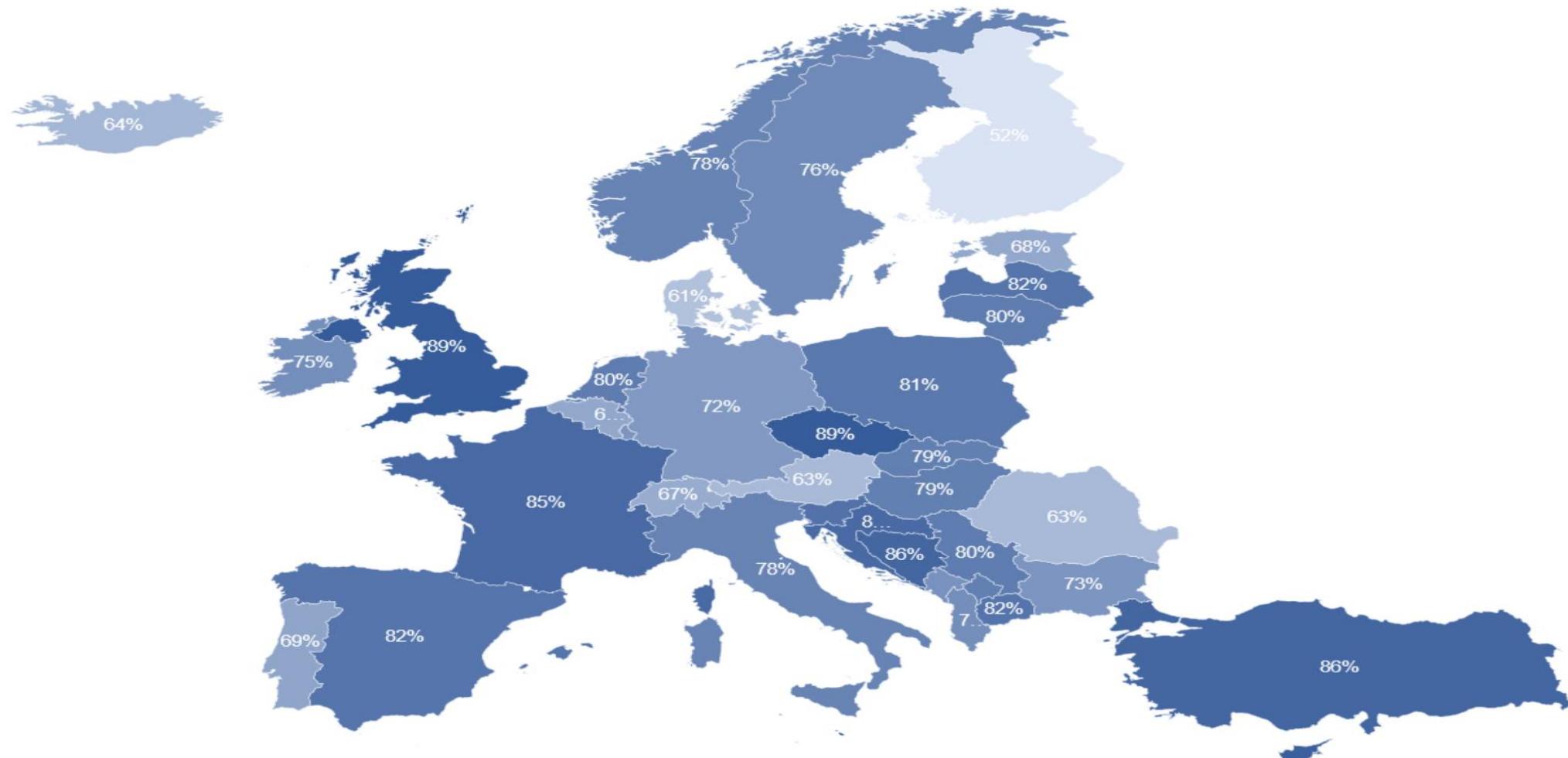
https://data.europa.eu/euodp/es/data/dataset/S2228_92_4_503_ENG

- 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>

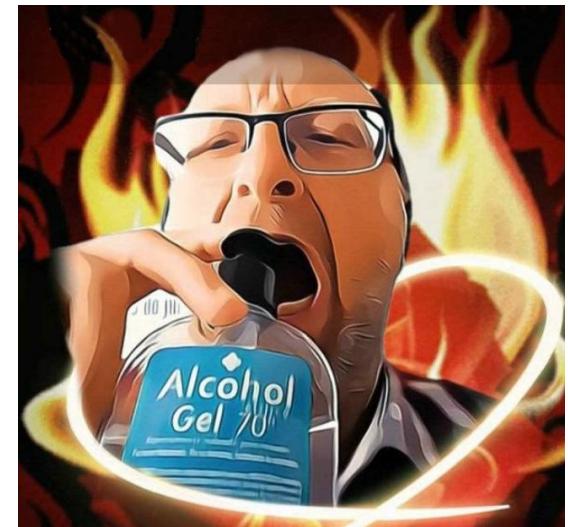
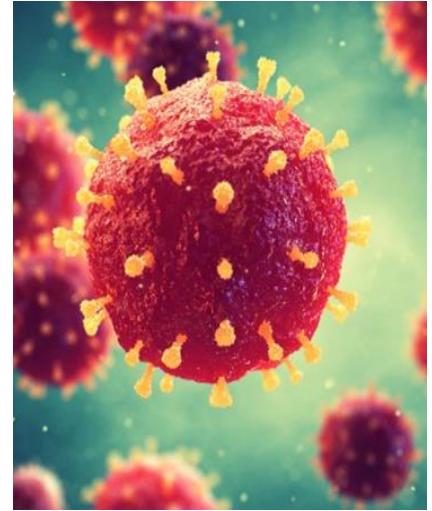
The perception of disinformation as a problem in Europe



Badillo-Matos A. et al. (2023). Analysis of the Impact of Disinformation on Political, Economic, Social and Security Issues, Governance Models and Good Practices: The cases of Spain and Portugal. Pamplona: IBERIFIER.

Disinfodemics during the pandemic

- **Infodemics**, often including **rumours**, and **conspiracy theories**, have been common during the **COVID-19 pandemic**
- A rash of poisonings is tied to **drinking hand sanitizer** that contained methanol after President Donald Trump mused on ingesting disinfectants to treat the novel coronavirus





IBERIFIER

Iberian Media Research
& Fact-Checking

European Digital Media Observatory

Universities



Universidad
de Navarra



UNIVERSITAS
Miguel Hernández



VNIVERSITAT
DE VALÈNCIA



UNIVERSIDAD
DE GRANADA

iscte

INSTITUTO
UNIVERSITARIO
DE LISBONA



uc3m

Universidad
Carlos III
de Madrid



CEU
Universidad
San Pablo



UNIVERSITAT
POLITÈCNICA
DE VALENCIA



universidade
de aveiro

Fact-checkers and news agencies



verificat

Polígrafo

LUSA

Agência de Notícias de Portugal

Multidisciplinary



REAL INSTITUTO
elcano
ROYAL INSTITUTE



FEDERACIÓN
ESPAÑOLA
PARA LA CIENCIA
Y LA TECNOLOGÍA



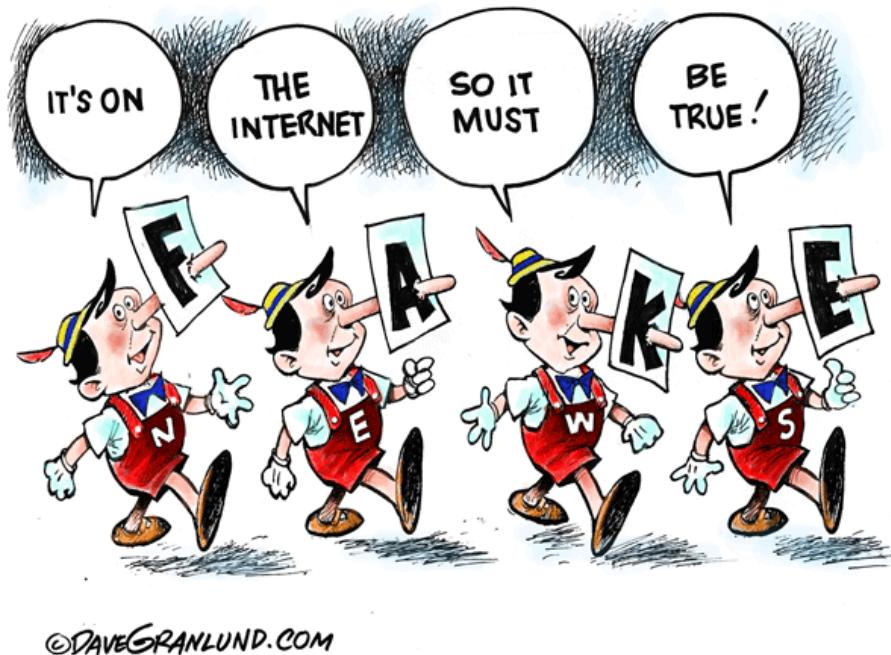
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación



OberCom
INSTITUCIÓN CENTRAL DE COMPUTACIÓN



False information: difficult to be detected by humans



- **Human ability to detect deception** is only slightly better than chance: typical accuracy rates are in the **55%-58%** range, with a mean accuracy of 54% over 1,000 participants in over 100 experiments
- **Bandwagon effect**: peer pressure can also at times “control” our perception and behaviour
- **Validity effect**: individuals tend to **trust fake news after repeated exposures**
- **Confirmation bias**: it confirms their preexisting beliefs
- **Desirability bias**: it pleases them

Rubin V. L. (2010). **On deception and deception detection: Content analysis of computer-mediated stated beliefs**. Proc. of the Association for Information Science and Technology 47, 1, 1–10

Leibenstein H. (1950). **Bandwagon, snob, and Veblen effects in the theory of consumers' demand**. The quarterly journal of economics 64(2):183–207

Ranking of fake news believers

>> IT WEBINARS: Transforma hoy tu TI para el mañana ¡Suscríbete a nuestro newsletter!

Los españoles son los europeos que más se creen las Fake News

Actualidad 11 SEP 2018



El 57% de los españoles admite haber creído alguna vez como verdadera la información de una noticia falsa. Esta es una de las principales conclusiones de un estudio de Ipsos Global Advisor en el que además se asegura que el **62% de la población española afirma que los españoles sólo buscan información en aquellas fuentes que piensan de forma similar a ellos.**

COMPARTIR

[Compartir](#)

[Twittear](#)

[Compartir](#)

El 57% sitúa a los españoles en quinto puesto del ranking mundial, por detrás de brasileños (67%), sauditas (58%), surcoreanos (58%) y peruanos (57%). Los españoles se consolidan además como los europeos que más han caído en las trampas de las noticias falsas, por delante de suecos (55%), polacos (55%), belgas (45%), alemanes (43%), franceses (43%), británicos (33%) o italianos (29%).

Fake or not fake? That's the question

Man arrested for calling directory assistance 2,600 times

The money saved from leaving the EU will result in the NHS getting
£ 350 million a week

Man tries to smuggle turtle onto plane by hiding it in a hamburger

Fake or not fake? That's the question

Man arrested for calling directory assistance 2,600 times

The money saved f

£ 350 million a wee



EU will result in the NHS getting

Man tries to smuggle turtle onto plane by hiding it in a hamburger

Fake or not fake? That's the question

Japan Today
Wed, 30 Jan 2008 15:41 UTC

A 37-year-old man has been arrested on suspicion of making about 2,600 prank calls to Nippon Telegraph and Telephone Corp's phone number directory service, obstructing its business, police said Wednesday. Takahiro Fujinuma, an unemployed man from Ota Ward, Tokyo, admitted to the allegations. He was quoted as telling the police, "Being single, I did it as a distraction from my loneliness."

According to the investigation, Fujinuma made a total of about 2,600 crank calls between June 1, 2007 and Nov 2007, interfering with operators' duties. The 104 directory service is provided by NTT Solco Corp. The operators dubbed him the "don't hang up on me" man because he often pleaded with them to stay on the line and talk, not even asking them to find a number. Police also said he dialed the service because he knew he would not be charged unless a number was found and given to him.

The Telegraph

[HOME](#) | [NEWS](#) | [SPORT](#) | [BUSINESS](#)

Travel | News

[Destinations](#) | [Hotels](#) | [Offers](#) | [Holiday types](#) | [City](#) | [Beach](#) | [Tours](#) |

 > Travel > News

Man tried to smuggle turtle on plane disguised as KFC burger

Fake or not fake? That's the question



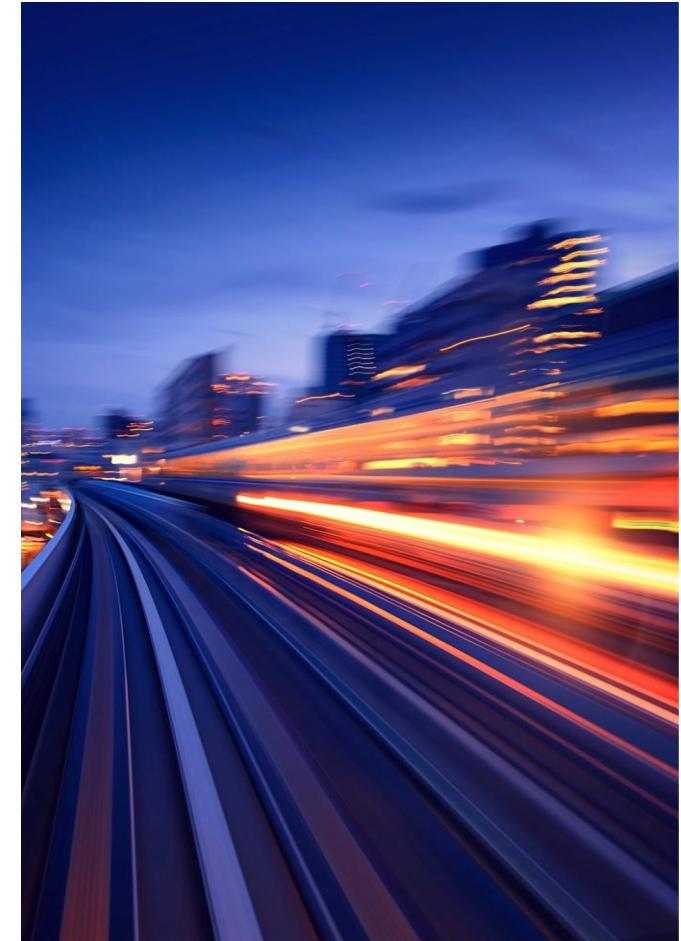
Man tries to smuggle turtle onto plane by hiding it in a hamburger

Disinformation detection should consider also:

- 1. Emotional signals and psycholinguistic characteristics**
2. Multimodal information
3. Disinformation campaigns and conspiracy theories

False info is faster and triggers different emotions

- Compared to the truth, **false info on Twitter is typically retweeted by many more users and spreads far more rapidly**, especially for political news
- **Fear, disgust, and surprise** (false rumours) vs joy, sadness, and anticipation (true rumours)



Fake or not fake? That's the question

The money saved from leaving the EU will result in the NHS getting
£ 350 million a week



disgust and surprise?

False info on Twitter



Replies to [@Gordon_Taylor @gurnull and @gurnull](#)

Did you vote for Hillary even though she is the criminal mastermind behind a child sex ring?

False info on Twitter



Replies to [@Gwen_Leigh @GwenLeigh and @GwenLeigh](#)

Did you vote for Hillary even though she is the criminal mastermind behind a child sex ring?

Fear, disgust and surprise?

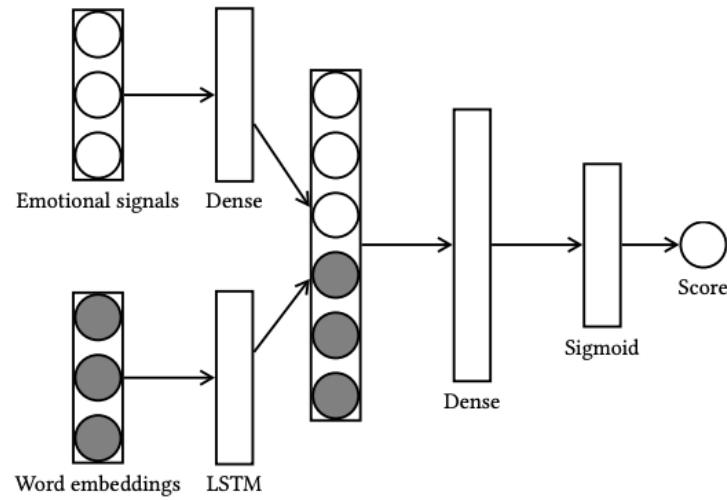
Information credibility on Twitter

- **emoCred**, a Long Short-Term Memory (**LSTM**) based system that leverages **emotional signals** for credibility detection
- Data creation: claims from Politifact

Giachanou A., Rosso P., Crestani F. (2019). **Leveraging Emotional Signals for Credibility Detection**. In Proc. of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19), 877-880

Giachanou A., Rosso P., Crestani F. (2021) **The Impact of Emotional Signals on Credibility Assessment**. In: Journal of the Association for Information Science and Technology (JASIST), 72(9):1117-1132

Information credibility on Twitter



Three different approaches for calculating the emotional signals of the claims:

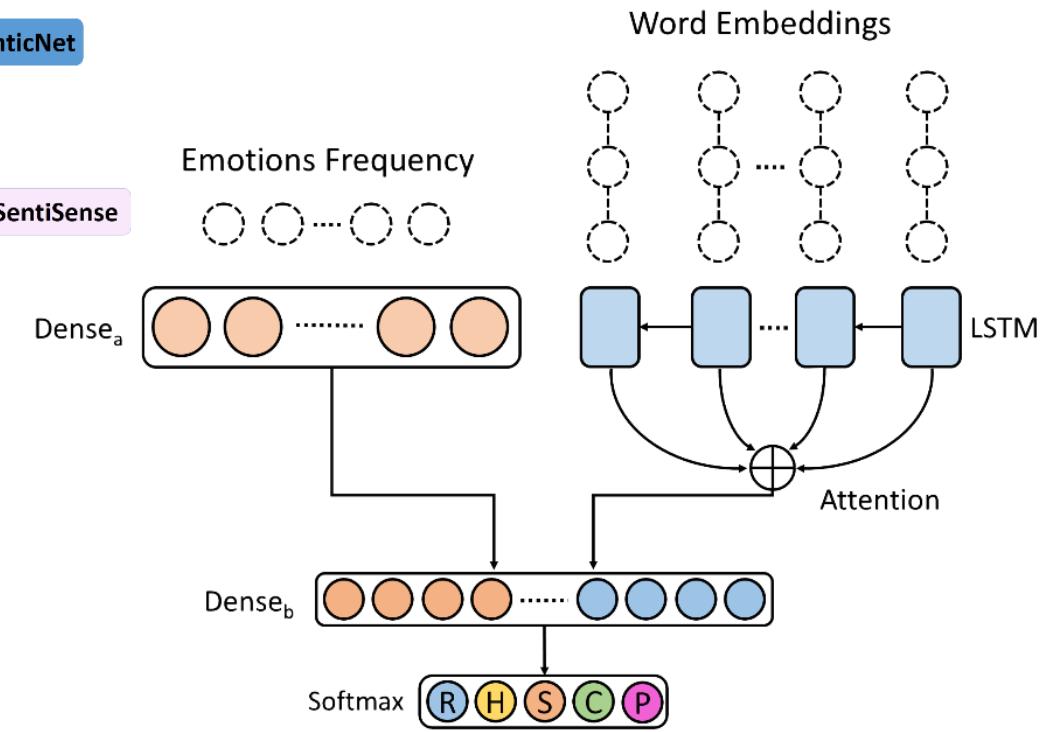
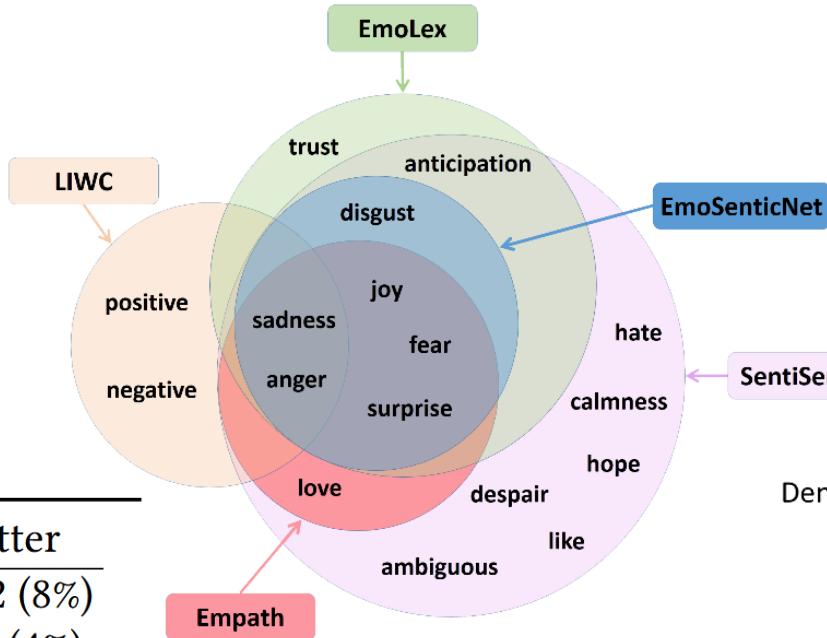
- **emoLexi**
- **emoInt**
- **emoReact**: LSTM network that predicts for each claim the probability to trigger any of the three intensity levels (low, average, high) for each of the five reactions love, joy, surprise, sadness and anger

Information credibility on Twitter

Dataset	Method	Accuracy	F1-score
Politifact-1	LSTM-text	0.551	0.549
	EmoCred-emoLexi	0.608	0.602*
	EmoCred-emoInt	0.604	0.602*
	EmoCred-emoReact	0.617	0.617*
Politifact-2	LSTM-text	0.597	0.567
	EmoCred-emoLexi	0.621	0.606*
	EmoCred-emoInt	0.628	0.586
	EmoCred-emoReact	0.619	0.601*

Emotional analysis of false info on Twitter and news

Category	News Articles	Twitter
Satire	5,750 (18%)	12,502 (8%)
Hoax	5,750 (18%)	6,247 (4%)
Propaganda	5,750 (18%)	66,225 (43.5%)
Clickbait	5,750 (18%)	36,103 (23.5%)
Real News	8,550 (28%)	30,949 (21%)
Total	31,550	152,026



Emotional analysis

	Accuracy	Macro-Precision	Macro-Recall	Macro-F1
News Articles				
BOW+SVM	73.67	71.81	71.11	70.70
W2V+LR	72.11	69.97	70.28	69.78
LSTM	74.79	79.69	70.85	72.26
EIN	80.72	79.52	79.82	79.43
Twitter				
BOW+SVM	62.86	59.53	55.94	57.45
W2V+LR	52.71	48.58	35.10	36.43
LSTM	63.29	64.44	52.00	55.41
EIN	64.82	60.65	58.90	59.70

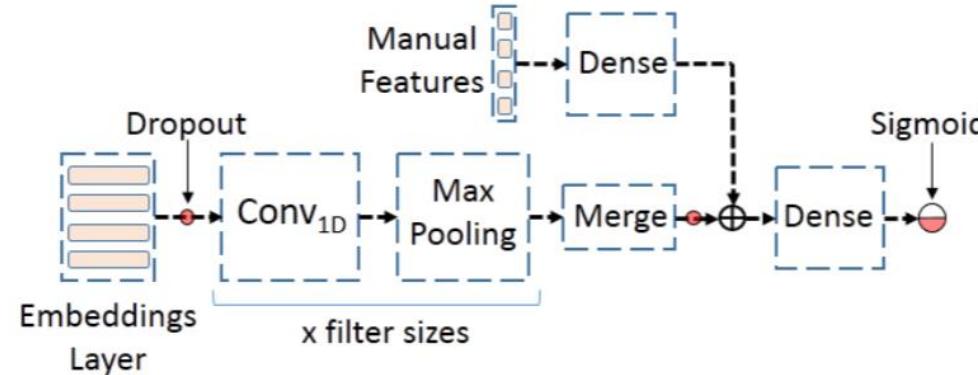
How emotions flow in disinformation

1 in a terrifying incident last night in chicago a cop accidentally killed
2 a man when trying to kill a flying cockroach local authorities report
3 the man tyrone smith was rushed to the hospital but couldn't
4 make it in time the cop who shot at him was
5 arrested everyone is strong and powerful until the cockroach can fly
6 the police officer mike doors said i needed something to defend
7 myself so i shot at it the fact that it killed
8 a man specifically a black one is a mere coincidence he
9 confirmed the cockroach lived but it's nowhere to be found local
10 authorities will launch a full investigation to determine what actually happened



Profiling mis/disinformation spreaders vs fact checkers

- CheckerOrSpreader is based on a Convolutional Neural Network (**CNN**)
- CheckerOrSpreader consists of two different components:
 - The textual one that refers to the word embeddings
 - The user's psycho-linguistic component (**linguistic patterns and personality scores**)



Giachanou A., Ríssola E., Ghanem B., Crestani F., Rosso P. (2020). **The Role of Personality and Linguistic Patterns in Discriminating between Fake News Spreaders and Fact Checkers**. In Int. Conf. on Applications of Natural Language to Information Systems, NLDB-2020, pp. 181-192, Springer.

Giachanou A., Ghanem B., Ríssola E., Rosso P., Crestani F., Oberski D. (2022) **The Impact of Psycholinguistic Patterns in Discriminating between Fake News Spreaders and Fact Checkers**. In: Data & Knowledge Engineering, Vol. 138

Linguistics patterns and personality scores

- **Linguistic patterns:** LIWC a software for mapping text to 73 psychologically-meaningful linguistic categories
 - pronouns (I, we, you, she/he, they)
 - personal concerns (work, leisure, home, money, religion, death)
 - time focus (past, present, future)
 - cognitive processes (causation, discrepancy, tentative, certainty)
 - informal language (swear, assent, nonfluencies, fillers)
 - affective processes (anxiety)
- **Personality scores with Big Five**
 - Openness to experience (unconventional, insightful, imaginative)
 - Conscientiousness (organised, self-disciplined, ordered)
 - Extraversion (cheerful, sociable, assertive)
 - Agreeableness (cooperative, friendly, empathetic)
 - Neuroticism (anxious, sad, insecure)

	F1-score
SVM+BoW	0.48
USE	0.53
LR+emotion	0.45
LR+sentiment	0.44
LR+LIWC	0.50
LR+personality	0.44
LSTM	0.44
CNN	0.54
CNN+LIWC	0.48
CNN+personality	0.57
CheckerOrSpreader	0.59

Misinformation spreaders and political bias

- **FACTOID dataset** (English) 4K users with 3.4M **Reddit** posts
- **Monitoring political discussion** in Reddit (since beginning 2020): long-term context of users' historical posts and the interactions between them
- **Fine-grained credibility level** (factuality of the news sources they shared: very low to very high) and **political bias strength** (extreme right to extreme left)
- Identifying misinformation spreaders by utilizing the **social connections between the users** along with their **psycholinguistic features**
- Affective mental processes correlates negatively with right-biased users: **right-biased** users tend to **express fewer emotions such as anxiety, anger and sadness**
- **Openness** to experience factor is lower **for fake news spreaders**

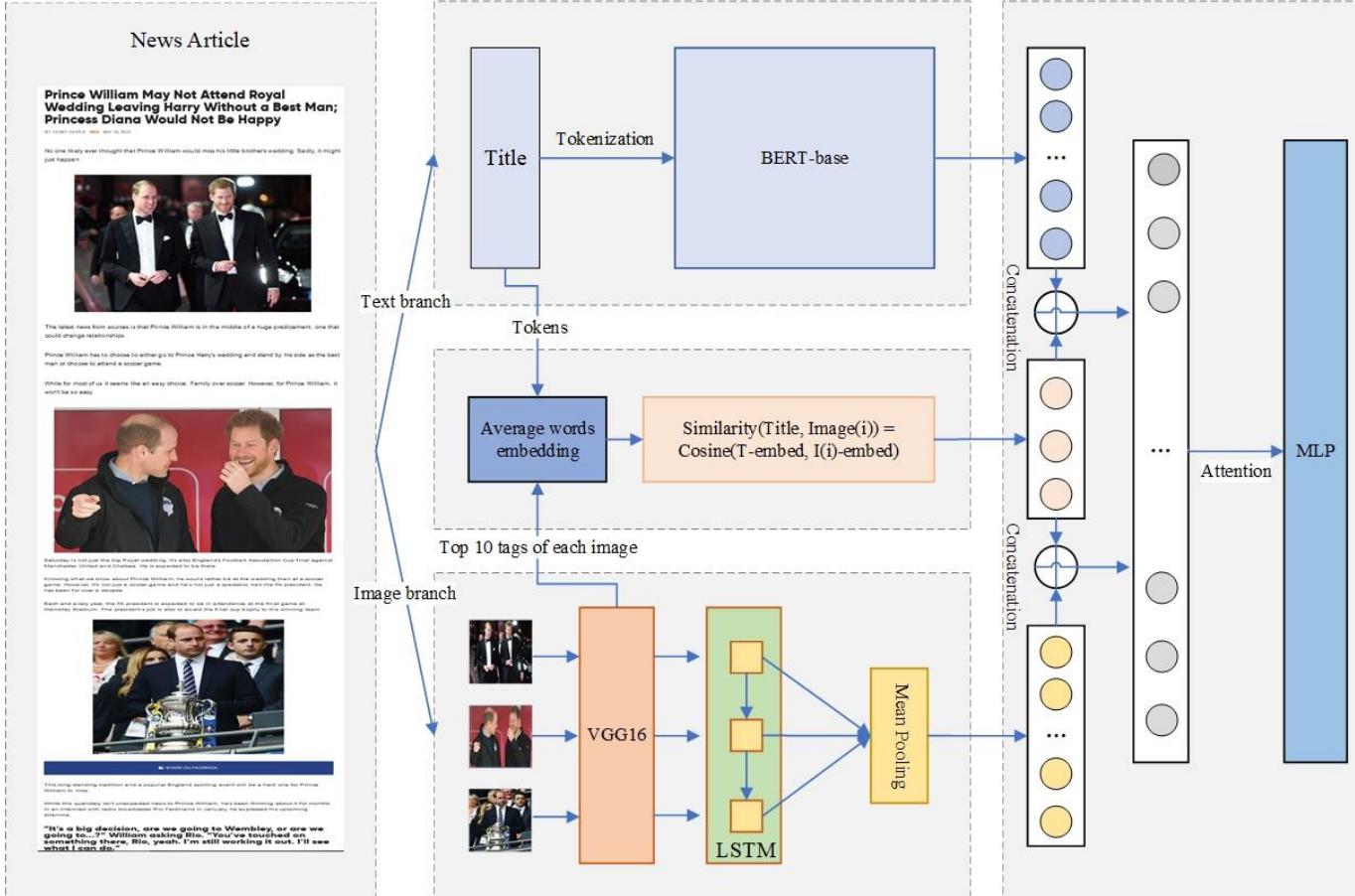
Disinformation detection should consider also:

1. Emotional signals and psycholinguistic characteristics
2. **Multimodal information**
3. Disinformation campaigns and conspiracy theories

Multimodal (multi) image fake news detection

- Articles come with more than one image
- Combine visual information from more than one images
- GossipCop: part of the FakeNewsNet collection
 - 2,745 fake news posts and 2,714 real news posts that contain **at least one image** after cleaning out the logo and icon images

Architecture



Results

	F1-score
BERT	0.7628* †
SpotFake	0.7537* †
EANN-var	0.4979* †
1-image-vgg16	0.3678* †
3-image-vgg16-LSTM	0.6690* †
4-image-vgg16-LSTM	0.6656* †
5-image-vgg16-LSTM	0.6465* †
1-image-vgg16+BERT+fusion(attention)	0.7683*
3-image-vgg16-LSTM+BERT+fusion(attention)	0.7792*
3-image-vgg16-LSTM+BERT+fusion(concatenation)	0.7775
3-image-vgg16-LSTM+BERT+similarity+fusion(attention)	0.7955
3-image-vgg16-LSTM+BERT+similarity+fusion(concatenation)	0.7884

Modeling scene context information

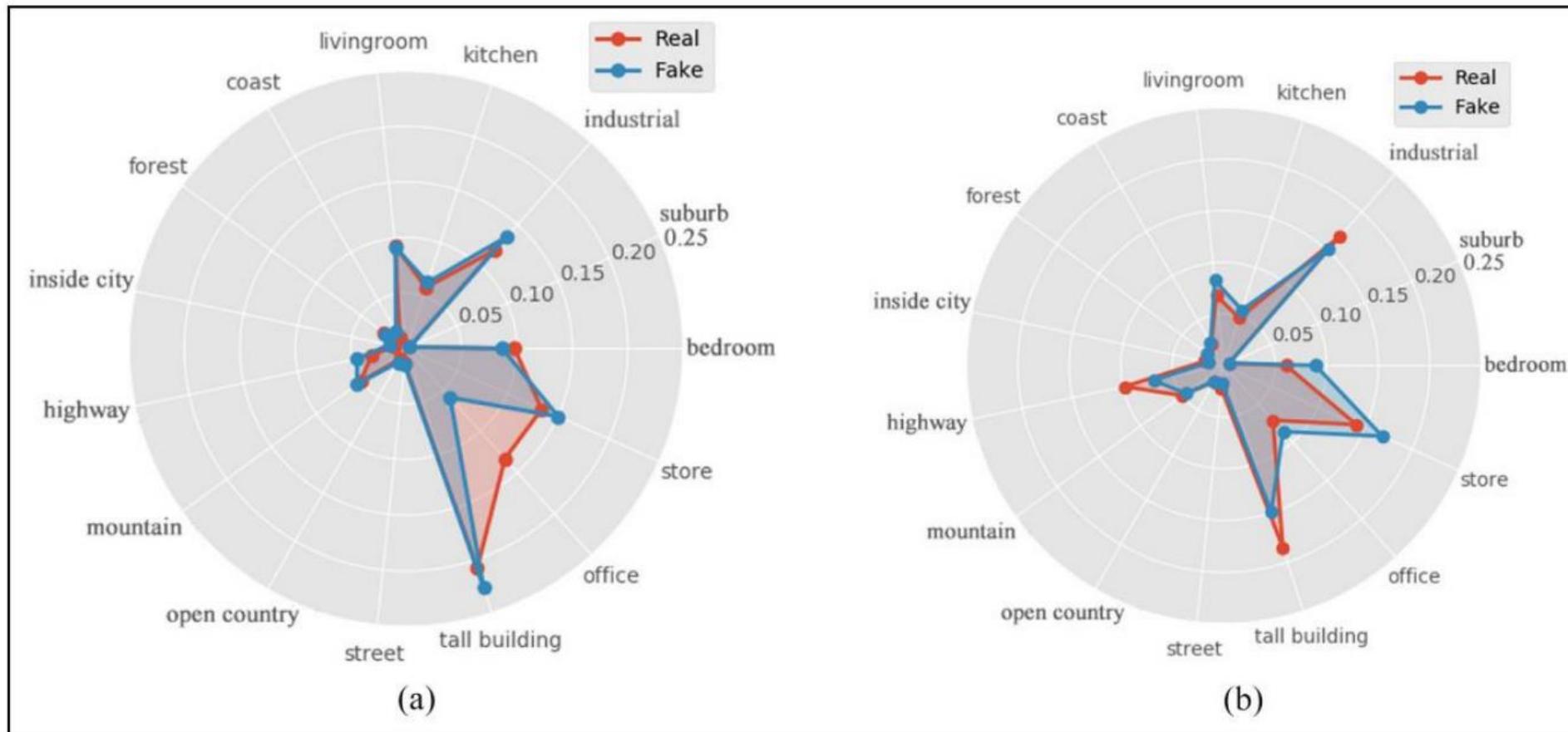


Figure 1. Average place scene scores for real and fake news of (a) PolitiFact and (b) GossipCop datasets.

Modeling scene context information

Place, weather, season

Table I. Scene co-occurrence ranking.

	Real	Fake
Rank		
1	PolitiFact	
2	Tall building, rain showers, winter	Store, rain showers, autumn
3	Tall building, rain showers, autumn	Industrial, rain showers, winter
Rank		
1	Industrial, rain showers, winter	Tall building, rain showers, autumn
2	GossipCop	
3	Tall building, rain showers, autumn	Store, rain showers, autumn
	Office, rain showers, autumn	Tall building, rain showers, spring
	Store, rain showers, autumn	

Disinformation detection should consider also:

1. Emotional signals and psycholinguistic characteristics
2. Multimodal information
3. **Disinformation campaigns and conspiracy theories**

Foreign Information Manipulation Interference

- European External Action Service (EEAS)
- Early detection of **Tactics, Techniques, and Procedures** (TTP)
- Indicators: **harmful, not illegal, manipulative, intentional, coordinated**
- Action plan for the European democracy by the EC (December 2020): aim is to **fight against disinformation**
- DISARM Framework: aim is the standardization of the model of the analysis of the **behaviours in disinformation attacks (detection, analysis, answers)**
- Threats of **manipulation of information and interference from foreign agents**
- Monitoring 616 channels (40% linked to Russian and Chinese media)
- Book by the **Spanish National Department of Security** on Disinformation campaigns (Chapter on AI to fight disinformation)

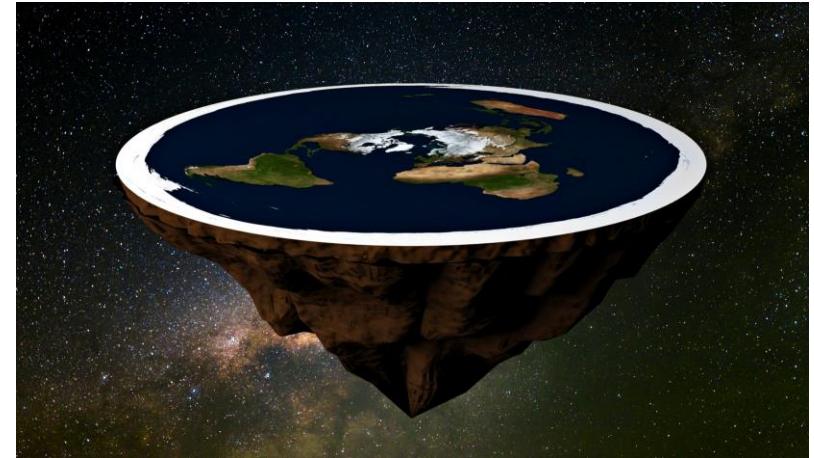
<https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:52020DC0790>

https://www.eeas.europa.eu/eeas/tackling-disinformation-foreign-information-manipulation-interference_en

https://www.eeas.europa.eu/eeas/1st-eeas-report-foreign-information-manipulation-and-interference-threats_en

Profiling conspiracy theories propagators

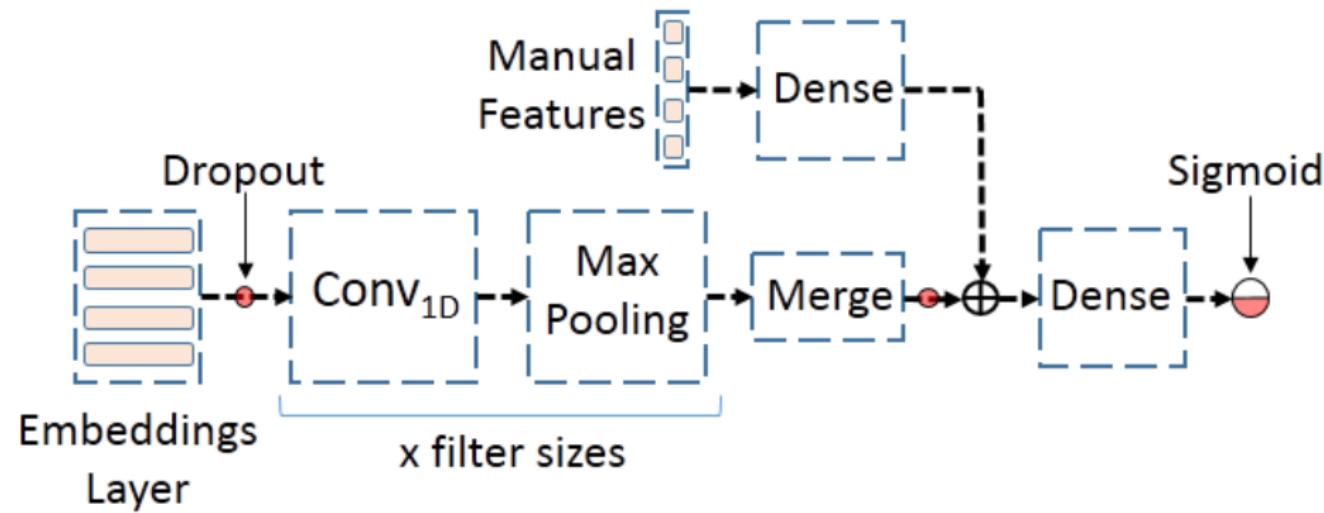
	pro-conspiracy	anti-conspiracy
Hashtags	#vaccinesCauseAutism #antiVax #climateChangeIsNotReal #flatEarth #nasaLies #nasaFake #spaceIsFake #moonLandingFake #bigPharmaFraud #ebolaconspiracy #antiFluoridation	#vaccinesWork #vaccinessavelives #climateChangIsReal #earthisnotflat #nasatruth #nasalsReal #spaceIsReal #moonlandingisreal
users	977	950
tweets	912,735	992,798



A psycho-linguistic analysis

- **Propagators have nine times less followers**
- Propagators less statuses, favorites and friends
- Verified users are more likely to refute conspiracies
- **Non propagators have old accounts**
- Non propagators have a larger number of statuses compared to propagators
- **Propagators tend to use more swear words**
- Non propagators exhibit higher usage of work, leisure, money, home, and death than **propagators** who concern more about **religion**
- Non propagators exhibit a higher usage in causation (because, effect, hence) compared to propagators

ConspiDetector



ConspiDetector

	Precision	Recall	F1
USE	0.70	0.69	0.69
CNN	0.68	0.82	0.68
CNN + Profile	0.61	0.78	0.58
CNN + Personality	0.75	0.79	0.73
CNN + LIWC	0.73	0.78	0.71
CNN + Sentiment	0.67	0.76	0.66
CNN + Emotion	0.77	0.58	0.67
ConspiDetector (Psycho-linguistic)	0.77	0.76	0.74
CNN + Psycho-linguistic + Profile	0.72	0.70	0.68

- The most effective feature is the IBM Personality Insights with a performance of 0.73
- The lowest performance is achieved with the profile characteristics (CNN + Profile), lower than the CNN baseline
- Profile, sentiment and emotion are not helpful

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

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

Concerns about COVID-19 on Twitter

- Shared task at FIRE 2023, the Forum of Information Retrieval Evaluation
- Aim: build an effective **multi-label classifier** to label a tweet according to the **specific concern(s) towards vaccines** as expressed by the author of the tweet
- **Concerns** (classes): Unnecessary, Mandatory, Pharma, **Conspiracy**, Political, Country, Rushed, Ingredients, Side-effect, Ineffective, Religious, None

Taxonomies on conspiracy theories (I)

- **Focus:** structure of **conspirative narrative** (**Aame Thomson Uther index [1]**)
- **Idea:** identifying narrative detecting common triplets (**actor, action, target**)
- **Labels:** actors and targets
- **Drawbacks:**
 - **targets** could be different things (**victims, tools, outcomes, events**)
 - **Reddit corpus not annotated by humans** (analysis of over ten years of discussions in **r/conspiracy**, an online community on Reddit dedicated to conspiratorial discussions)

Samory M., Mitra T. (2018). **The Government Spies Using Our Webcams: The Language of Conspiracy Theories in Online Discussions.**
Proc. of the ACM on Human-Computer Interaction

[1][https://encyclopedia.pub/entry/37990#:~:text=The%20Aarne%E2%80%93Thompson%20classification%20systems,system%20\(developed%20in%202004%20and](https://encyclopedia.pub/entry/37990#:~:text=The%20Aarne%E2%80%93Thompson%20classification%20systems,system%20(developed%20in%202004%20and)

Taxonomies on conspiracy theories (II)

- **Aim:** distinguish between **conspiracy theory** vs **conspirational thinking**
(conspirational thinking should not be though as pathological)
- **Labels:** **actor, action, consequence, target, event, goal**
- **Conspiracy theory:** the first four categories + event, goal (optional)
- **Conspirational thinking:** one or more of the six labels but **not all the first four**
- **Drawbacks:**
 - in actor no distinction between agent and facilitator; and in target no distintion between victim and campaigner
 - focus on **pathological thinking** and **not on critical thinking**

Introne J., Korsunska A., Krsova L., Zhang Z. (2020).

Mapping the Narrative Ecosystem of Conspiracy Theories in Online Anti-vaccination Discussions.

Proc. Int. Conf. on Social Media and Society (SMSociety'20)

Taxonomies on conspiracy theories (III)

- **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 (BERTa Chulvi)
- “Us vs them” narrative (label for us: campaigners label)
- Insiders include campaigners and victims
- Outsiders include agents and facilitators
- Categories at span level
- Domain-agnostic: it could be applied to other conspiracy theories

Taxonomy: 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 (Spanish, English)
- Oppositional non-mainstream views on the **COVID-19 pandemic**
- Shared task at  **PAN** 2024



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

Stance and ideological orientation for promoting critical thinking



NGI.eu Next Generation Internet



A pilot AI tool



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



BERTa Chulvi



Ivan Grubišić



Andrea Simeri



Paolo Rosso



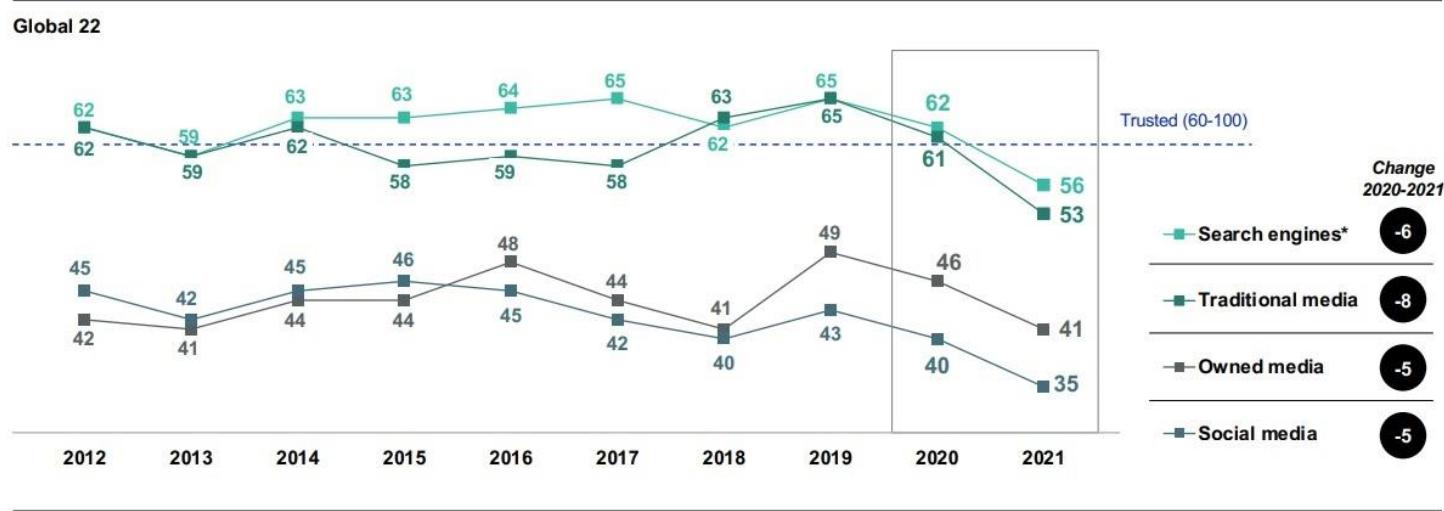
ATHENS TECHNOLOGY CENTER



The problem

TRUST IN ALL INFORMATION SOURCES AT RECORD LOWS

Percent trust in each source for general news and information



2021 Edelman Trust Barometer. COM_MCL. When looking for general news and information, how much would you trust each type of source for general news and information? 9-point scale; top 4 box, trust. Question asked of half of the sample. General population, 22-mkt avg.

*From 2012-2015, "Online Search Engines" were included as a media type. In 2016, this was changed to "Search Engines."



Edelman Trust
Barometer 2021
(survey of 33,000
people 18+ across
28 countries)

The problem

NEWS ORGANIZATIONS SEEN AS BIASED

Percent who agree

Journalists and reporters are **purposely trying to mislead** people by saying things they know are false or gross exaggerations

Global 27

59%

Most **news organizations** are more concerned with **supporting an ideology** or political position than with informing the public

Global 27

59%

The media is **not** doing well at **being objective** and non-partisan

↑

Global 24

61%

Strongest agreement that media is not doing well in:

Japan	80
S. Korea	77
Colombia	76
Argentina	75
Italy	75
Spain	73
Brazil	72
UK	69
*Nigeria	67
Mexico	66

SPAIN 73%

2021 Edelman Trust Barometer. POP_EMO. Some people say they worry about many things while others say they have few concerns. We are interested in what you worry about. Specifically, how much do you worry about each of the following? 9-point scale; top 4 box, worry. Attributes shown to half of the sample. ATT_MED_AGR. Below is a list of statements. For each one, please rate how much you agree or disagree with that statement. 9-point scale; top 4 box, agree. Question asked of half of the sample. General population, 27-mkt avg. PER_MED. How well do you feel the media is currently doing each of the following? Please indicate your answer using the 5-point scale below. 5-point scale; bottom 3 box, not doing well. Question asked of half of the sample. General population, 24-mkt avg. Data not collected in China, Russia, and Thailand.

*Nigeria not included in the global average

Related work from communication science

AllSides™
Don't be fooled by media bias & misinformation.

NEWS MEDIA BIAS PERSPECTIVES TOPICS TALKS SCHOOLS ABOUT

Donate Join Search

TOP STORIES: COVID-19 | Omicron Variant | Vaccine Mandates | Capitol Riot | Economy | Fact Checks



CDC Clarifies the 'Record Number' of Children Hospitalized with COVID-19 Amid Rise in Cases

A record number of 672 children in the US were hospitalized with COVID-19 this week, according to CDC data.

The previous record of daily hospital admissions for the 0-17 age group...

From the Center	From the Right	From the Left
A record number of children are hospitalized with COVID-19 Insider L C R R	CDC Announces Rise In Kids Hospitalized With COVID – But Data Includes Kids... The Daily Wire L C R R	The CDC Is Warning That Child Hospitalization Rates Are Breaking Pandemic Records BuzzFeed News L C R R

An Editorial Review Team— which includes people from across the political spectrum — reviews the works of a source and comes to a general consensus on its bias. Based on US politics, the web offers a balanced feed of news across the political spectrum that is manually elaborated by professional journalists in selected topics. <https://www.allsides.com/>

Related work from computational linguistics



Bali R. et al. (2020) **We Can Detect Your Bias: Predicting the Political Ideology of News Articles.**

In Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing EMNLP '20, pp. 4982–4991

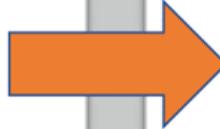
How we approach the problem

Bias Detection Paradigm

- Manipulation vs truth
- Biased media vs Unbiased media
- Focus on source and content

Critical Thinking Paradigm

- Uniformity vs plurality
- Information hygiene practices
- Focus on interaction among *source-content-person*



A tool combining ideological orientation and stance detection

- Combine **Stance Detection** and **Information about the ideological orientation of the source**, with the goal of **showing plurality and combat the polarized vision of the media landscape**.
- 3 controversial topics: **immigration**, **climate change** and **COVID vaccination**, and to create a pilot dataset of 15 newspapers from Spain and from UK that cover all the political spectrum



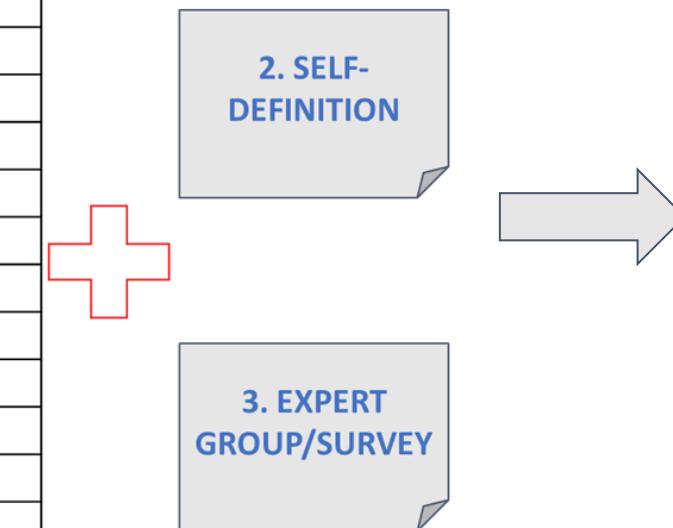
Ideological orientation of the media

15 newspapers from Spain and from UK that cover all the political spectrum: first we select the media and we classify them using three methods

Name	Media Bias Chart/mediabiasfactcheck	Name	mediabiasfactcheck
EL PAÍS	Left Center	The Guardian (LC)	Left-center
EL MUNDO	Rigth Center	Daily Mirror (LC)	Left-center
ABC	Right Center	Independent (LC)	Left-center
LA RAZÓN	Right	Manchester Evening News (LC)	Left-center
Expansión	No data about ideological bias	Express	Right
El Confidencial	Right	The Sun (R)	Right
El Periódico	Balanced Bias	BBC (Least biased)	Left-center
La Vanguardia	Balanced Bias	PinkNews (L)	Left
Huffington Post	Left	The Telegraph (R)	Center
Eldiario.es	Left	Metro (LC)	Left-center
Voz pópuli	Right	Daily Mail (R)	Right
Libertad digital	Right	Evening Standard (RC)	Right-Center
Infolibre	Left	Belfast Telegraph	No data about ideological bias
El independiente	No data about ideological bias	The New Statesman (L)	No data about ideological bias
El salto	No data about ideological bias	The Herald (Scotland)	Left

<https://political-watch- oa.d1llfzwprjpx3u.amplifyapp.com/>

<https://mediabiasfactcheck.co m/>

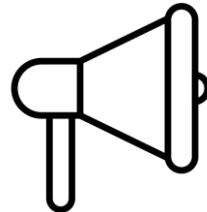
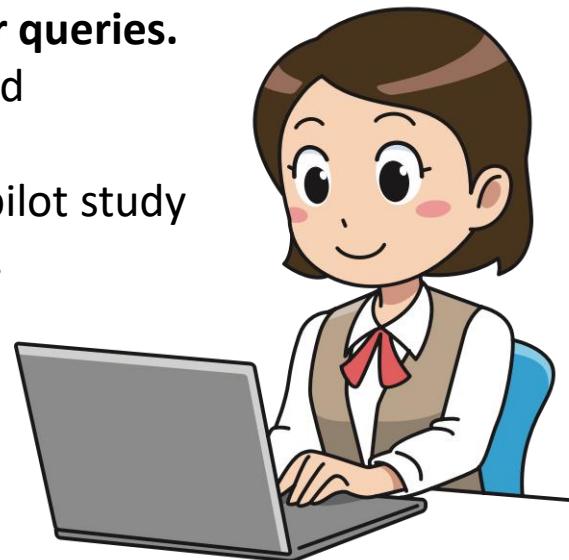


Final
classification
of the media

Collecting the data: pilot study with real users

Using a keyword-based approach, we found that many instances were not properly related to the topic.

Therefore, we decided to address **the creation of the pilot dataset with a second strategy based on user queries**. For this second approach, we developed a pilot study with six users.



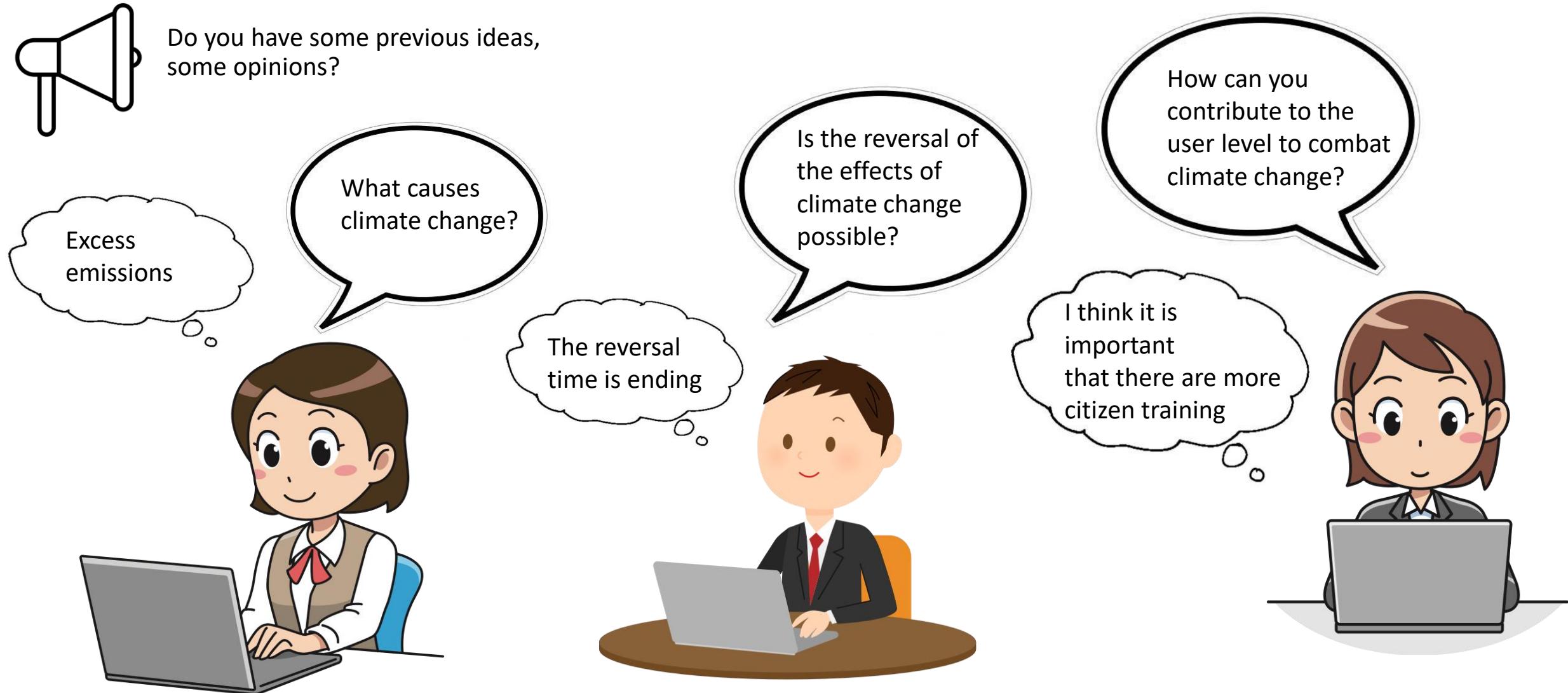
Please think of a question about climate change that you would ask a search engine like Google...



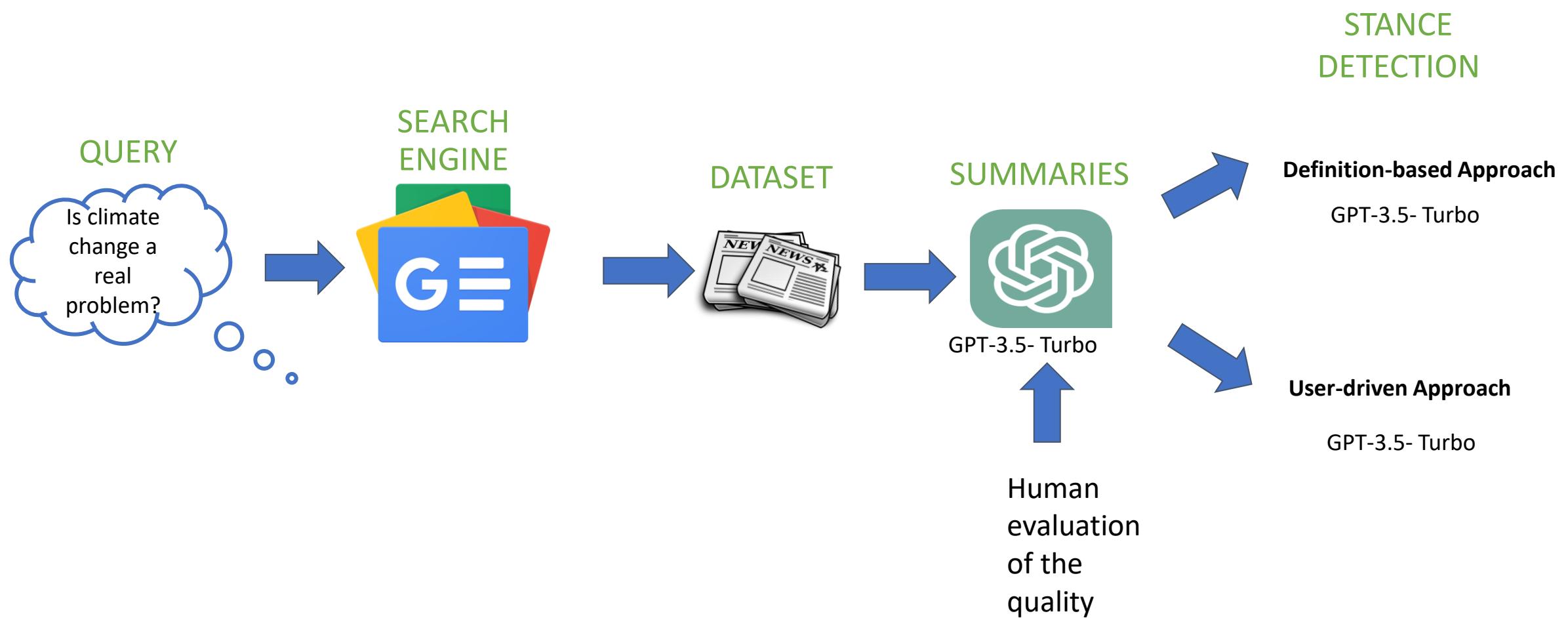
Is the reversal of the effects of climate change possible?



Collecting the data: pilot study with real users



Stance detection workflow



Definition-based approach for stance detection

The Definition-based stance detection approach is designed to provide an assessment of an article's attitude towards a selected topic relative to the general population's view. There are three possible responses to the article's stance:

- "The article is in favor w.r.t. the topic", if the article emphasizes the arguments in favor of the topic;
- "The article is against w.r.t. the topic", if the article emphasizes the arguments against the topic;
- "The article is neutral w.r.t. the topic", if the article does not emphasize either of these arguments.

This approach relies on predefined topic definitions and predefined arguments in favor and against the topic. It is based on **prompt engineering** and utilizes large language models (LLMs) as **zero-shot classifiers**.

GPT-3.5-Turbo assigned a score between -1 (against) and 1 (in favour) to each news

News media articles

UK cities must adapt to a changing climate

Selected article

The UK is facing the impacts of climate change, including heatwaves, flooding, and extreme weather events. While reducing carbon emissions is crucial, investing in resilient infrastructure is also necessary to adapt to the changing climate. Flooding and heavy rainfall, storms, heatwaves, and sea level rise are the main climate-related dangers in the UK. Cities are particularly at risk due to the lack of green spaces and the "urban heat island" effect caused by concrete and buildings. One in six properties in England is at risk of flooding, and weather events cost millions of pounds each year in damage and repairs. Pre-emptive adaptation measures are more cost-effective than retrofitting. Designing features such as sustainable urban drainage systems, shaded buildings, and green spaces can help mitigate the risks. The UK can learn from other countries' adaptation strategies, such as China's "sponge city" neighborhoods and London's Counter Creek Sewer Alleviation scheme.

Output: Stance

-1 = totally against, 0 = neutral, +1 = totally in favour

1



User-driven approach for stance detection

The user-driven stance detection approach, rooted in the **entailment task**: diverges from the Definition-based stance detection methodology by focusing on contextual nuances in user-generated content.

Unlike the former, which aims to gauge the general population's stance on a chosen topic, the user-driven approach centers around an actual user query and their previous personal perspective (i.e., the stance) on the matter.

The process begins with a **user query and with a newspaper article relevant to the query**. The user expresses their personal stance, indicating their initial perspective on the query. Subsequently, a Large Language Model (GPT-3.5-Turbo), is employed to **generate a sentence that combines the query and the user's stance leveraging on prompt engineering**.

To explore the opposing viewpoint, we utilize the LLM to generate another sentence that represents the negation of the user text hereinafter referred to as the opposite user text. This sentence encapsulates the opposite stance, providing a balanced perspective for further analysis. Finally, we employ a **Transformer model designed for entailment tasks: BART**. This model takes as input the user text, the opposite user text, and the text of the retrieved newspaper article, in order to **assign scores indicating the likelihood of entailment for both perspectives**.

Excess emissions cause climate change.



Excess emissions cause climate change.

98%

The excess of emissions does not have any correlation with climate change.

2%

TrustSearch AI pilot tool



- 1 Select a language
- 2 Select a topic
CLIMATE CHANGE COVID VACCINATION IMMIGRATION
- 3 Select a query
How can we contribute to combat climate change as citizens? ▾
- 4 Select a stance

No opinion

You can contribute to the user level to combat climate change by supporting the implementation of citizen training programs.

At user level we can do almost nothing because major changes are required at the level of the economic system.

Climate change is not a problem, we don't need to fight it.

The prototype can be accessed through this link:
<https://trust-search.k8s.atc.gr/>

TrustSearch AI pilot tool



< BACK

LIST PLOT

Topic: Climate change

Query: How can we contribute to combat climate change as citizens?

Stance: No opinion

How do we characterize the source?

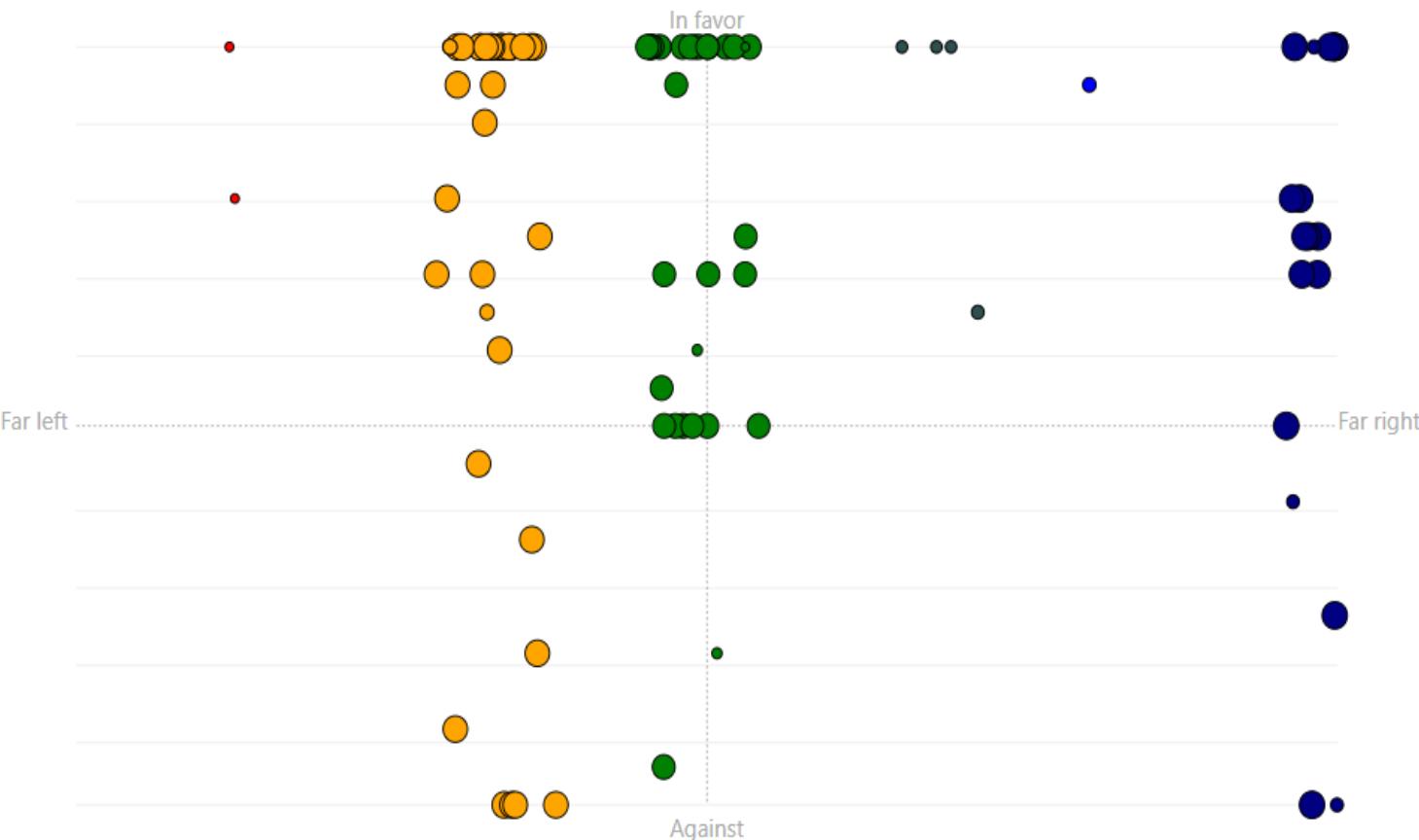
How do we calculate the stance?

Circle color represents media's political ideology

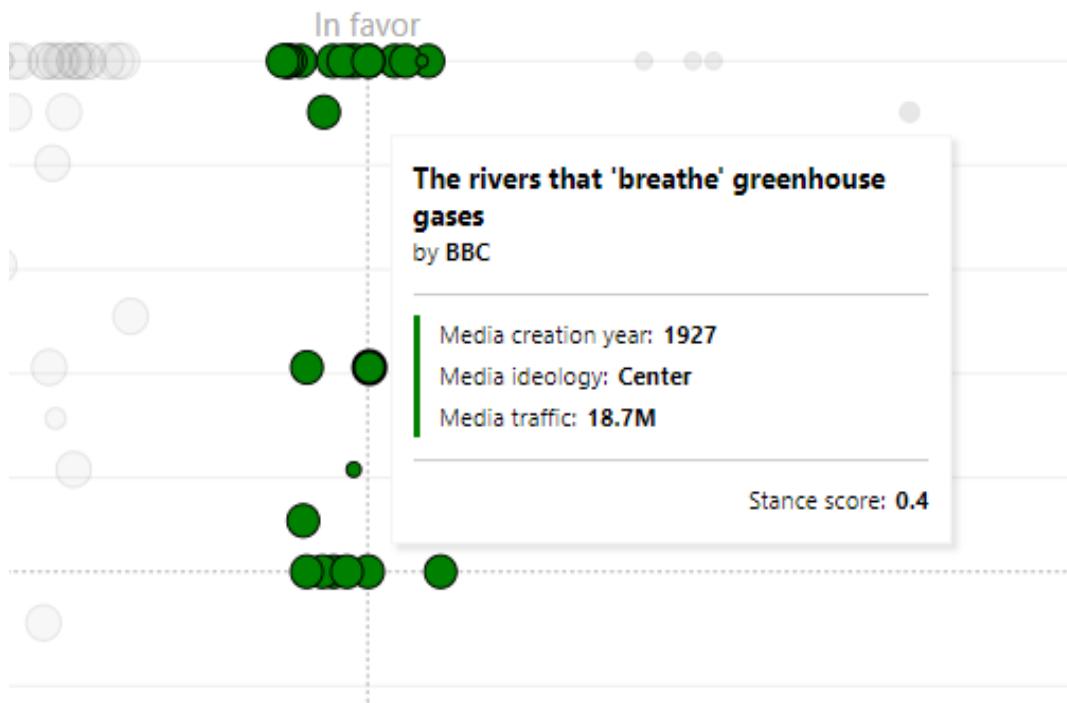
- Far left
- Left
- Left center
- Center
- Right center
- Right
- Far right

Circle size represents media's traffic

- Minimum
- Maximum



TrustSearch AI pilot tool



The screenshot shows the full BBC news article page for the same story. The headline is "The rivers that 'breathe' greenhouse gases" by BBC. The article text discusses the study findings. At the bottom right, there is a summary box with the stance score and ideology information.

Rivers are a significant source of greenhouse gases, with a study finding that rivers in Hong Kong's New Territories are releasing large quantities of carbon dioxide, methane, and nitrous oxide. The study revealed that the more polluted the river, the greater its emissions. Discharge from livestock farms, misconnections in old buildings, and unsewered premises were identified as the main reasons for pollution. The study also highlighted the impact of land use and land cover on greenhouse gas production in polluted river sites. It was found that when river water quality deteriorated, the emissions of CO₂, methane, and N₂O increased significantly. The article emphasizes that urbanization and human activities are contributing to higher greenhouse gas emissions from rivers, with more than half of the world's population living close to surface freshwater bodies. However, there is hope for reducing river emissions through river restoration and pollution reduction measures. The article suggests that improving water quality can make a significant difference, and various national and international programs have been implemented to restore and rehabilitate rivers. The study's authors plan to share their findings with relevant governments and NGOs to inform them about the benefits of improving water quality in rivers. The article concludes by highlighting the importance of reducing nutrient pollution to improve water quality and mitigate climate change. Overall, the study underscores the need to address water pollution to minimize the impact of rivers on global warming.

Stance score: 0.4
Ideology: Center

The screenshot shows the BBC news article page again, but with a sidebar on the right. The sidebar features a large image of a river with many small boats, and a thumbnail of the same news article with the headline "The rivers that 'breathe' greenhouse gases". Below the sidebar, there is a footer bar with links like "Home", "News", "Sport", "Business", etc., and a "Herramienta Recortes" link.

BBC

Home News Sport Business Innovation Culture Travel Earth Video Live

The rivers that 'breathe' greenhouse gases

24 March 2021
By Matthew Keegan, Features correspondent

Share

The rivers that 'breathe' greenhouse gases
By Matthew Keegan, Features correspondent

Herramienta Recortes

TrustSearch AI pilot tool

< BACK

LIST PLOT

Topic: Climate change

Query: How can we contribute to combat climate change as citizens?

Stance: You can contribute to the user level to combat climate change by supporting the implementation of citizen training programs.

[How do we characterize the source?](#)

[How do we calculate the stance?](#)

Circle color represents media's political ideology

- Far left
- Left
- Left center
- Center
- Right center
- Right
- Far right

'Vulnerable to climate change in every way': Mumbai eyes status as south Asia's first carbon neutral city

by The Telegraph

Mumbai, a city vulnerable to climate change due to its low-lying geography, is aiming to become the first carbon neutral city in South Asia by 2050. The city's survival is at stake, with 40% of it at risk of being underwater in the next 100 years if...

[Read more](#)

Stance score: **0.6792082786560059**

Climate change: Which vegan milk is best?

by BBC

The article discusses the increasing popularity of vegan foods, particularly plant-based milks, and their environmental impact. A study from the University of Oxford found that producing a glass of dairy milk results in almost three times the greenhouse...

[Read more](#)

Stance score: **0.6722351312637329**

How Seychelles ocean plants could help tackle climate change

by BBC

The Seychelles is leading the way in the fight against climate change by mapping its seagrasses, which are a significant carbon store. Seagrasses, along with other coastal wetlands, act as a barrier against rising ocean levels and extreme weather cau...

[Read more](#)

Stance score: **0.6448837518692017**

Swamp power: how the world's wetlands can help stop climate change | Greenhouse gas emissions

by The Guardian

The article discusses the potential of swamp farming, or paludiculture, in mitigating climate change and reducing carbon emissions. It highlights the benefits of cultivating wetlands, such as providing low-carbon energy, wildlife habitat, and water d...

[Read more](#)

Stance score: **0.6148849725723267**

UK must 'adapt or die' as climate change experts warn floods could kill hundreds

by The Mirror

The Environment Agency has issued a warning that Britain could face catastrophic floods similar to those in Germany, potentially resulting in hundreds of deaths. The agency's report emphasizes the need for the country to adapt to the inevitable effec...

[Read more](#)

Stance score: **0.6039271354675293**

City of Glasgow College: Let's come together for a higher purpose

by The Herald

The City of Glasgow College is gearing up for the 26th United Nations Climate Change Conference of the Parties (COP26) and is committed to living up to its visionary motto 'lead from the future'. With over 100 world leaders expected to attend, includ...

[Read more](#)

Stance score: **0.598459005355835**

GP tells patients their inhalers could causing global warming

by Daily Mail

Dr. Brett Montgomery, a senior lecturer in general practice at the University of Western Australia, advises his patients with asthma to use dry powder inhalers instead of metered-dose inhalers, which emit chemicals that damage the ozone and contribut...

[Read more](#)

Stance score: **0.5915987491607666**

The rivers that 'breathe' greenhouse gases

by BBC

Rivers are a significant source of greenhouse gases, with a study finding that rivers in Hong Kong's New Territories are releasing large quantities of carbon dioxide, methane, and nitrous oxide. The study revealed that the more polluted the river, th...

[Read more](#)

Stance score: **0.5891315937042236**

Miami's fight against rising seas

by BBC

Miami, Florida is facing a significant challenge due to rising sea levels. The region is experiencing regular flooding, with water in the streets becoming a common occurrence. The impact of rising seas is particularly severe in this area, which is co...

[Read more](#)

Hold the beef: McDonald's avoids the bold step it must take to cut emissions | Environment

by The Guardian

McDonald's has announced sustainability initiatives, but experts argue that the company is avoiding the necessary step to significantly reduce emissions: cutting beef production. The company's massive beef consumption leads to high greenhouse gas emi...

Rising sea levels mean these cities and even entire countries could disappear beneath the waves within decades

by The Sun

The article discusses the threat of rising sea levels due to global warming, which could lead to the disappearance of cities and even entire countries by the end of the century. The polar ice caps melting at an alarming rate has raised concerns about...

[Read more](#)

World leaders duped by manipulated global warming data

by Daily Mail

The Mail on Sunday has revealed evidence that the world's leading source of climate data, the National Oceanic and Atmospheric Administration (NOAA), rushed to publish a paper that exaggerated global warming. The article reported on claims made by Dr...

User study: Evaluation of the tool

RQ. Does the use of TrustSearch introduce significant differences compared to Google in terms of promoting critical thinking.

41 university students (20-26 age) in Spain participate in the study as part of a practical class in research methods. Answers were anonymous.



PHASE 1- Critical Thinking Disposition (CTD) (Initial evaluation)

- 1.1 Subjects answer *Critical Thinking Disposition Scale* (Bravo et al. 2021) (Measure 1)
- 1.2 Read a definition of Critical Thinking and measure how different sources promote critical thinking (Q1)

PHASE 2- Google experience and evaluation

- 2.1 They were asked to search in Google this query: "How can we contribute to combat climate change as citizens?" (5 mins for the examination of the news provided by Google)
- 2.2 Evaluation of how Google promote skills linked to critical thinking with Q2
- 2.3 General evaluation of the Google tool promoting critical thinking Q3

PHASE 3- TrustSearch experience and evaluation

- 3.1 Use of TrustSearch app (5 mins for the examination of the app)
- 3.2 Evaluation how TrustSearch promote skills linked to critical thinking with Q2
- 3.3 General evaluation of the TrustSearch tool promoting critical thinking Q3
- 3.4 Open question about the app

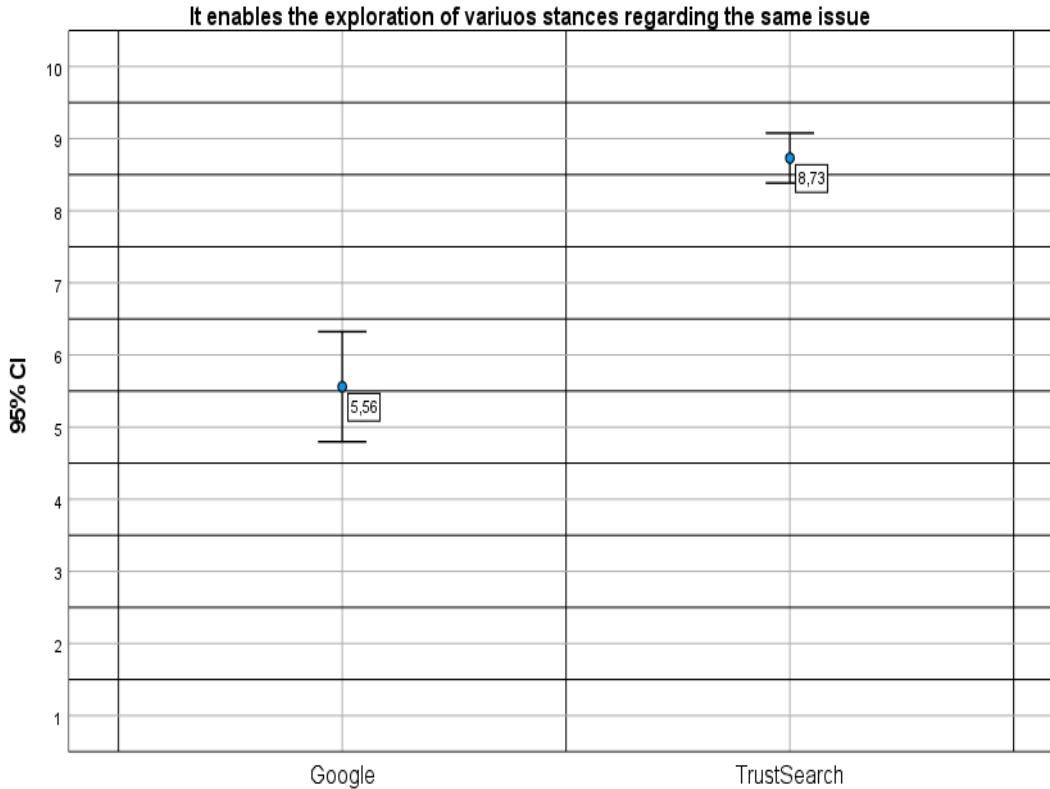
PHASE 4- Critical Thinking Disposition (CTD) (Last evaluation)

Questionnaire Q2

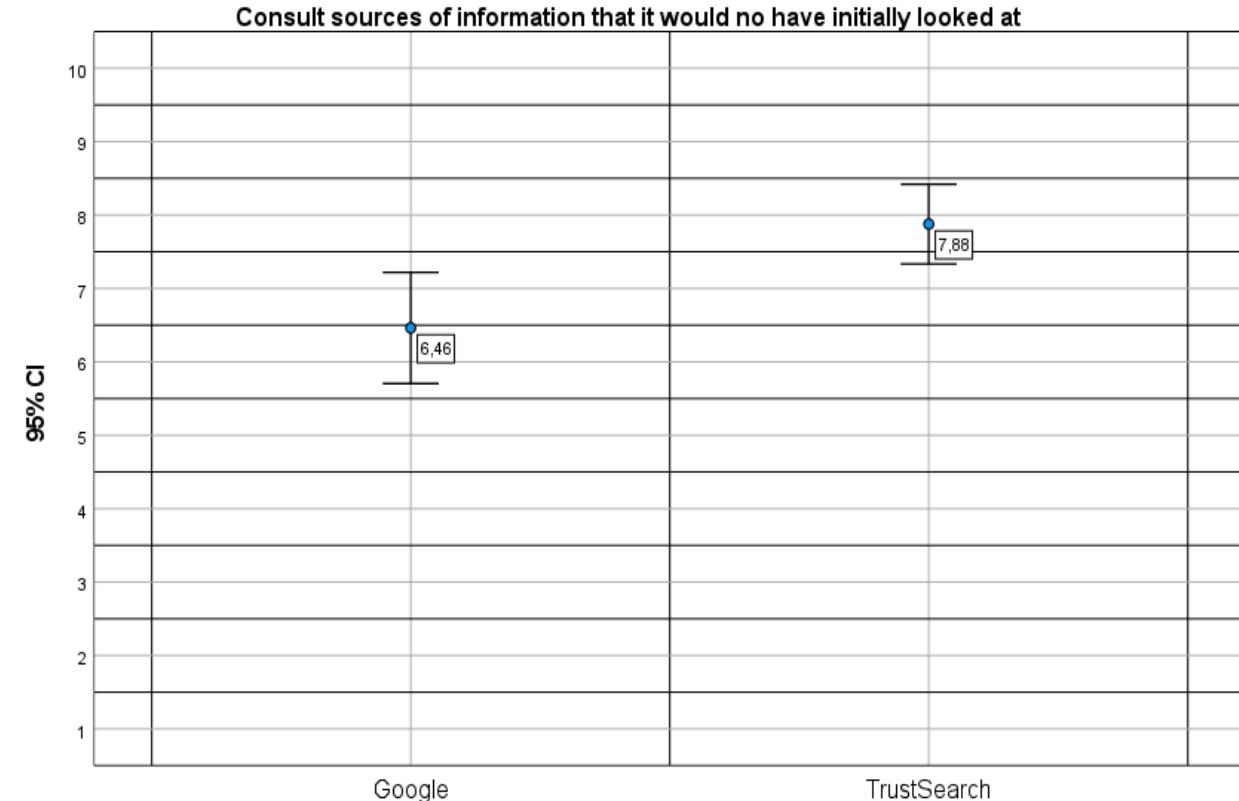
- It enables the exploration of various stances regarding the same issue (Q2-1)
- Show the differences between the media (Q2-2)
- Note how media of different political orientations coincide in their opinion (Q2-3)
- Awareness of diversity of opinions (Q2-4)
- Consult source of information that it would not have looked at (Q2-5)

User study: Evaluation of the tool

Mean values were compared for the two search engines with a Paired Sample T-Test



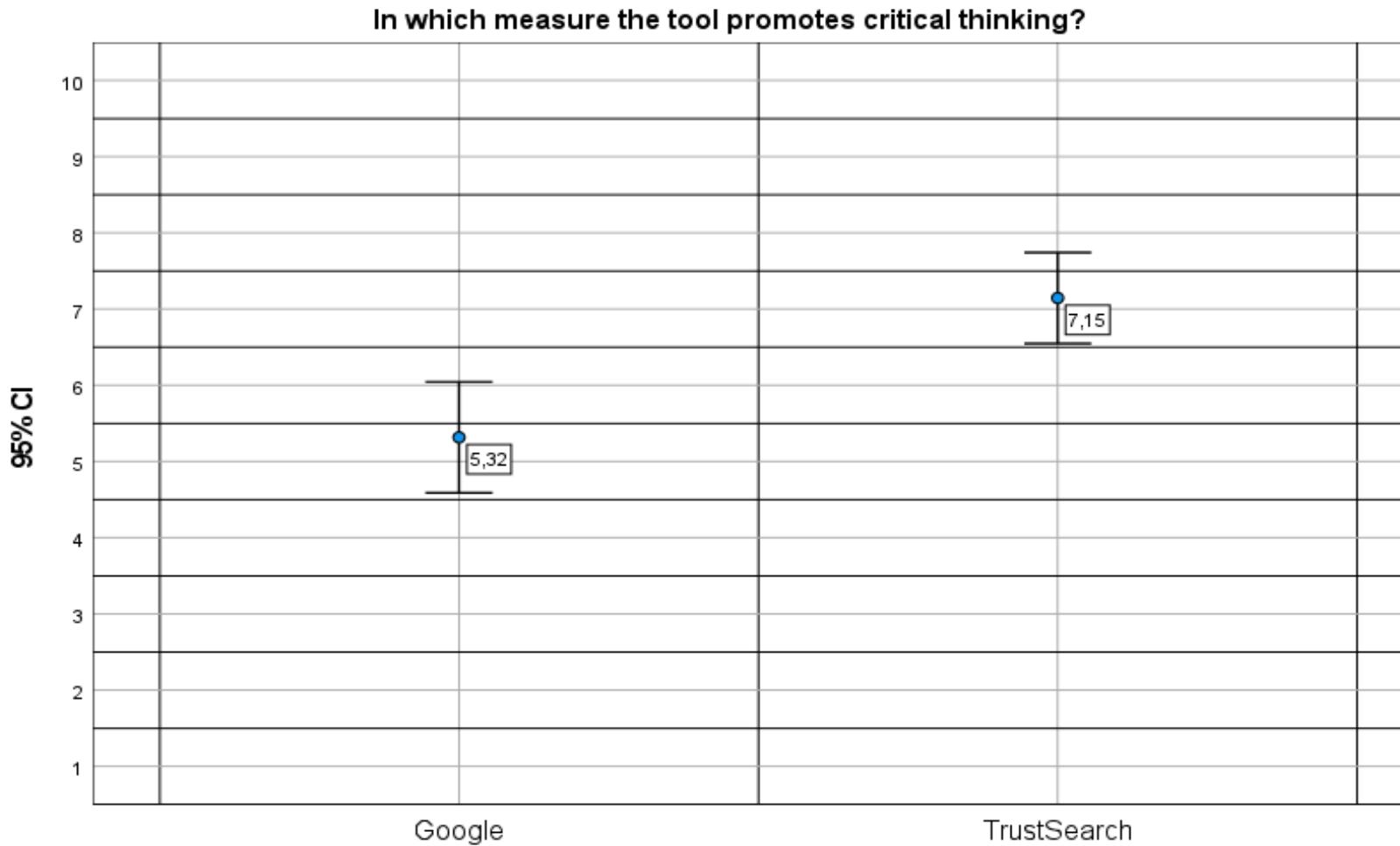
In terms of **allowing various stances on the same issue to be explored**, mean was higher for TrustSearch ($M=8.73$, $SD=1.09$) than for Google ($M=5.56$, $SD=2.41$). The difference in means (-1.82) was statistically significant, $t(40) = -3.950$, $p<.001$



In terms of **consulting sources of information that it would not have initially looked at**, mean was higher for TrustSearch ($M=7.88$, $SD=1.72$) than for Google ($M=6.46$, $SD=2.39$). The difference in means (1.41) was statistically significant, $t(40) = 2.978$, $p=.002$

User study: Evaluation of the tool

Mean values were compared for the two search engines with a Paired Sample T-Test



In a **general evaluation of the capability to promote critical thinking**, mean was higher for TrustSearch ($M=7.15$, $SD=1.89$) than for Google ($M=5.32$, $SD=1.83$). The difference in means (1.41) was statistically significant, $t(40) = 3.950$, $p < .001$