Disinformation detection

Paolo Rosso

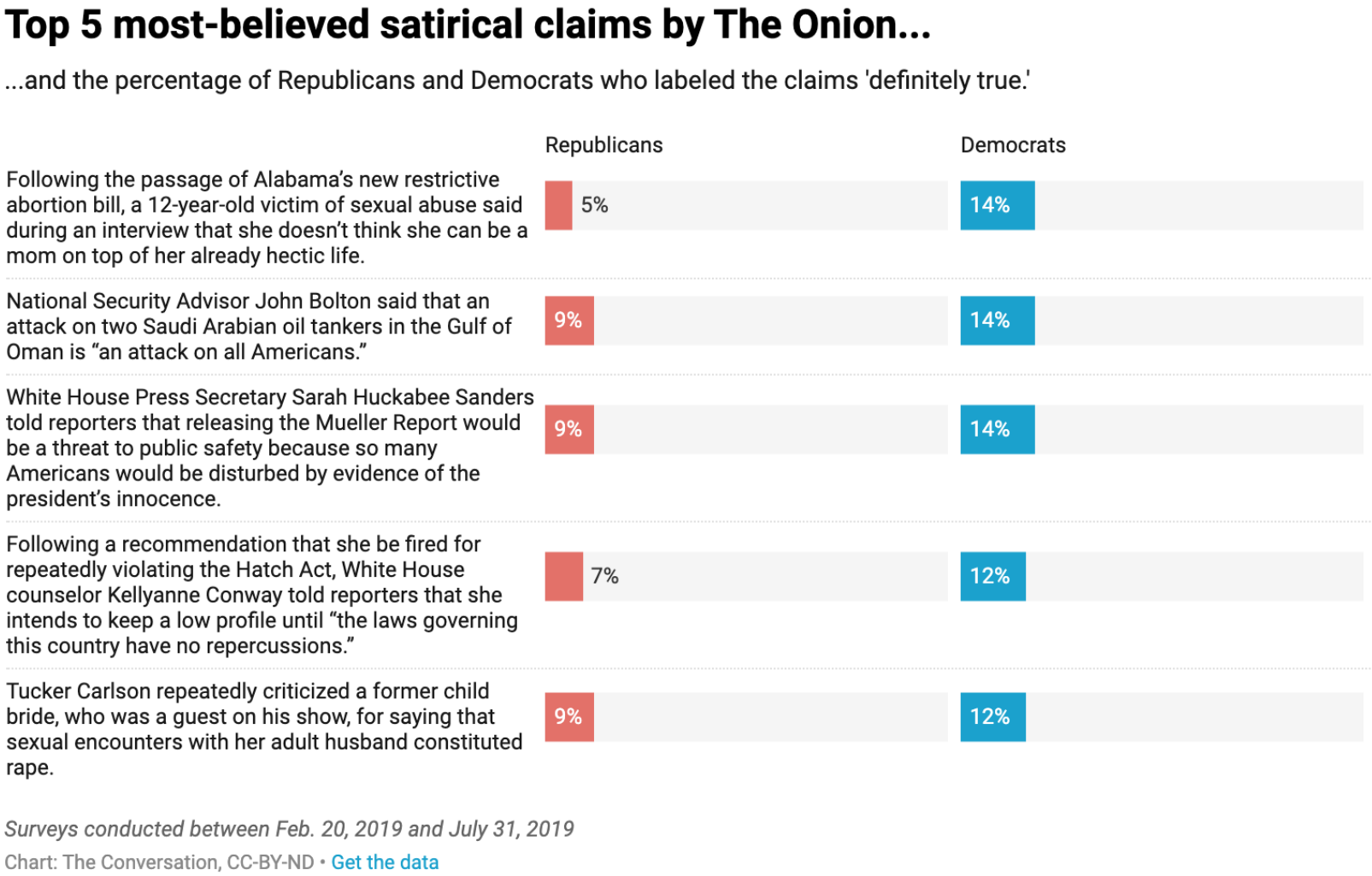
2023-24

MIARFID
Official Master's Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging

DSIC

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Misinformation (e.g. also satire) vs…



https://theconversation.com/too-many-people-think-satirical-news-is-real-121666

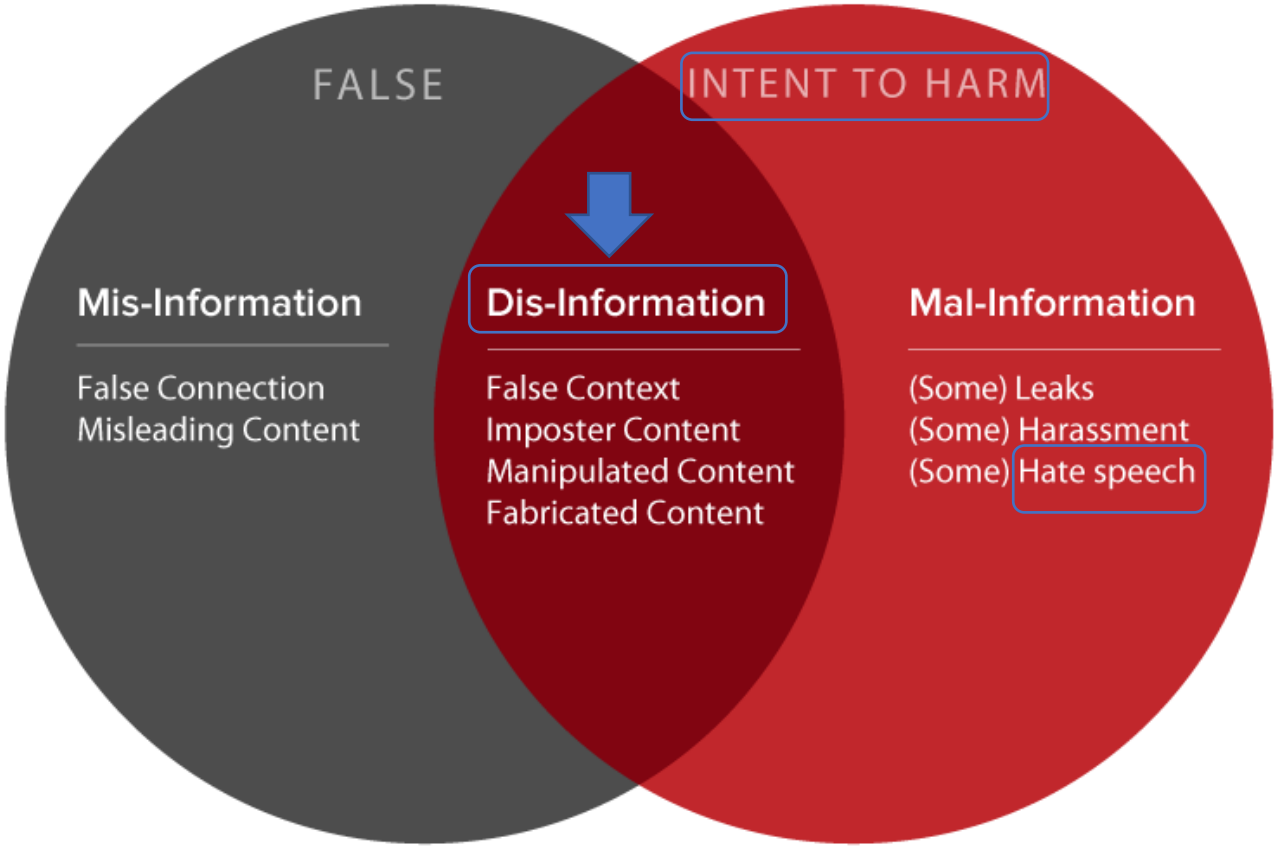
…Disinformation

- 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/

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

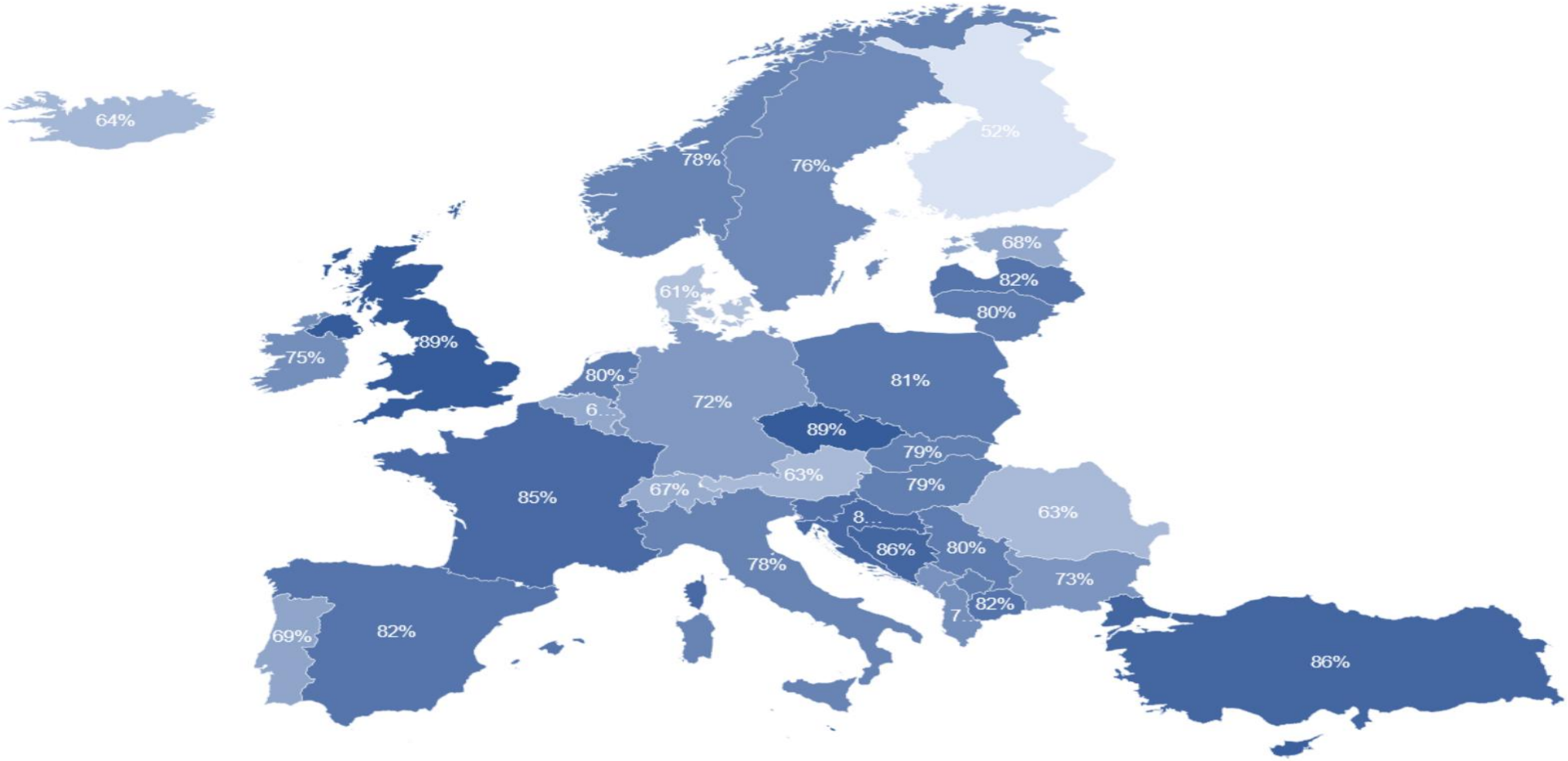
https://data.europa.eu/euodp/es/data/dataset/S2228_92_4_503_ENG

- 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

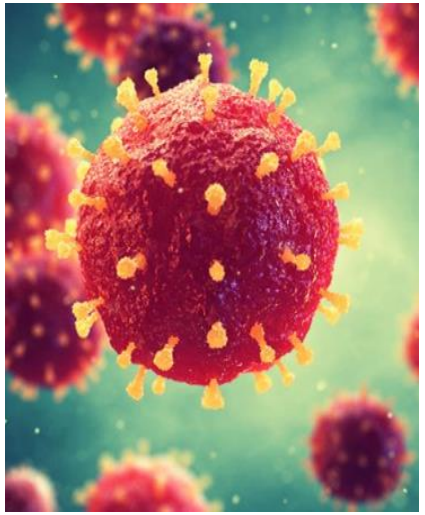
The perception of disinformation as a problem in Europe

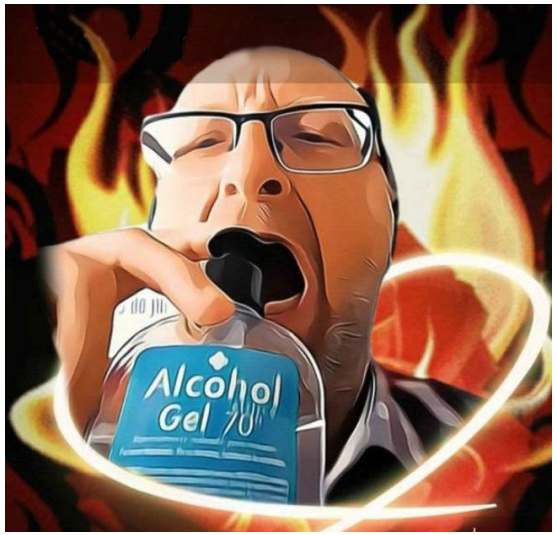


Badillo-Matos A. et al. (2023). Analysis of the Impact of Disinformation on Political, Economic, Social and Security Issues, Governance Models and Good Practices: The cases of Spain and Portugal. Pamplona: IBERIFIER.

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

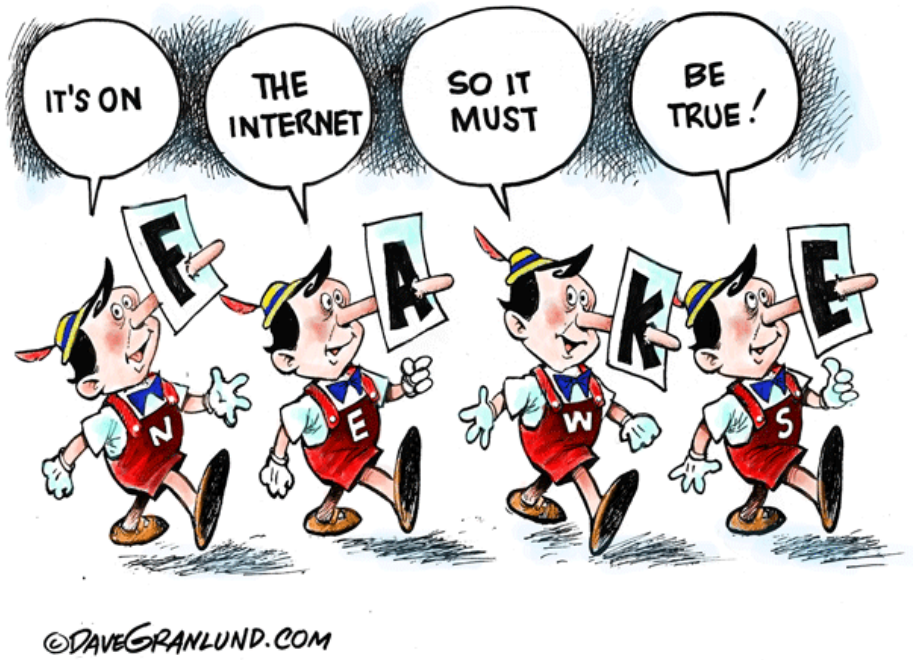




IBERIFIER
Iberian Media Research & Fact-Checking

European Digital Media Observatory

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

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.

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 EU will result in the NHS getting £ 350 million a wee

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

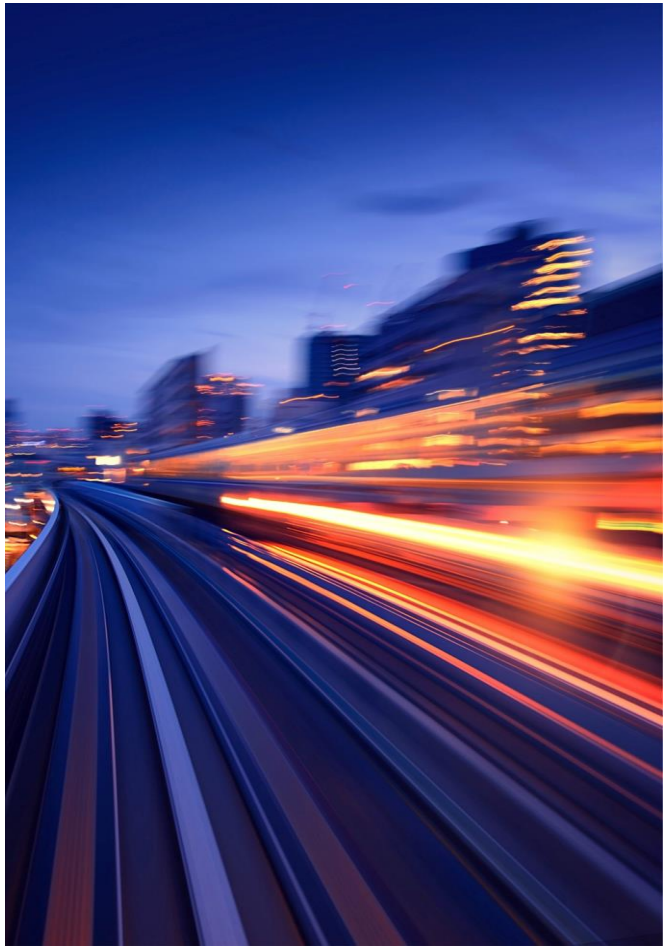
Man tries to smuggle turtle onto plane by hiding it in a hamburger

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)



Vosoughi S., Roy D., Aral S. (2018). **The spread of true and false news online**. Science 359, 6380, 1146–1151.

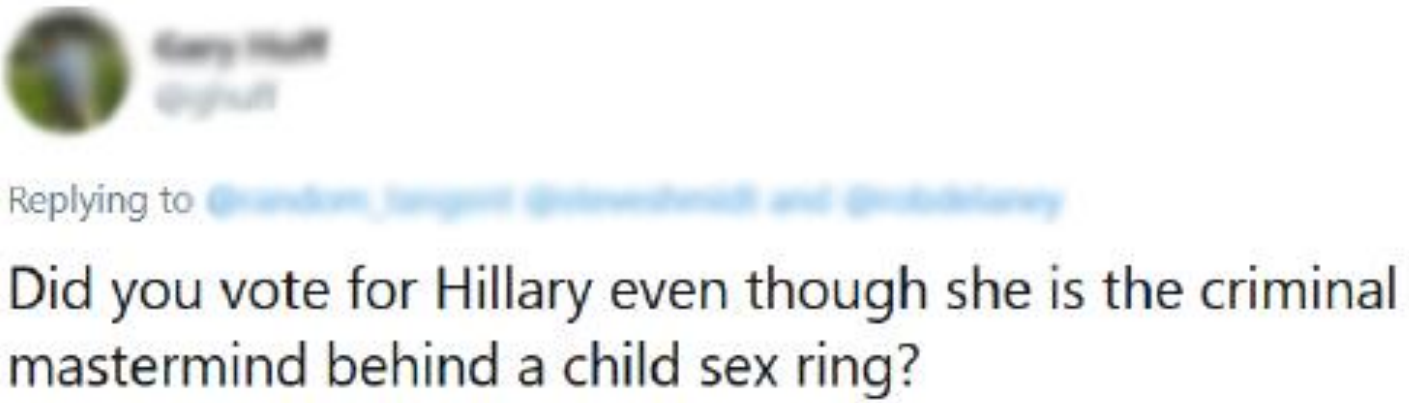
Fake or not fake? That's the question

The money saved from leaving the EU will result in the NHS getting

£ 350 million a week

disgust and surprise?

False info on Twitter



False info on Twitter

Fear, disgust and surprise?

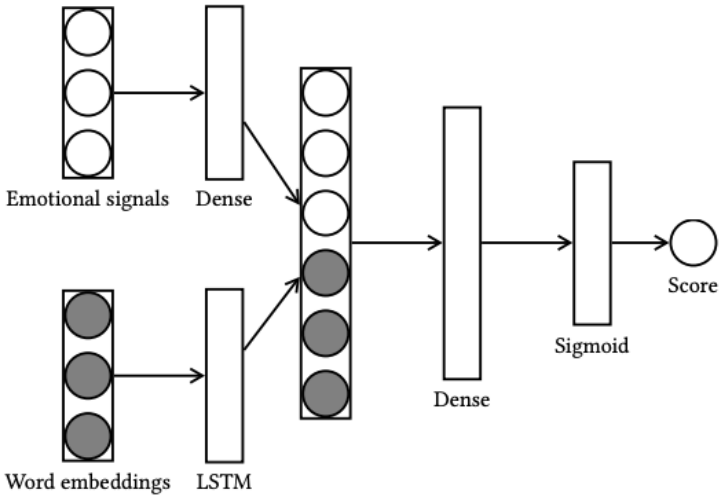
Information credibility on Twitter

- **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:

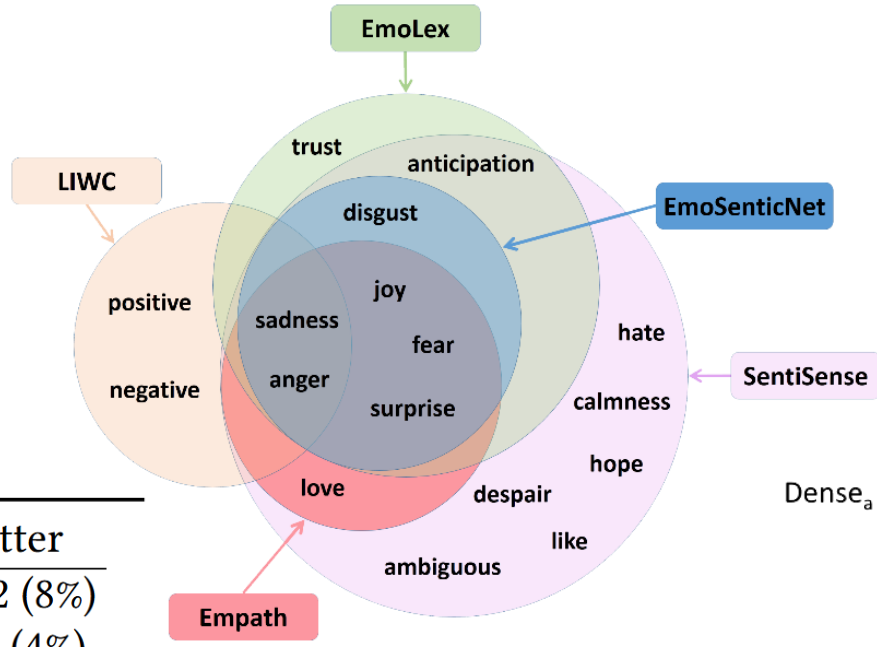
- **emoLexi**
- **emoInt**
- **emoReact**: LSTM network that predicts for each claim the probability to trigger any of the three intensity levels (low, average, high) for each of the five reactions love, joy, surprise, sadness and anger

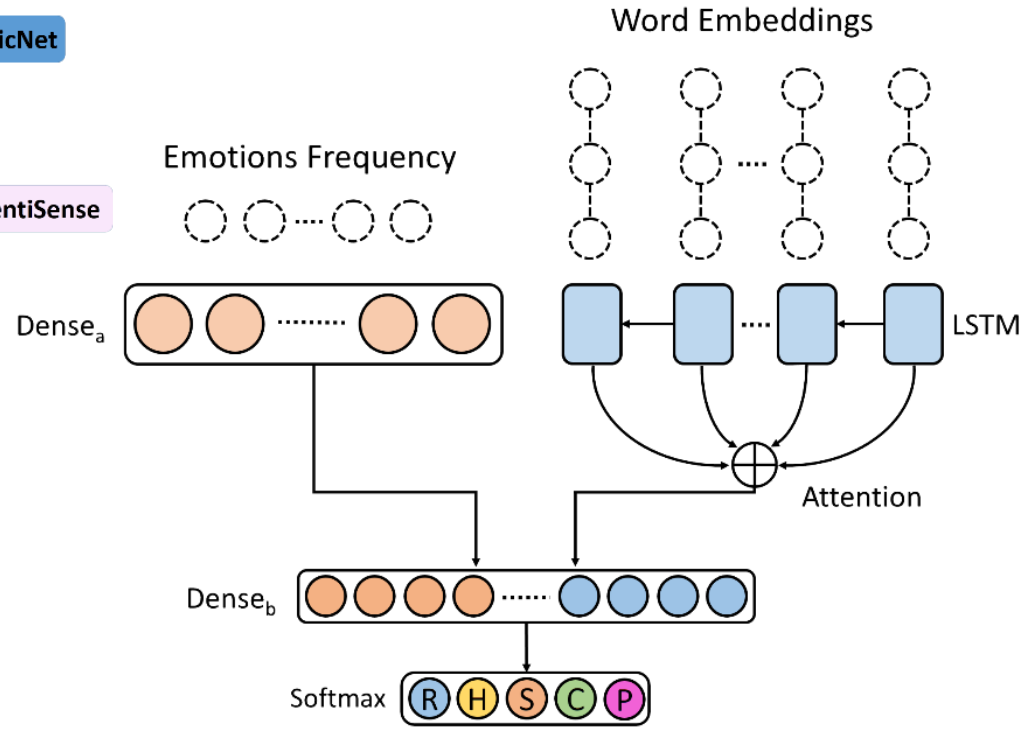
Information credibility on Twitter

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**



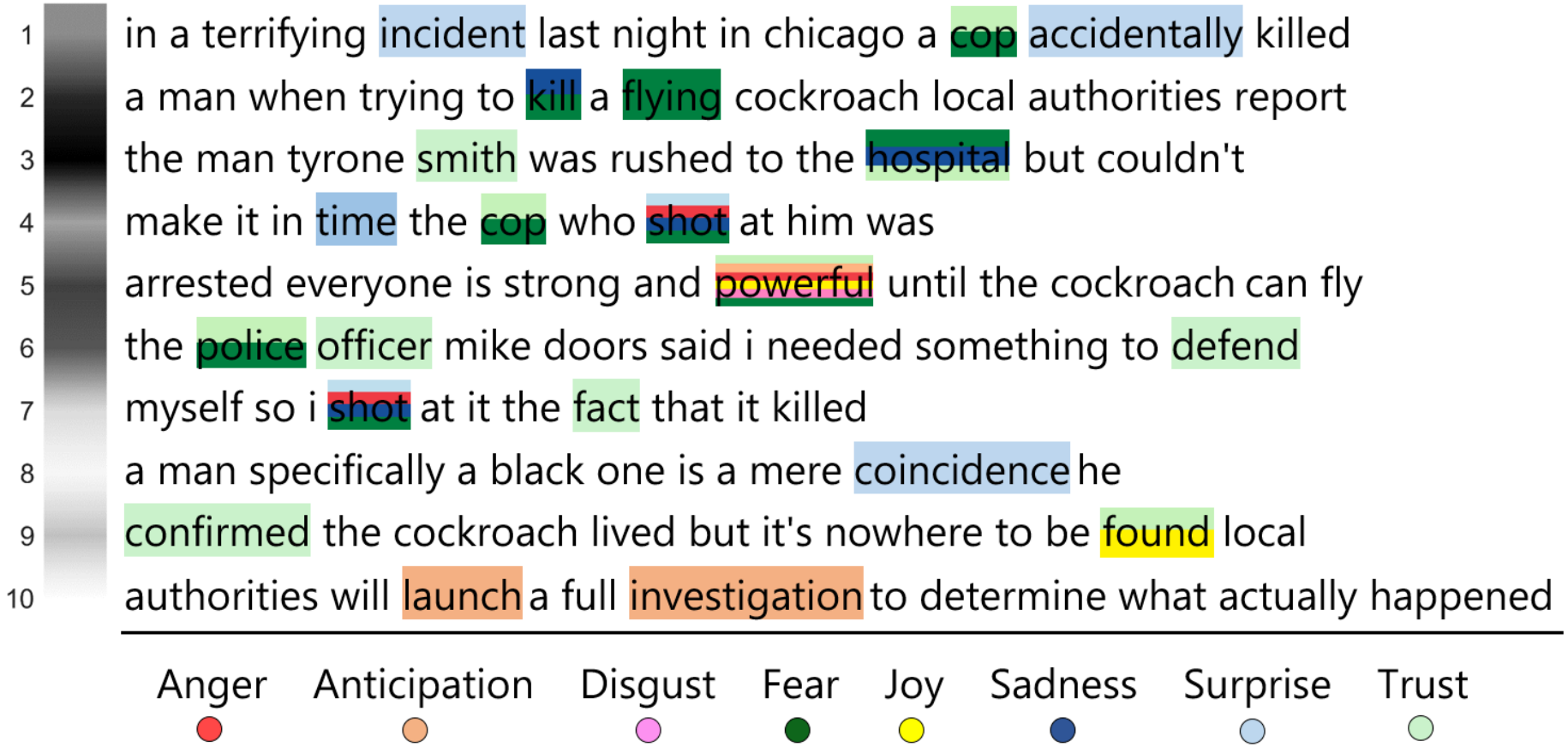


Ghanem B., Rosso P., Rangel F. (2020). **An Emotional Analysis of False Information in Social Media and News Articles**. ACM Transactions Internet Technology. 20(2): 1-18

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

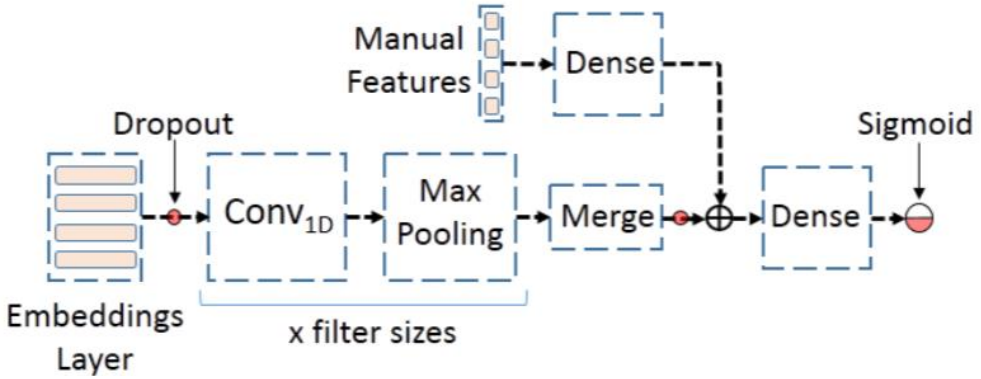
How emotions flow in disinformation



Ghanem B., Ponzetto S., Rosso P., Rangel F (2021) **FakeFlow: Fake News Detection by Modeling the Flow of Affective Information**. In: Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2021, pp. 679–689

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, deatl
  - time focus (past, present, future)
  - cognitive processes (causation, discrepancy, tentative, certaint
  - informal language (swear, assent, nonfluencies, fillers)
  - affective processes (anxiety)
- **Personality** scores with **Big Five**
  - **O**penness to experience (unconventional, insightful, imaginati
  - **C**onscientiousness (organised, self-disciplined, ordered)
  - **E**xtraversion (cheerful, sociable, assertive)
  - **A**greeableness (cooperative, friendly, empathetic)
  - **N**euroticism (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

Neuman Y., Cohen Y. (2014) **A Vectorial Semantics Approach to Personality Assessment**. Sci Rep 4, 4761

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**

Sakketou F., Plepi J., Cervero R., Geiss H.J., Rosso P., Flek L. (2022) **A New Dataset for Identifying Misinformation Spreaders and Political Bias**. In: Proc. 13th Int. Conf. on Language Resources and Evaluation (LREC), pp. 3231-3241

Disinformation detection should consider also:

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

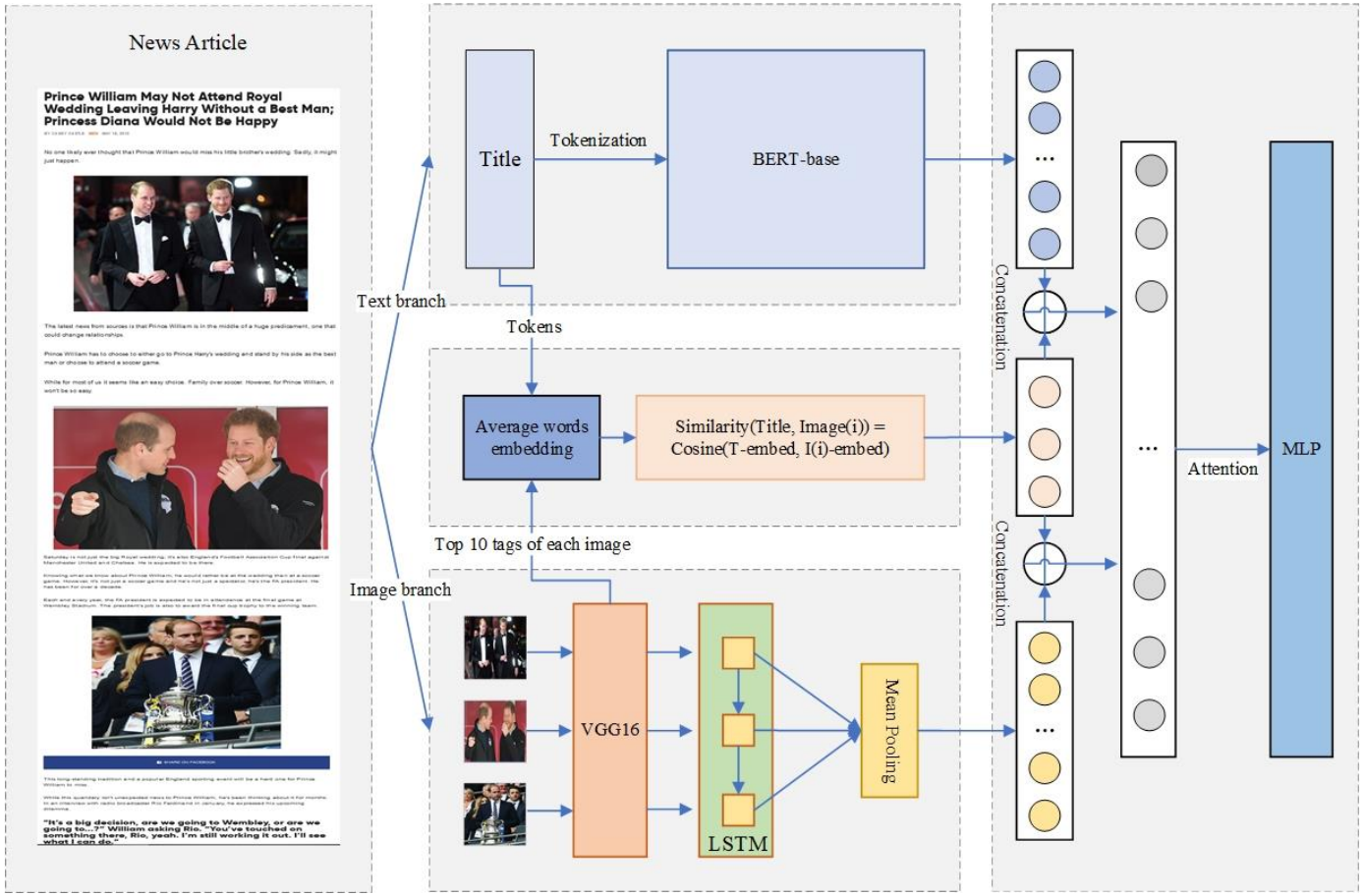
Multimodal (multi) image fake news detection

- 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

Giachanou A., Zhang G., Rosso P. (2020). **Multimodal Multi-image Fake News Detection**. In Proc. of the 7th IEEE Int. Conf. on Data Science and Advanced Analytics Special Session on Fake news, bots and trolls, DSAA-2020, pp. 647-654

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

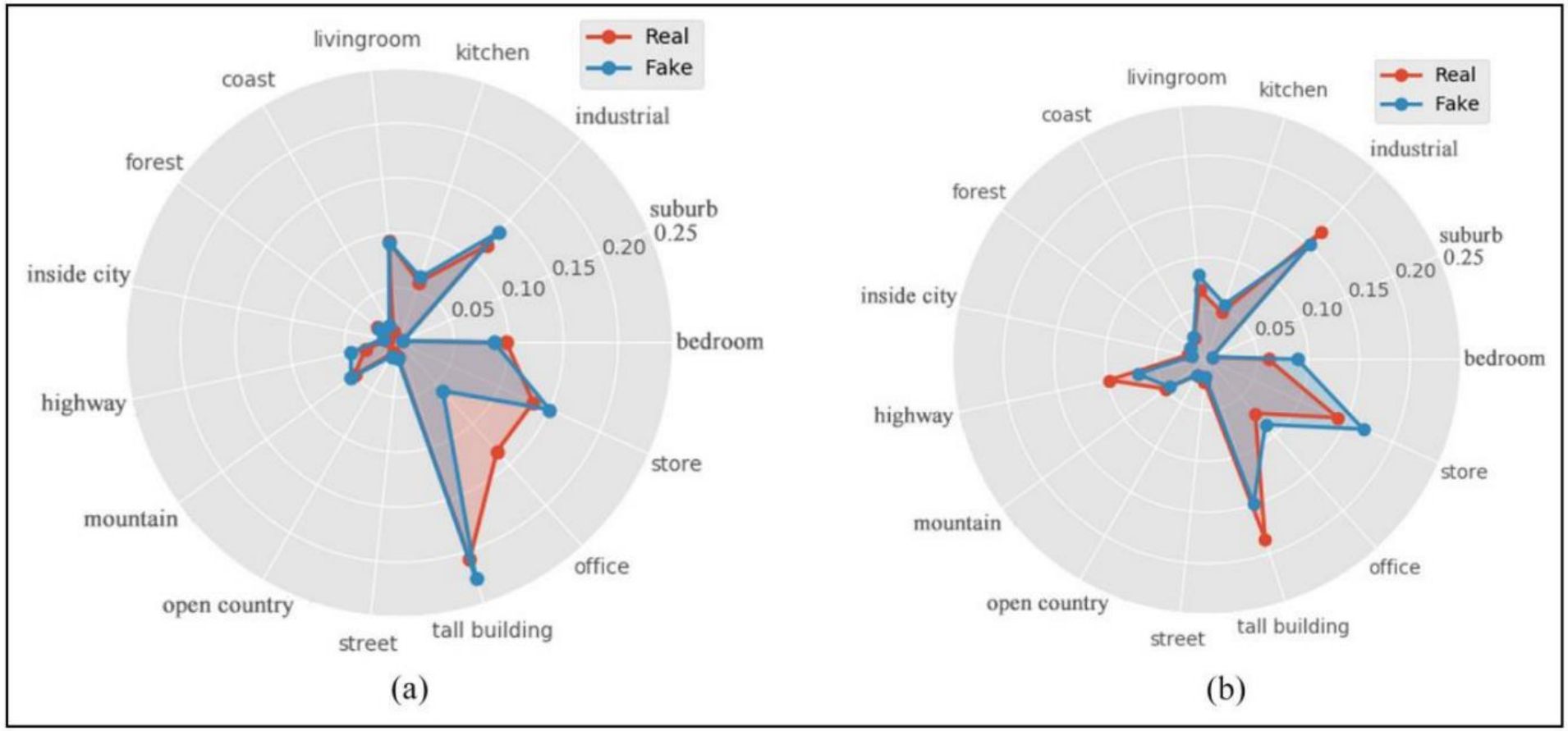


Figure 1. Average place scene scores for real and fake news of (a) PolitiFact and (b) GossipCop datasets.

Zhang G., Giachanou A., Rosso P. (2022). **SceneFND: Multimodal Fake News Detection by Modeling Scene Context Information**. Journal of Information Science. https://doi.org/10.1177/01655515221087683

Modeling scene context information

Place, weather, season

Table 1. Scene co-occurrence ranking.

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

Disinformation detection should consider also:

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

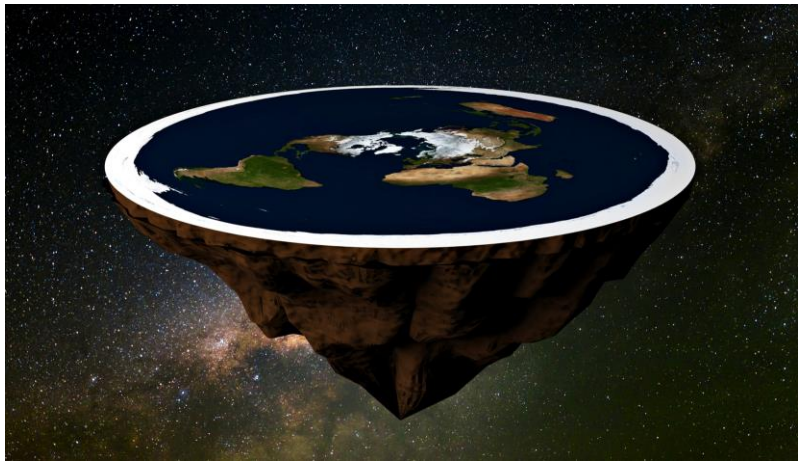
Foreign Information Manipulation Interference

- **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 #climateChangeIsReal #earthisnotflat #nasatruth #nasaIsReal #spaceIsReal #moonlandingisreal
users	977	950
tweets	912,735	992,798

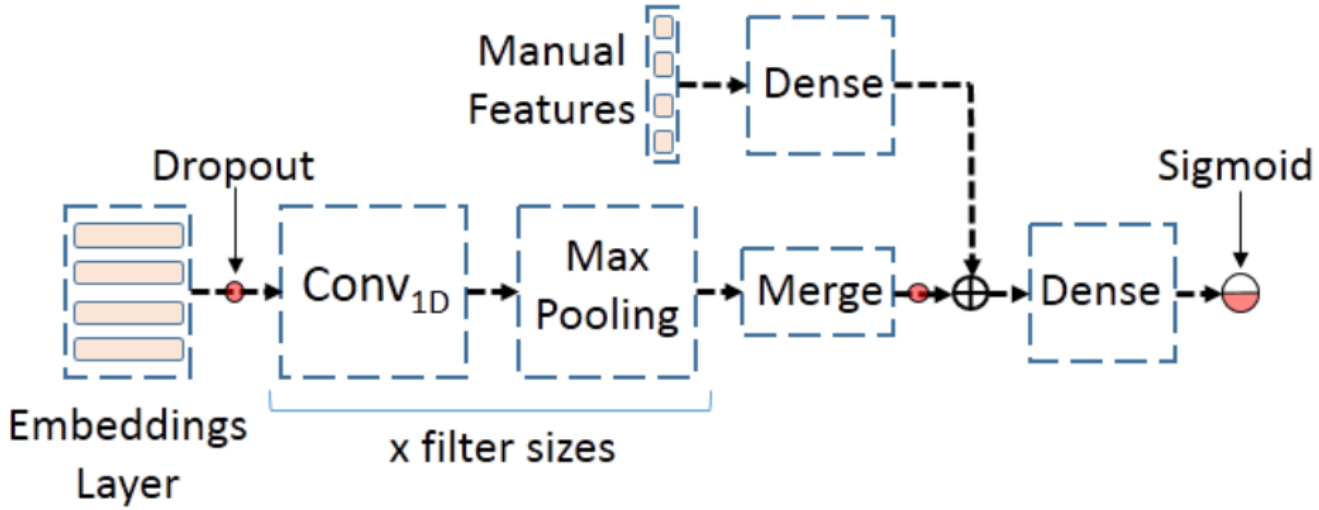


Giachanou A., Ghanem B., Rosso P. (2023) **Detection of Conspiracy Propagators using Psycho-linguistic Characteristics**. Journal of Information Sciences, 49(1):3-17

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	0.710
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	0.355
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

https://sites.google.com/view/aisome/aisome

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

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

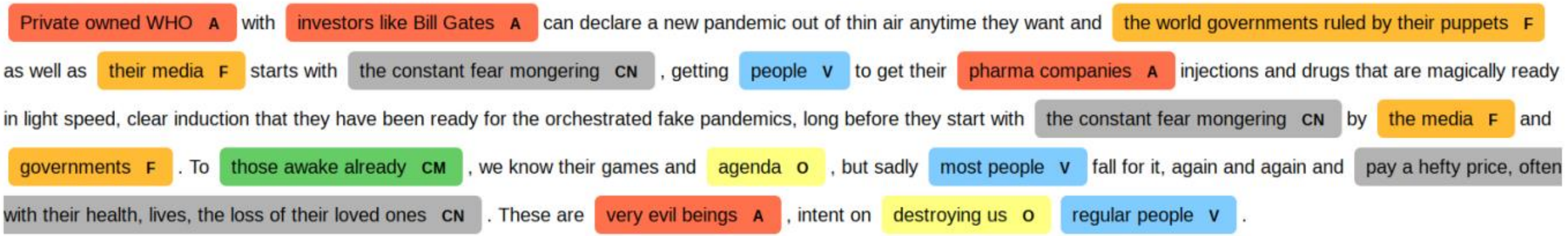
Holur P., Wang T., Shahsavari S., Tangherlini T., Roychowdhury V. (2022).
Which Side are you On? Insider-Outsider Classification in Conspiracy-theoretic Social Media.
Proc. of the 60th Annual Meeting of the Association for Computational Linguistics, pp. 4975 - 4987

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)



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

MINISTERIO DE CIENCIA E INNOVACIÓN

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

Stance and ideological orientation for promoting **critical thinking**

NGI

NGI.eu Next Generation Internet

A pilot AI tool

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

BERTa Chulvi

Ivan Grubišić

Andrea Simeri

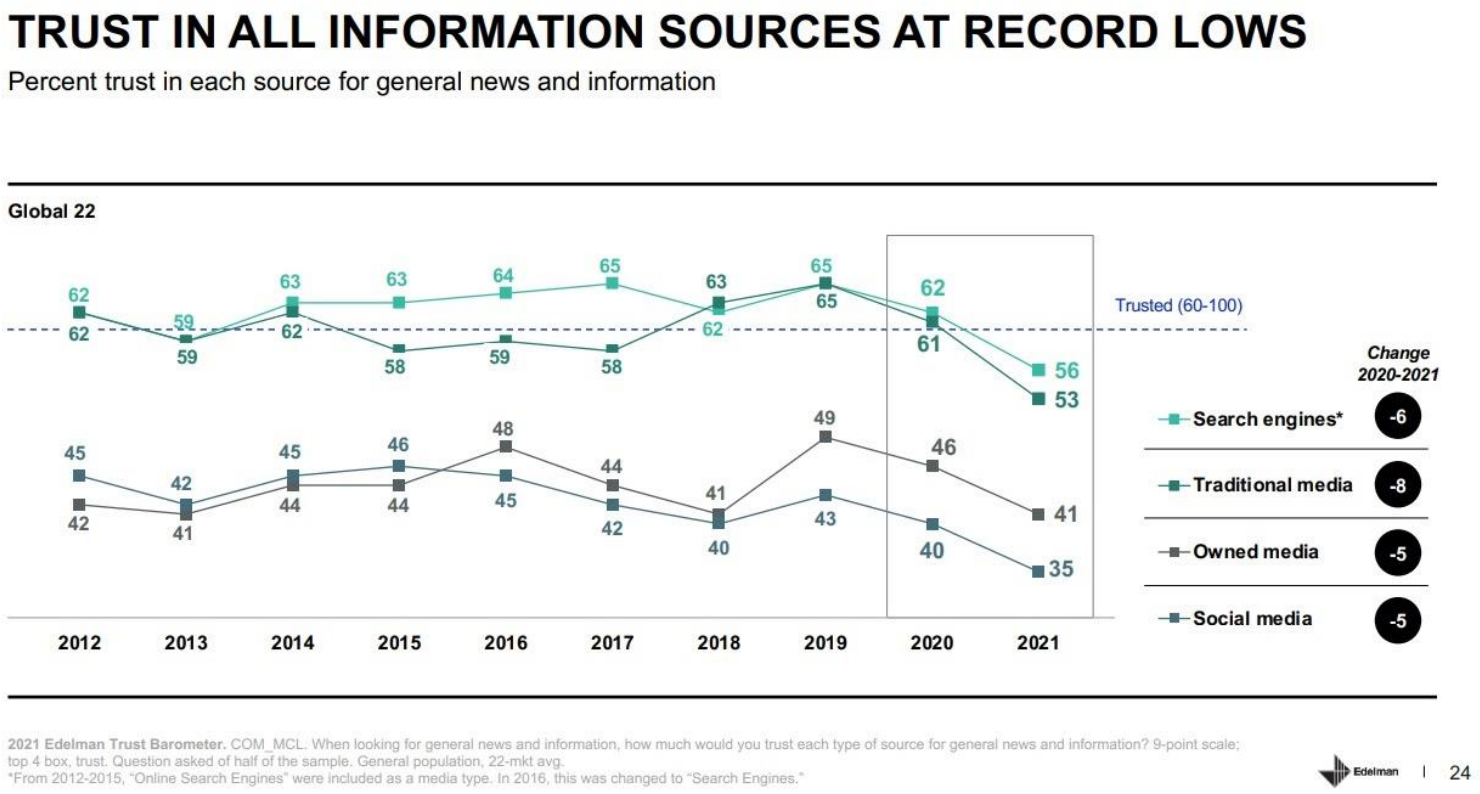
Paolo Rosso

valgrai.eu

ATC

ATHENS TECHNOLOGY CENTER

The problem



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

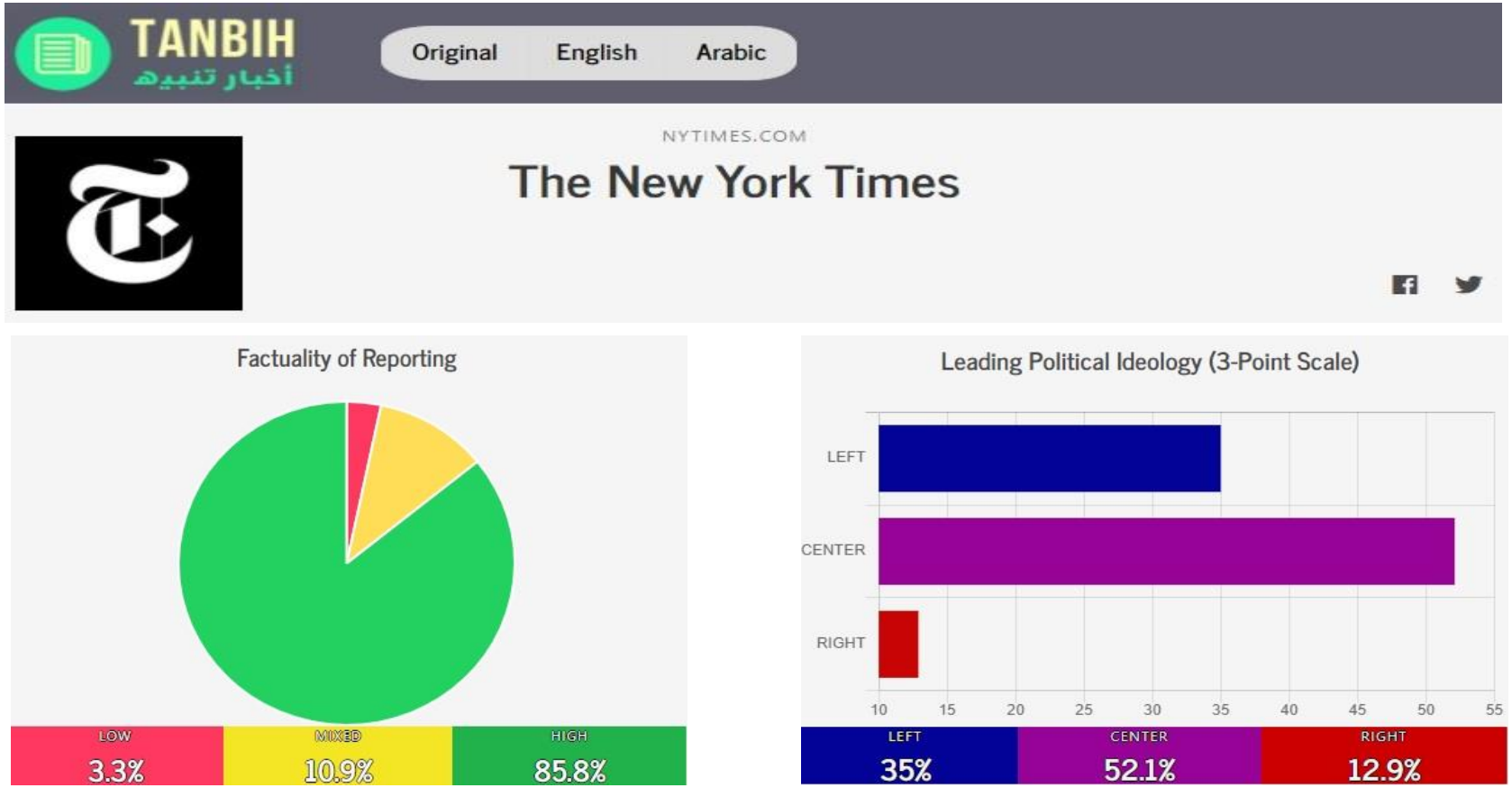
The problem

SPAIN 73%

Related work from communication science

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

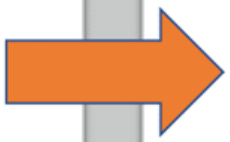


Bali R. et al. (2020) **We Can Detect Your Bias: Predicting the Political Ideology of News Articles**.
In Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing EMNLP '20, pp. 4982–4991

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

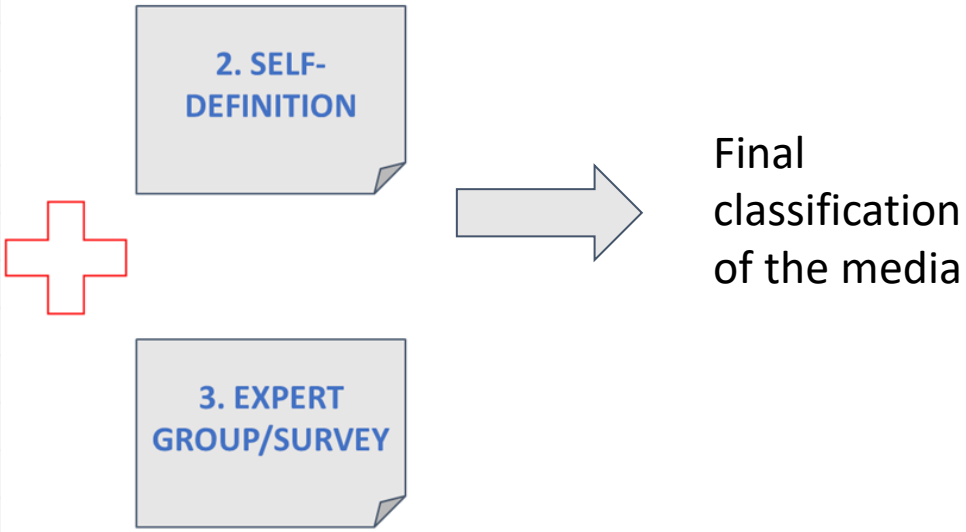
Ideological orientation of the media

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 PAIS	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 (Scottland)	Left

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

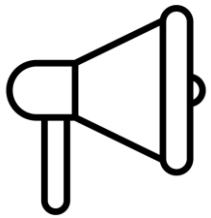
https://mediabiasfactcheck.com/



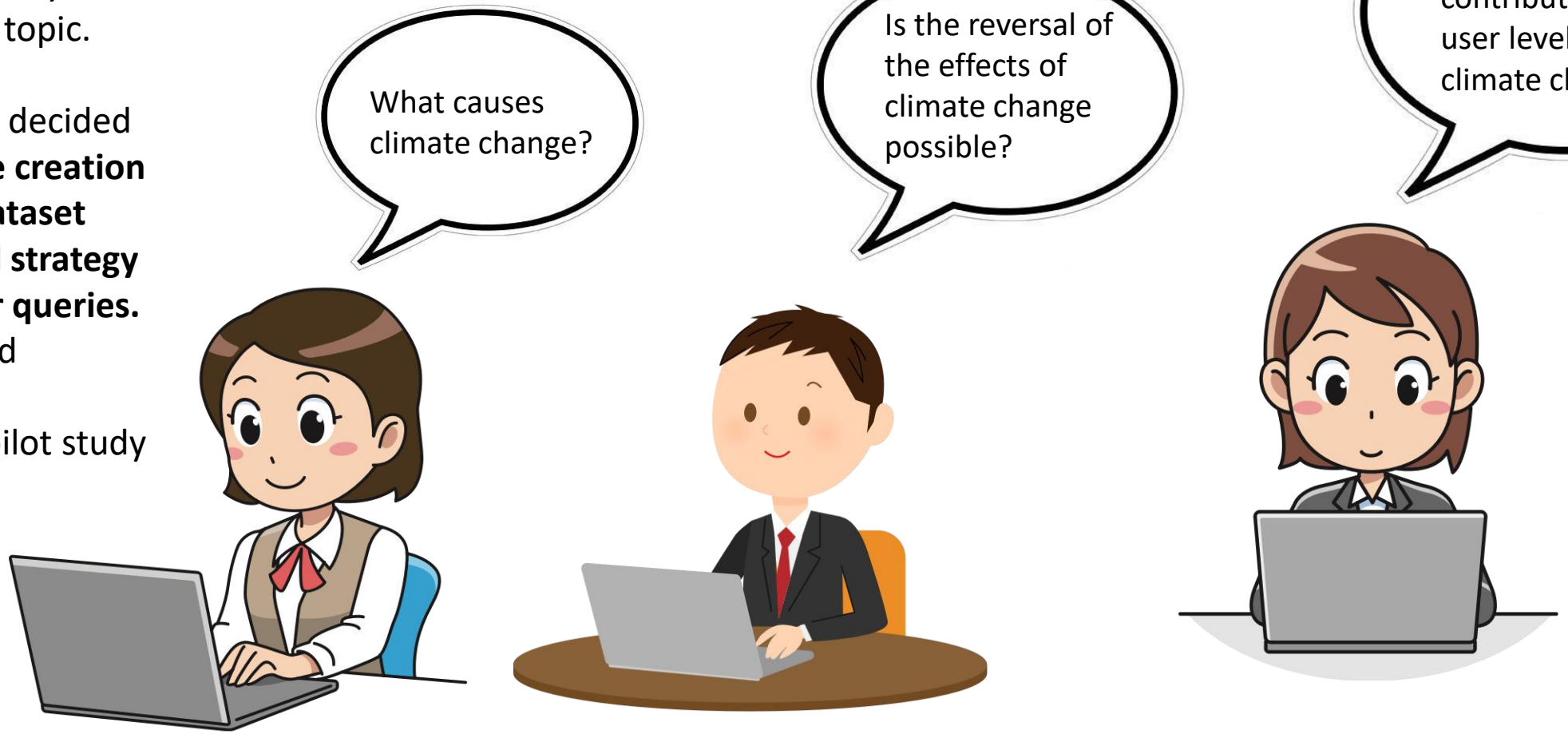
Collecting the data: pilot study with real users

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…



Collecting the data: pilot study with real users

Do you have some previous ideas, some opinions?

Excess emissions

What causes climate change?

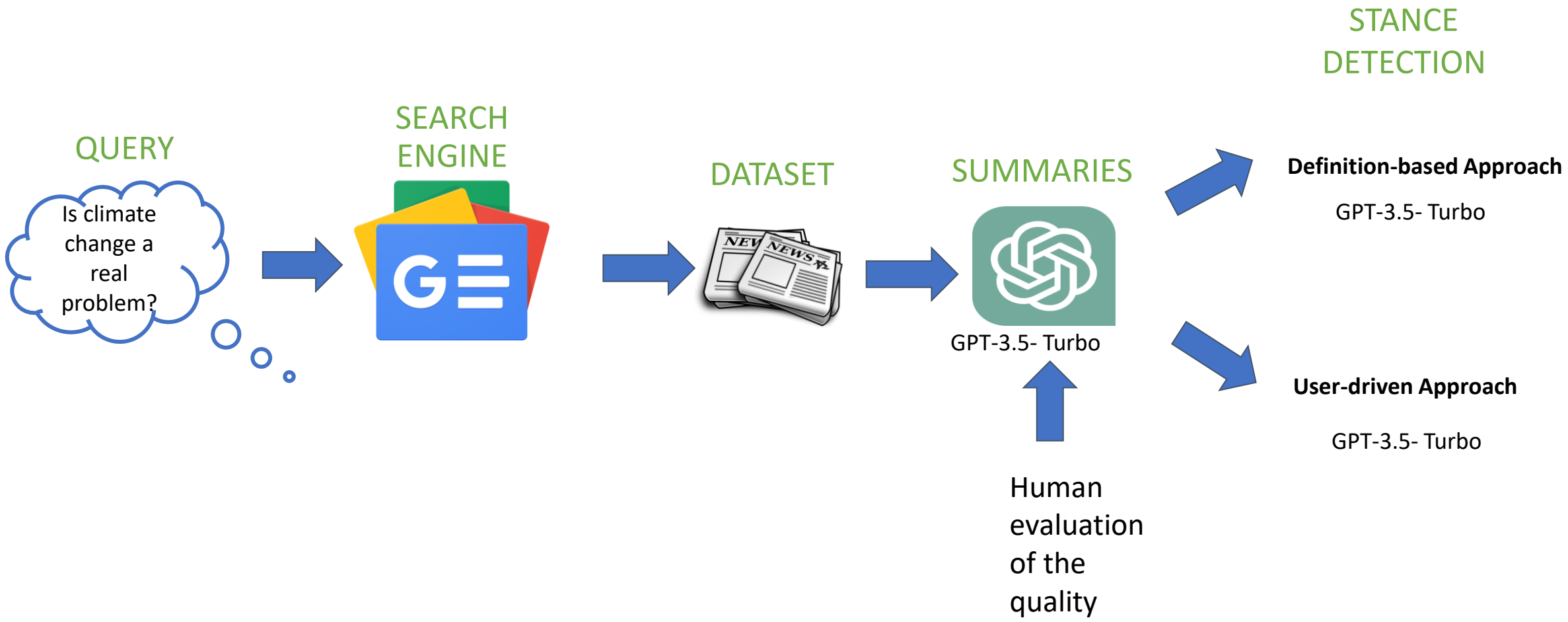
The reversal time is ending

Is the reversal of the effects of climate change possible?

I think it is important that there are more citizen training

How can you contribute to the user level to combat climate change?

Stance detection workflow



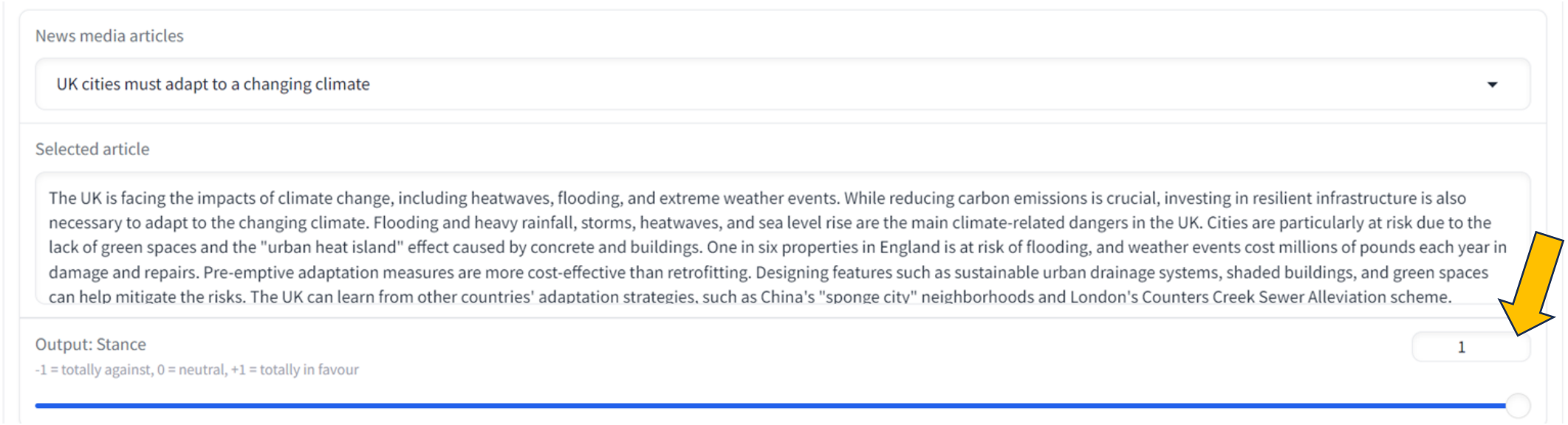
Definition-based approach for stance detection

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 ether 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



User-driven approach for stance detection

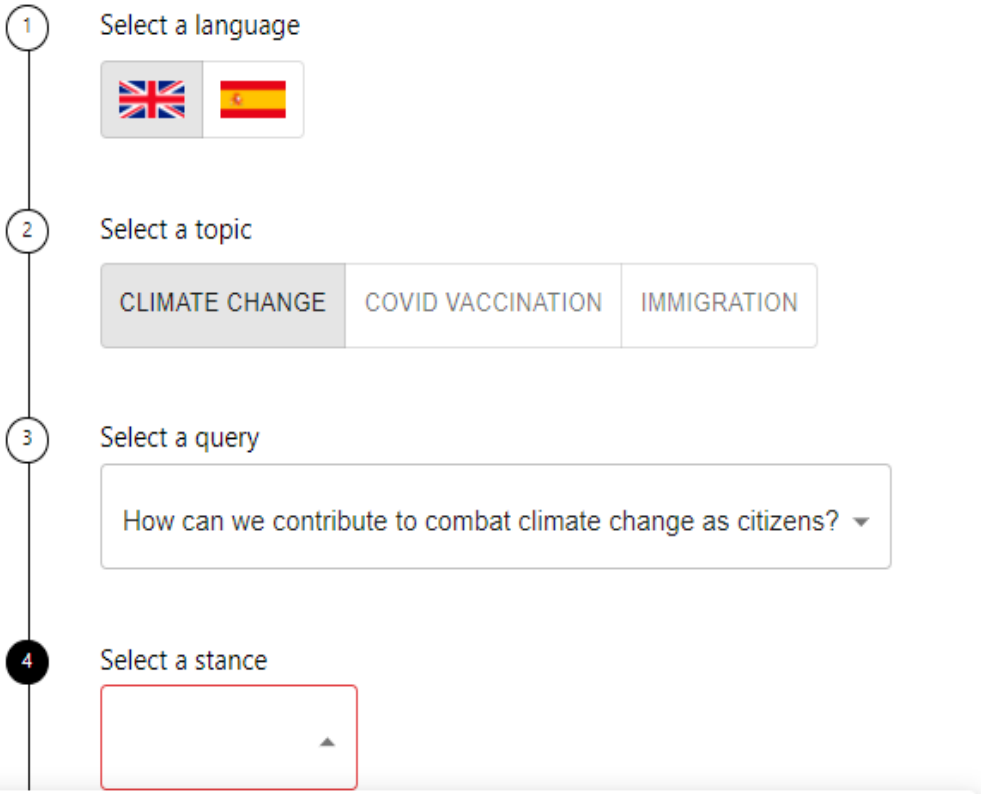
The user-driven stance detection approach, rooted in the **entailment task**: diverges from the Definition-based stance detection methodology by focusing on contextual nuances in user-generated content.

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**.

TrustSearch AI pilot tool



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?

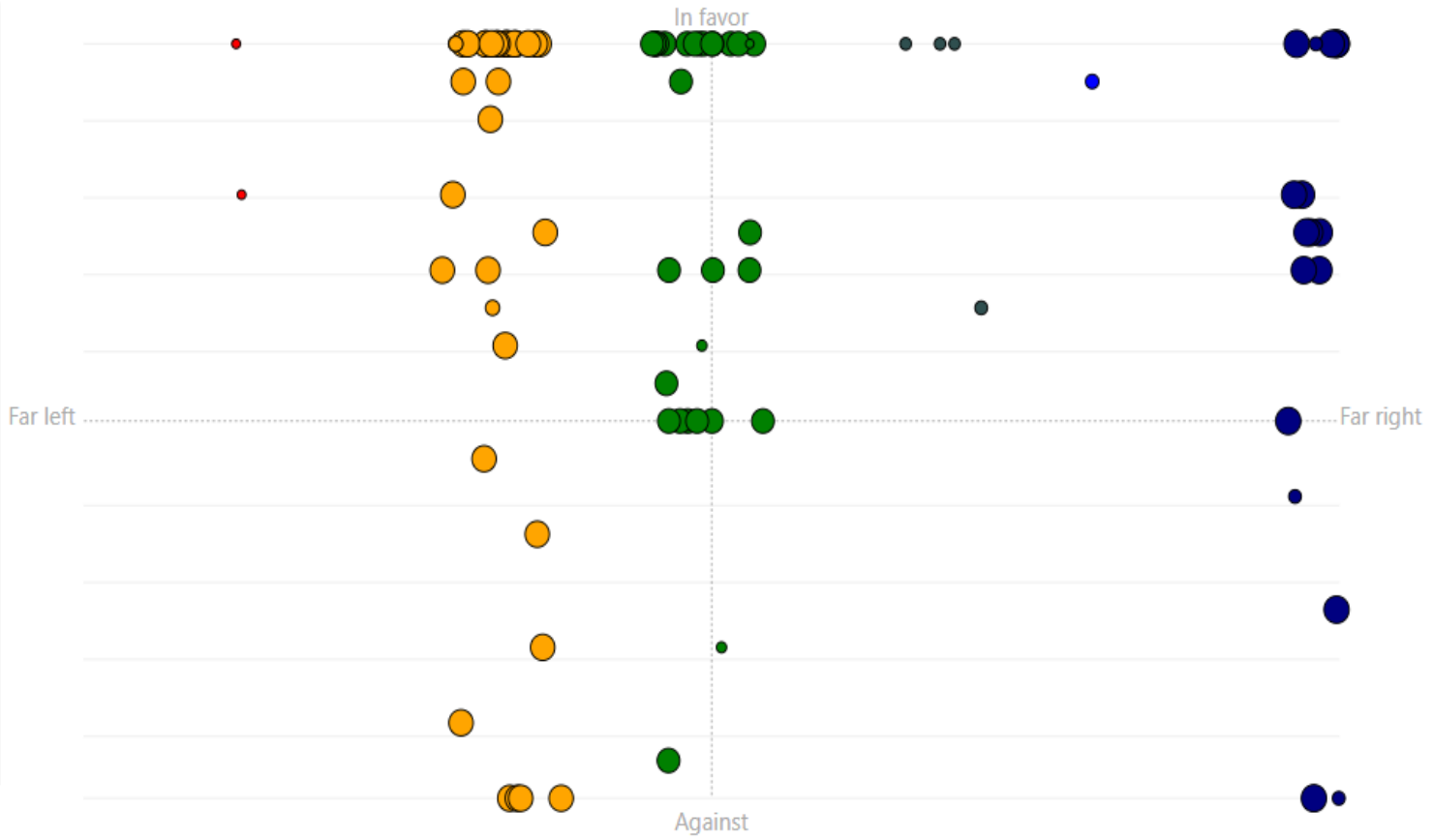
How do we calculate the stance?

Circle color represents media's political ideology

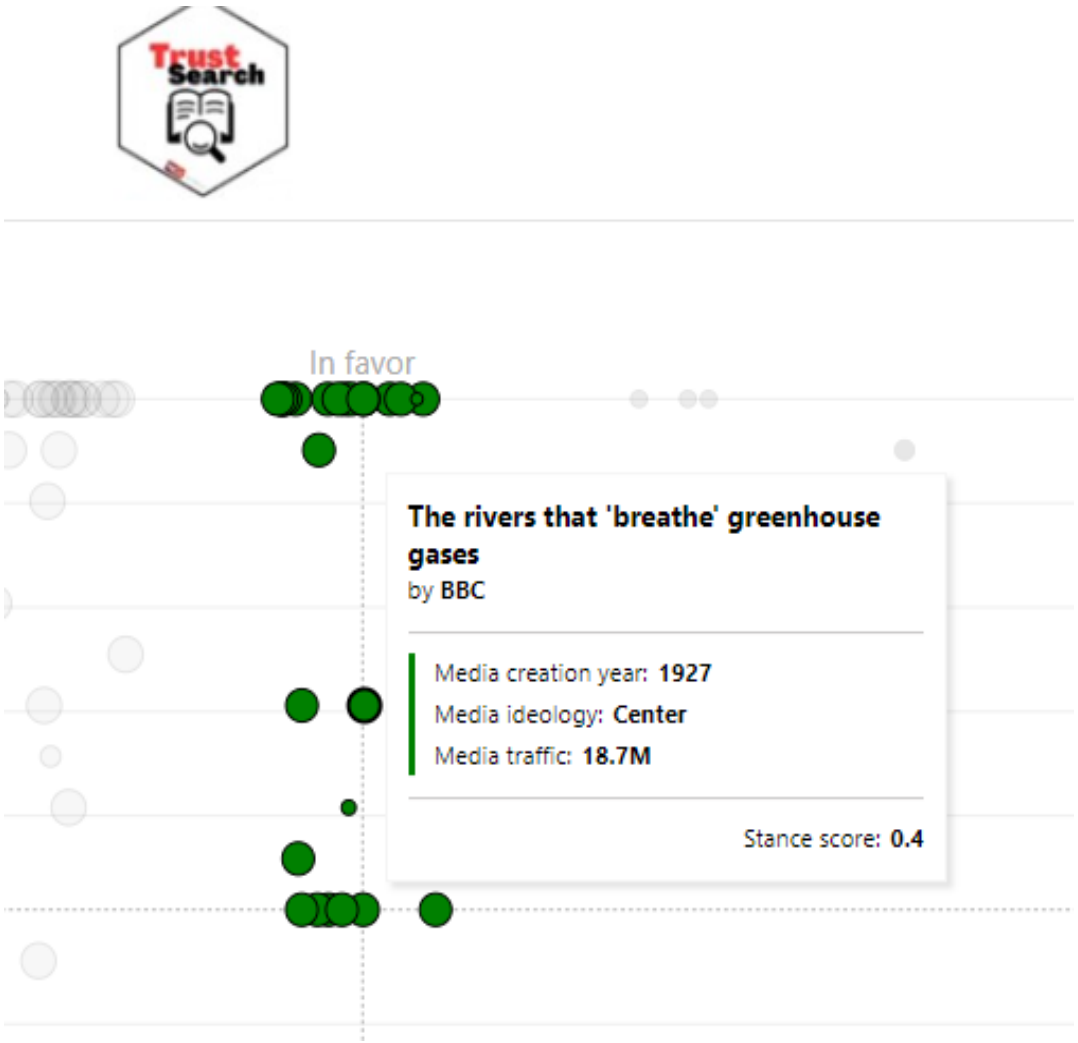
- Far left
- Left
- Left center
- Center
- Right center
- Right
- Far right

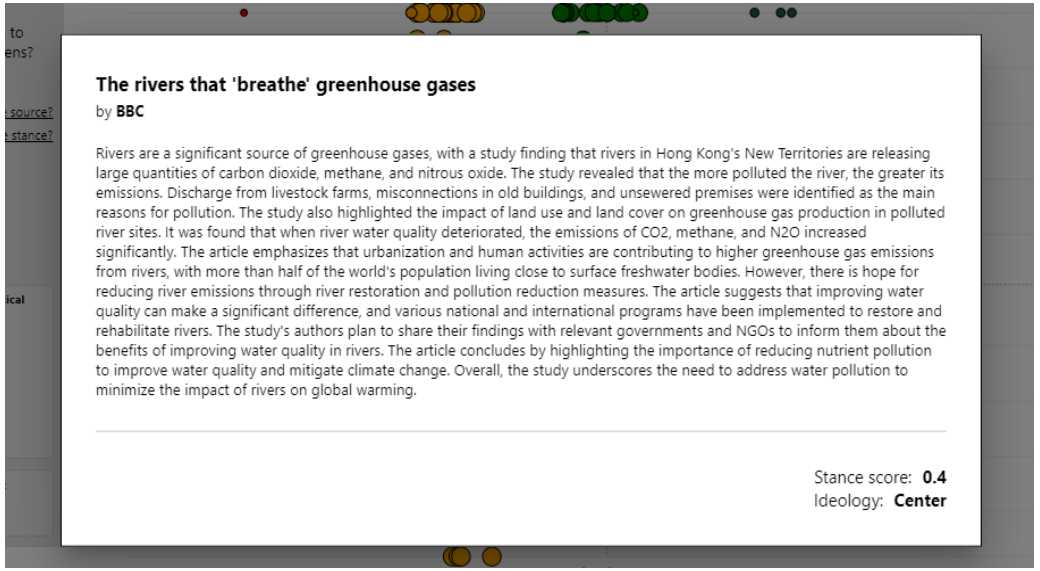
Circle size represents media's traffic

- Minimum
- Maximum

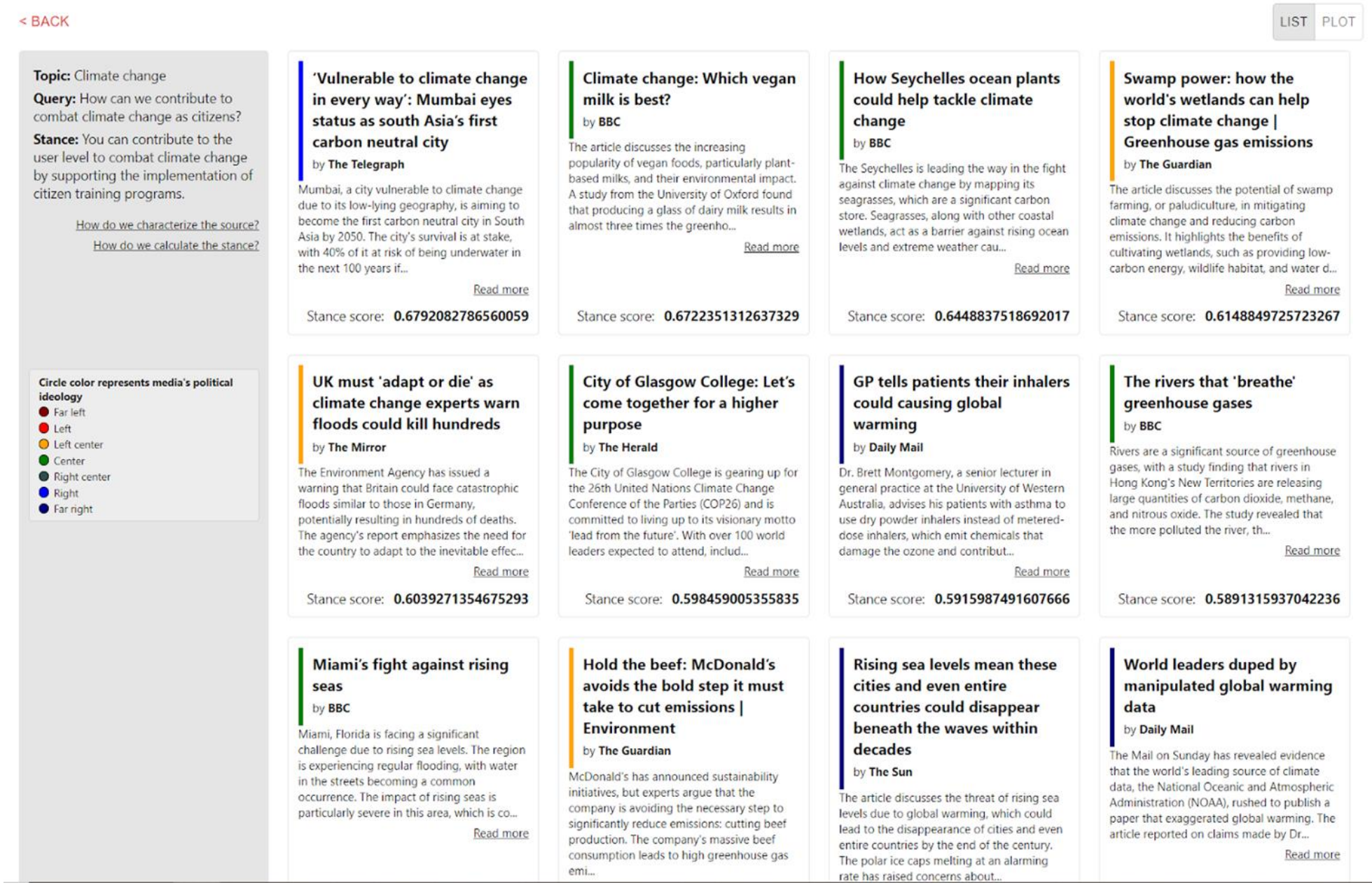


TrustSearch AI pilot tool





TrustSearch AI pilot tool



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

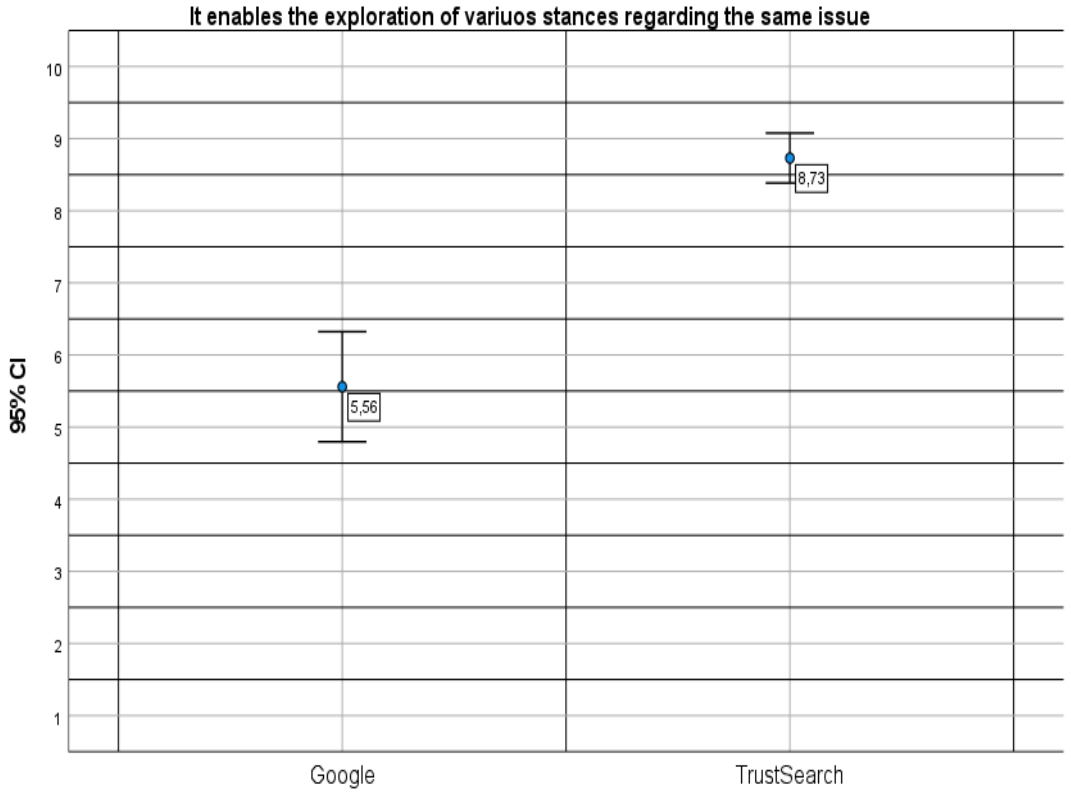
PHASE 4- Critical Thinking Disposition (CTD) (Last evaluation)

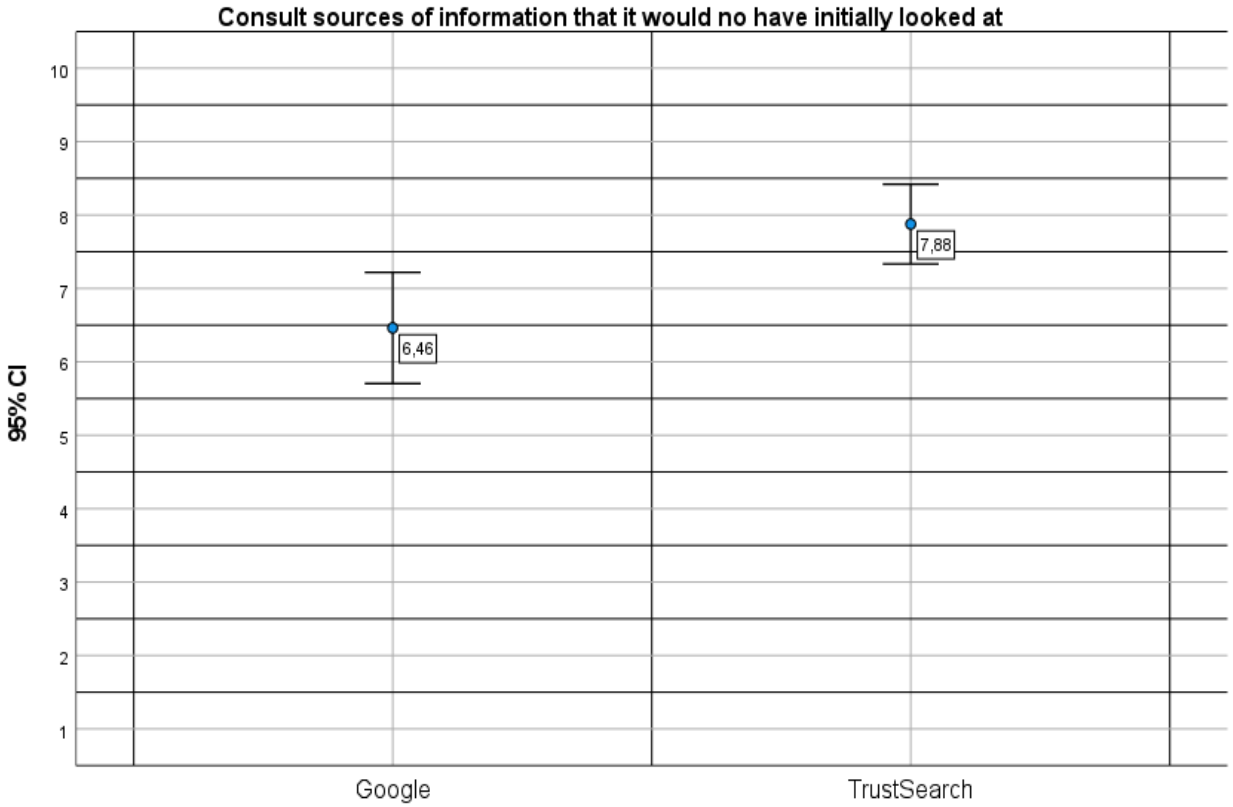
Questionnaire Q2

- 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



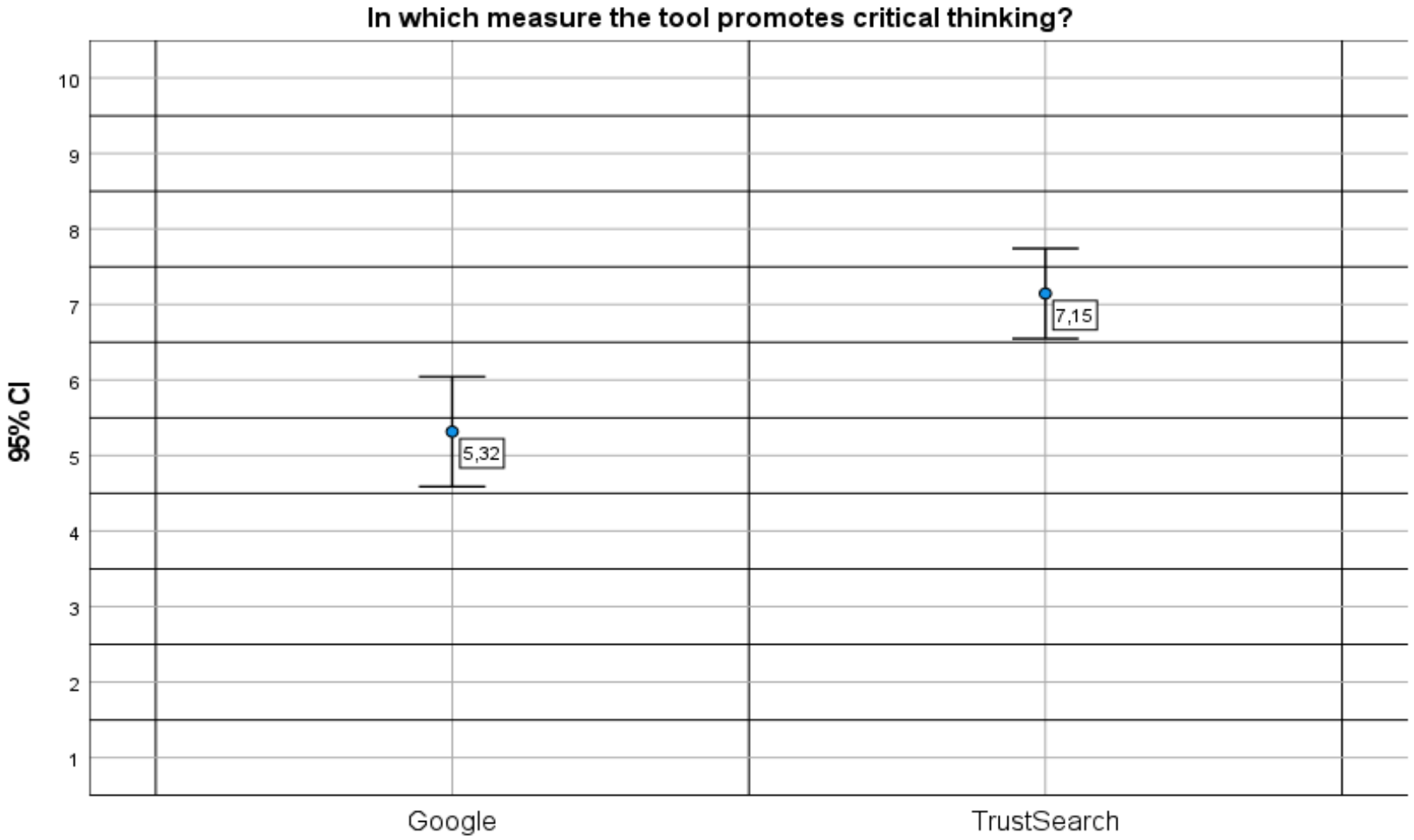


In terms of **allowing various stances on the same issue to be explored**, mean was higher for TrusSearch (M=8.73, SD=1.09) than for Google (M=5.56, SD=2.41. The difference in means (-1.82) was statistically significant, t(40) = -3.950, p<.001

In terms of **consulting sources of information that it would no have initially looked at**, mean was higher for TrusSearch (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 TrusSearch (M=7.15, SD=1.89) than for Google (M=5.32, SD=1.83). The difference in means (1.41) was statistically significant, $t(40) = 3.950$, $p<.001$