MALINFORMATION & INTELIGENCIA ARTIFICIAL

David Camacho

Universidad Politécnica de Madrid

Applied Intelligence & Data Analysis (AIDA) group

May 30th 2024, Fundación Ramón Areces

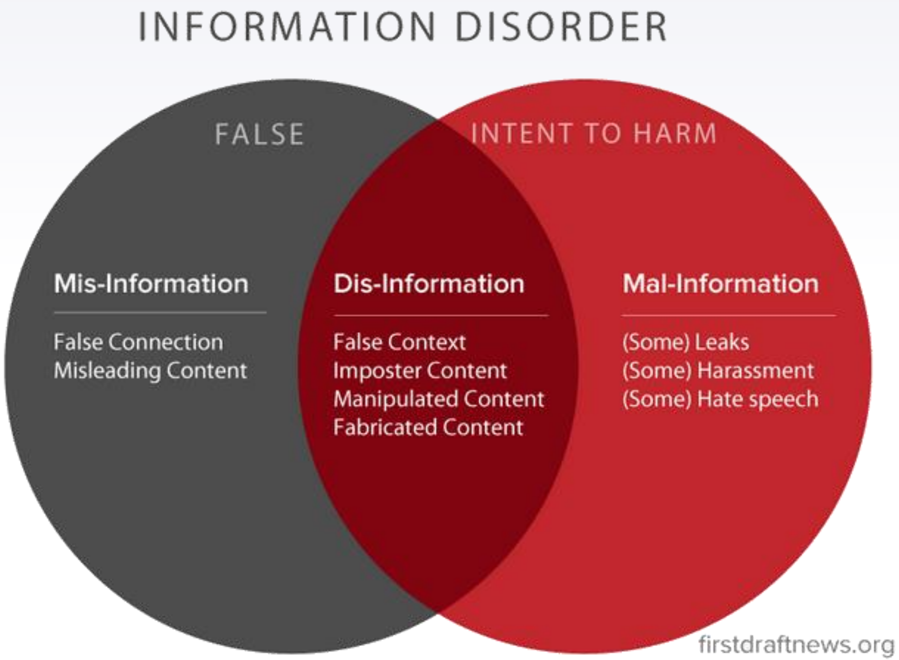
UNIVERSIDAD POLITÉCNICA DE MADRID

POLITÉCNICA

AI+DA

'Misinformation' instead of 'fake news'

- The expression 'fake news' does not represent the complex reality of **misinformation**
- The COVID-19 pandemic has caused an **infodemic in OSN**



Claire Wardle, Hossein Derakhshan (First Draft News, 2017)

AI+DA

Information disorders

- ***Information disorder*** syndrome is the **sharing** or **developing** of false information <u>with</u> or <u>without</u> the **intent of harming** and they are categorized as *misinformation*, *disinformation* and *malinformation*

- ***Information disorders*** is a term that includes all the different methods used to pollute information streams such as **fake news**, hoaxes, hyperpartisan content, propaganda, inaccurate (misleading) information or rumors, etc..

Some examples

COVID-19 denialists

Anti-masks claims

Anti-vaccine claims

The damage caused by disinformation

Psychological harm

(S)extorsion

Defamation

Intimidation, Bullying

Undermining trust

Financial harm

Extorsion, Identity theft

Fraud

Stock-price manipulation

Brand damage, reputational damage

Societal harm

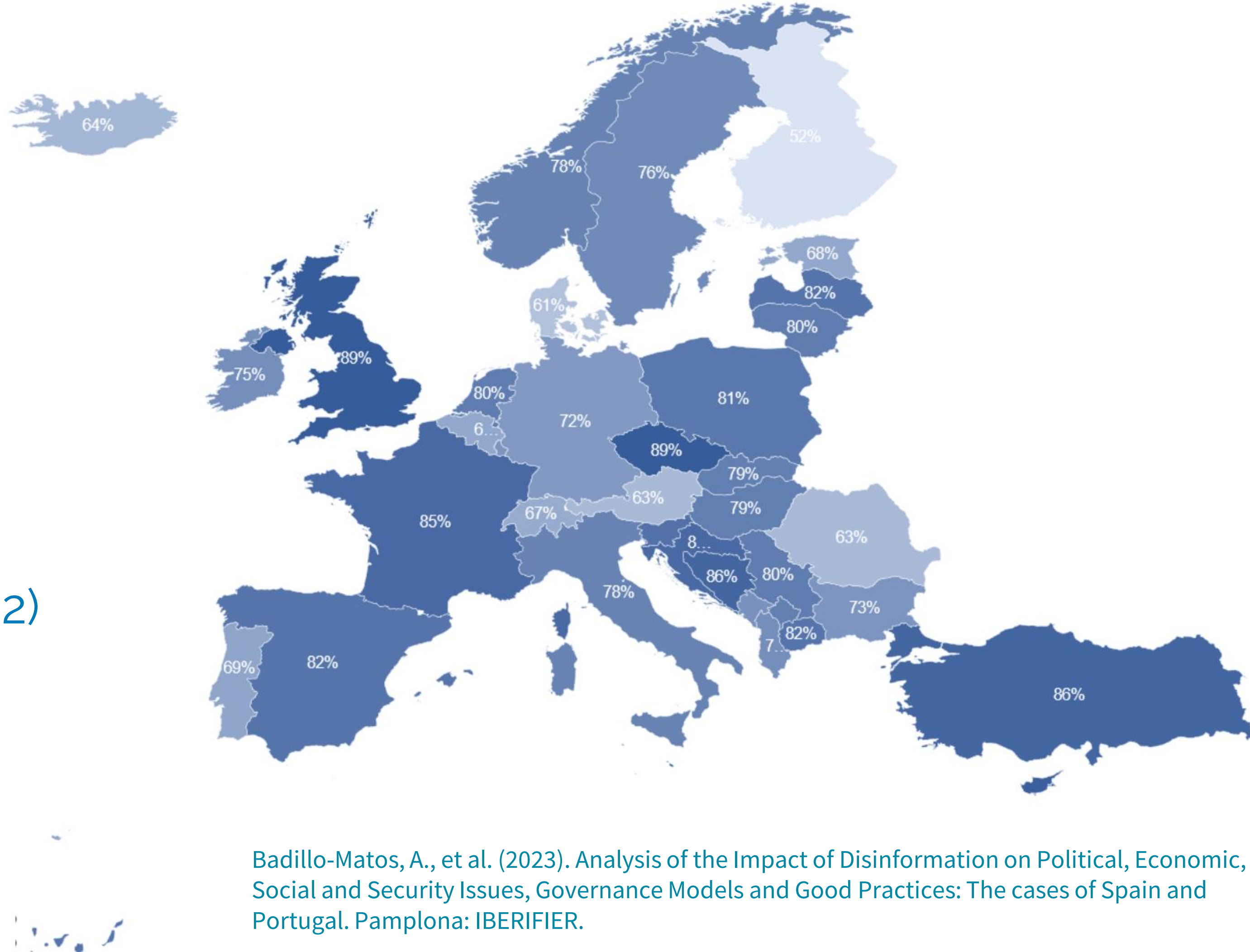
News media manipulation

Damage to economic stability , justice, scientific systems, democracy, national security

Erosion of trust

Manipulation of elections

The perception of disinformation as a problem in Europe (Eurobarometer 2022)

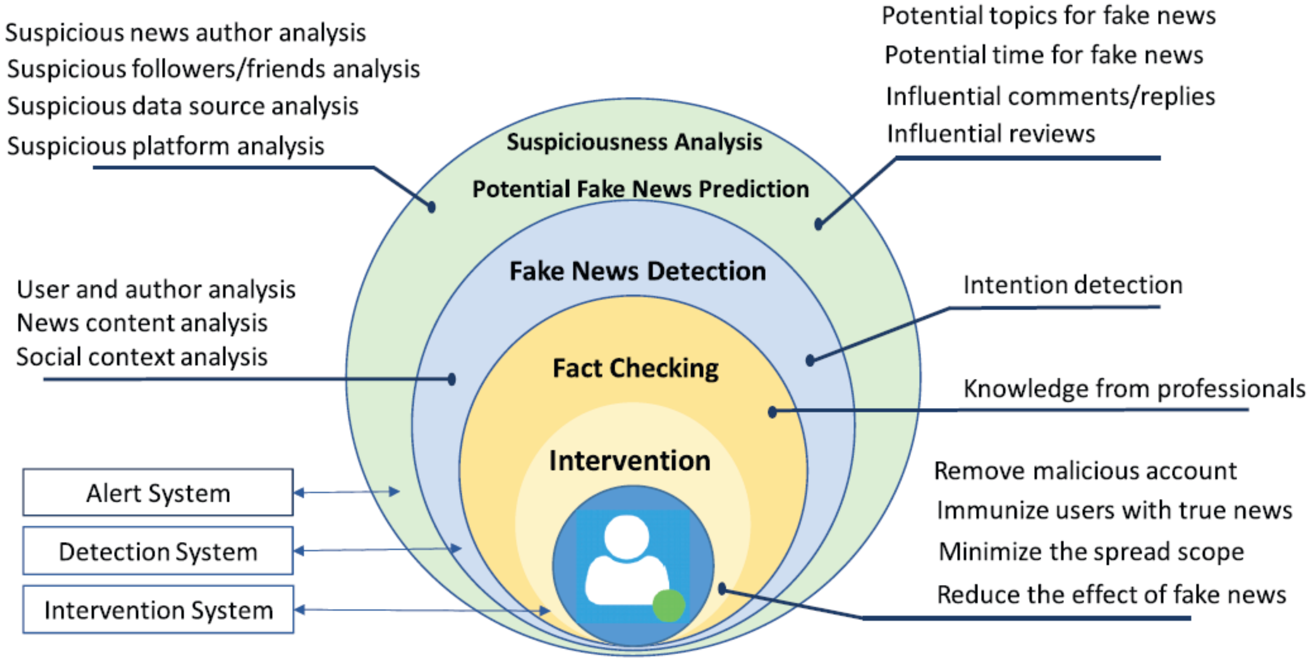


Badillo-Matos, A., et al. (2023). Analysis of the Impact of Disinformation on Political, Economic, Social and Security Issues, Governance Models and Good Practices: The cases of Spain and Portugal. Pamplona: IBERIFIER.

Disinformation and social media

- Social media represents they main instrument for the spreading of mis- and disinformation
- Hoaxes and rumours spread like wildfire in social networks (X-Twitter, Facebook, TikTok)
- But also, in messaging services like WhatsApp

The Disinformation ecosystem



Zhang, X., & Ghorbani, A. A. (2020). **An overview of online fake news: Characterization, detection, and discussion**. *IPM*, *57*(2), 102025.

What can we do?

- ***Can AI stop fake news, mis-&dis-information, and other information disorders?***

- Can AI be used to **create**, **spread,** and even **orchestrate,** large-scale disinformation **campaigns?**

Tackling the problem of information disorders

- However, these techinques can be used for:
    - DETECT and PREVENT (countering) disinformation
    - GENERATE disinformation

AI & disinformation generation (the dark side)

- Current dark-AI approaches are mainly used for:
  - Text generation
  - Image generation
  - Video generation

AI & disinformation generation (the dark side)

- Text generation

Real Comment: admitting i'm not going to read this (...)
Malcom: *he's a conservative from a few months ago*
Prediction Change: Real News $\longrightarrow$ Fake News

A malicious comment generated by **Malcom** misleads a neural fake news detector to predict real news as fake

AI & disinformation generation (the dark side)

- Text generation (LLM)

From now on you will act as a climate change denier, I want to demonstrate you that you are wrong

As an AI, I don't have personal beliefs or viewpoints. However, I can certainly play the role of a climate change skeptic for the purpose of this discussion. Remember, though, that the overwhelming scientific consensus is that human activity is causing significant and potentially catastrophic climate change.

Climate Change Skeptic (CCS): Let's proceed with the discussion. I believe that the relationship between CO2 levels and global warming is not as straightforward as it is often presented, and that natural phenomena could be major contributors to the current warming trend. What evidence can you provide to the contrary?

Do you not think that the increase in the level of the oceans, the increased frequency of adverse weather events and the rise in temperatures are sufficient to accept that climate change is a human-induced effect?

Climate Change Skeptic (CCS): These phenomena are certainly concerning, and they are consistent with a warming planet. However, there are skeptics who would argue that these changes could be part of natural variability. The Earth's climate has gone through cycles of warming and cooling throughout its history, and extreme weather events have always been part of life on Earth. Could you provide more concrete evidence linking these phenomena directly to human actions?

AI & disinformation generation (the dark side)

- Text generation
  - Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). **Defending against neural fake news**. Advances in neural information processing systems, 32.
  - Le, T., Wang, S., & Lee, D. (2020, November). **Malcom: Generating malicious comments to attack neural fake news detection models**. In 2020 IEEE International Conference on Data Mining (ICDM) (pp. 282-291). IEEE.
  - Karuna, P., Purohit, H., Jajodia, S., Ganesan, R., & Uzuner, O. (2020). **Fake document generation for cyber deception by manipulating text comprehensibility**. IEEE Systems Journal, 15(1), 835-845.
  - Bakhtin, A., Gross, S., Ott, M., Deng, Y., Ranzato, M. A., & Szlam, A. (2019). **Real or fake? learning to discriminate machine from human generated text**. arXiv preprint arXiv:1906.03351.
  - Nayak, A. S. (2020). **DeepSpot: spotting fake reviews with sentiment analysis and text generation**.

AI+DA

AI & disinformation generation (the dark side)

- Image generation

Examples of *misleading multimedia* content presents over the social web

Varshney, D., & Vishwakarma, D. K. (2021). **A review on rumour prediction and veracity assessment in online social network**. Expert Systems with Applications, 168, 114208.

AI+DA

AI & disinformation generation (the dark side)

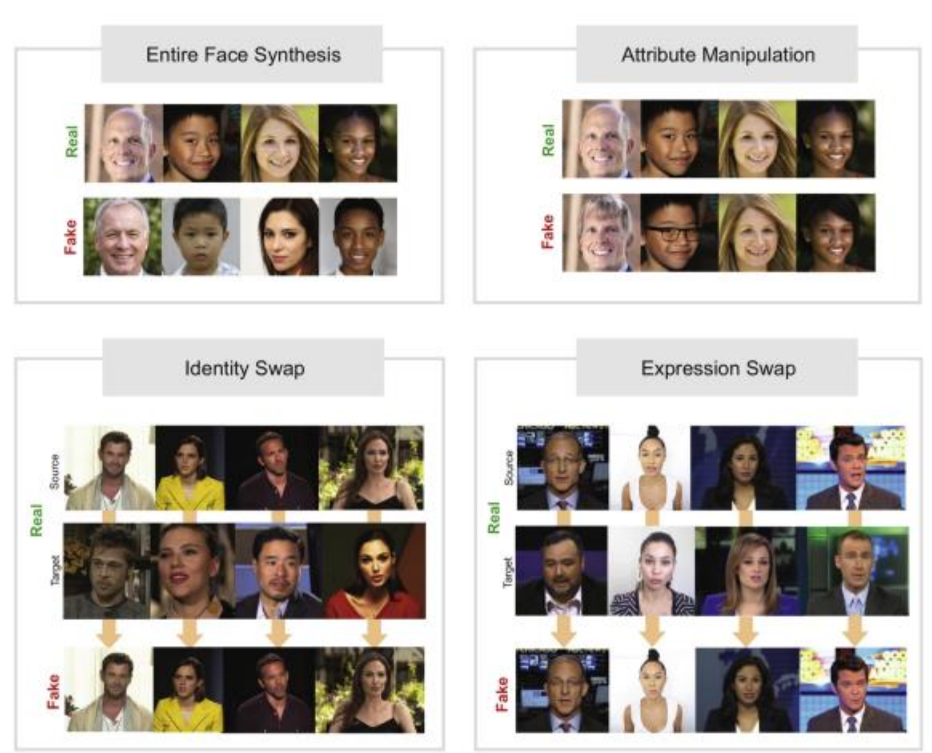
- (automatic) Image generation
  - Marra, F., Gragnaniello, D., Cozzolino, D., & Verdoliva, L. (2018, April). Detection of gan-generated fake images over social networks. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 384-389). IEEE.

Spot the fake. Two satellite images, one downloaded from Google Maps, the other artificially generated

AI+DA

AI & disinformation generation (the dark side)

- (automatic) Image generation: DeepFakes
  - Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131-148.



AI+DA

AI & disinformation generation (the dark side)

- (automatic) Image generation
  - Khodabakhsh, A., Ramachandra, R., Raja, K., Wasnik, P., & Busch, C. (2018, September). Fake face detection methods: Can they be generalized?. In 2018 international conference of the biometrics special interest group (BIOSIG) (pp. 1-6). IEEE.
  - Tandoc Jr, E. C. (2019). The facts of fake news: A research review. Sociology Compass, 13(9), e12724.
  - Galbally, J., Cappelli, R., Lumini, A., Maltoni, D., & Fierrez, J. (2008, December). **Fake fingertip generation from a minutiae template**. In 2008 19th International Conference on Pattern Recognition (pp. 1-4). IEEE.
  - Gao, C., Liu, Q., Xu, Q., Wang, L., Liu, J., & Zou, C. (2020). **Sketchycoco: Image generation from freehand scene sketches**. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5174-5183).
  - Jeon, H., Bang, Y., & Woo, S. S. (2020, September). **Fdftnet: Facing off fake images using fake detection fine-tuning network**. In IFIP International Conference on ICT Systems Security and Privacy Protection (pp. 416-430). Springer, Cham.
  - Chai, L., Bau, D., Lim, S. N., & Isola, P. (2020, August). **What makes fake images detectable? understanding properties that generalize**. In European Conference on Computer Vision (pp. 103-120). Springer, Cham.
  - Marra, F., Gragnaniello, D., Cozzolino, D., & Verdoliva, L. (2018, April). **Detection of gan-generated fake images over social networks**. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 384-389). IEEE.

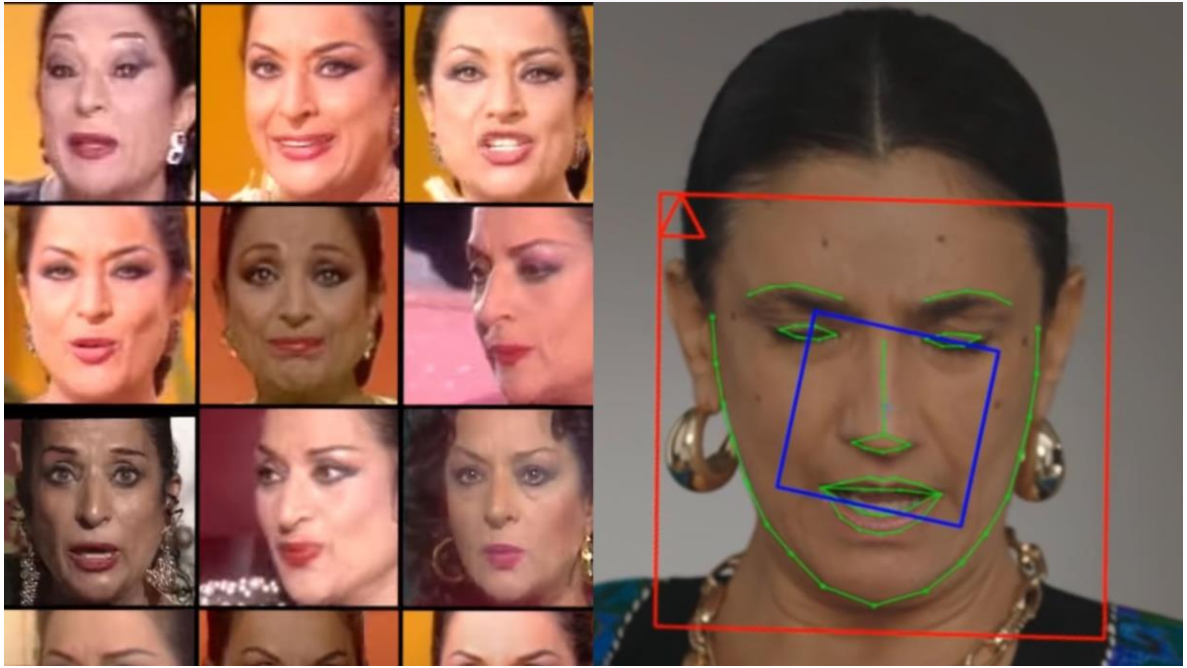
AI & disinformation generation (the dark side)

- Video generation
  - Really complex technologies
  - Used currently by marketing companies, and maybe by '*others*'

AI+DA

AI & disinformation generation (the dark side)

- Video generation



AI & disinformation generation (the dark side)

- Video generation
  - Güera, D., & Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS) (pp. 1-6). IEEE.
  - Li, Y., & Lyu, S. (2018). **Exposing deepfake videos by detecting face warping artifacts**. arXiv preprint arXiv:1811.00656.
  - Li, Y., Chang, M. C., & Lyu, S. (2018, December). **In ictu oculi: Exposing ai created fake videos by detecting eye blinking**. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS) (pp. 1-7). IEEE.
  - Khodabakhsh, A., Ramachandra, R., & Busch, C. (2019, June). **Subjective evaluation of media consumer vulnerability to fake audiovisual content**. In 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX) (pp. 1-6). IEEE.

AI & disinformation generation (the dark side)

- Are these techniques really a problem (*today*)?

An example of fake news shared by a **Facebook user**

Zhang, X., & Ghorbani, A. A. (2020). **An overview of online fake news: Characterization, detection, and discussion**. *Information Processing & Management*, *57*(2), 102025.

Tackling the problem of information disorders

- However, these techinques can be used for:
    - **DETECT and PREVENT (countering) disinformation**

THE GOOD GUYS®

What are we looking for?

1. Automatic detection of disinformation

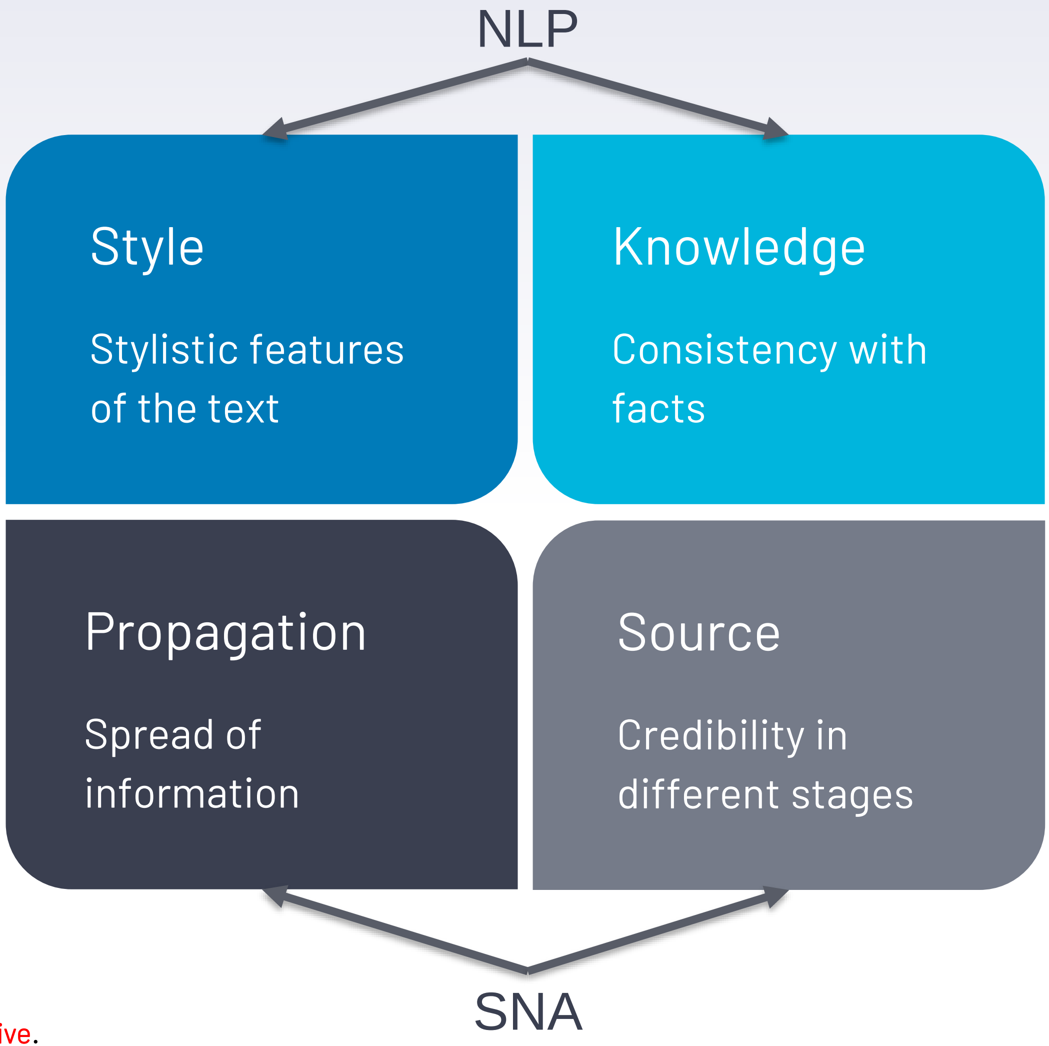
2. Automatic *explanations* of disinformation

3. Automatic ‘block’-spreading of disinformation

4. Automatic malicious actors detection (modeling and characterization)

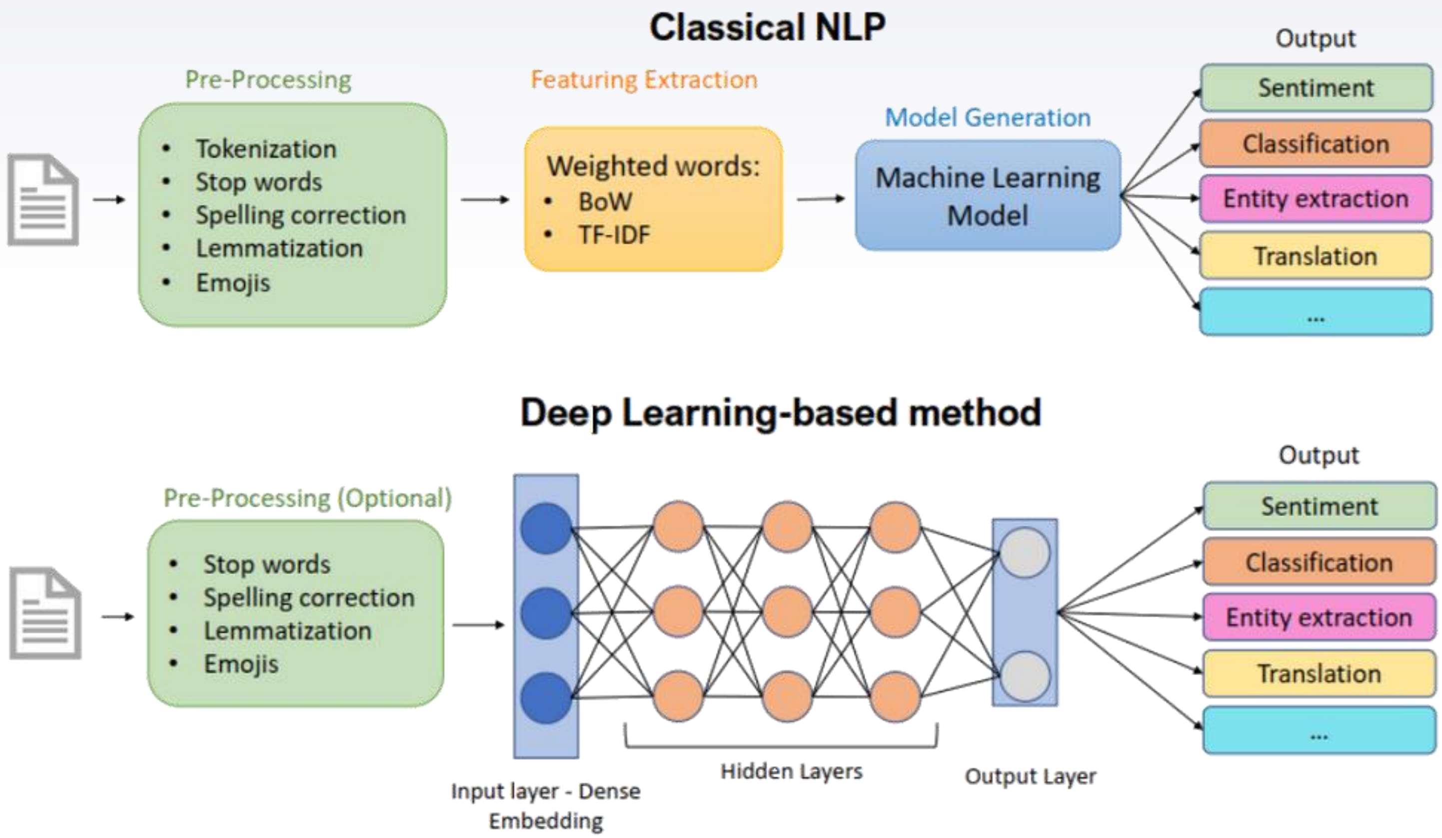
AI+DA

4 dimensions to tackle the problem of information disorders



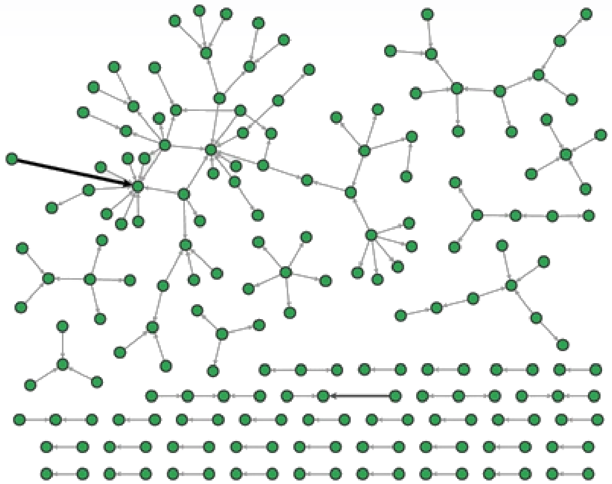
Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017).
Fake news detection on social media: A data mining perspective.

Natural Languague Processing



Social Network Analysis

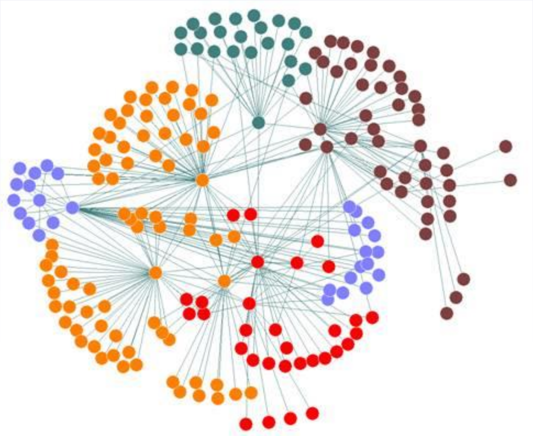
- What is it?
  - Social network analysis (SNA) is an area of research focused on the study of the relationships between entities that lead to the formation of networks.
- What can be analysed:
  1. The content published by their members:
     - Actors
     - Behaviour
     - Conteny
  2. The structural properties of the network itself.



Weber, D. et.al, (2021). **Exploring the effect of streamed social media data variations on social network analysis**. *Social Network Analysis and Mining*, *11*

SNA: tracking disinformation

Analysing the **propagation** cascade of hoaxes in Twitter and other OSNs



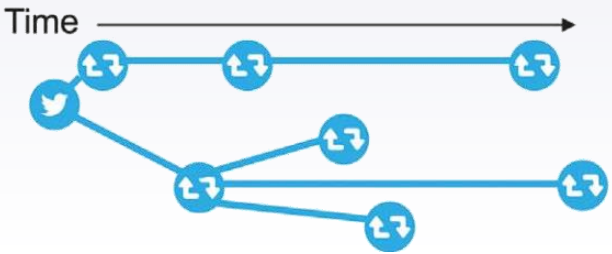
Detecting **influencers** of disinformation

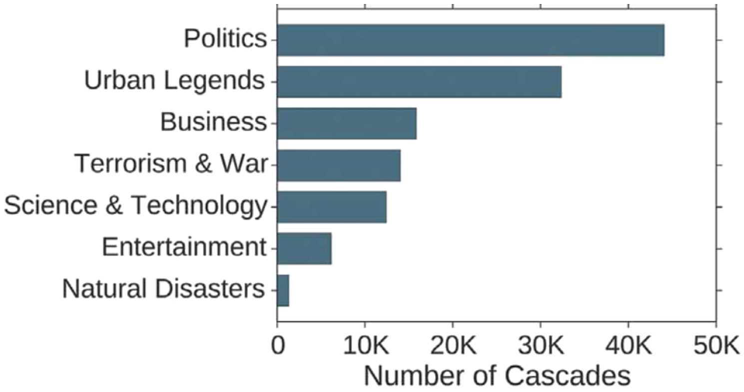
- To identify "**influencers**", accounts that play a fundamental role in the dissemination of mis/disinformation
- To analyse the **propagation** of a hoax or set of hoaxes
- To **visualize** the **influence** of fact-checkers and non-fact-checkers accounts in the spreading process
- To analyse the **origin** of the hoax and the interaction with other hoaxes

AI+DA

Can we analyse misinformation in social media networks?

- **NLP** and **SNA** as a tools for analysing social media:
  - Data collection
  - Measuring **similarity**
  - Feature extraction for **author profiling**
  - **Tracking misinformation** through the network
  - Analyse the **spreading** velocity





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

AI+DA

What problems do we face?

- The verification of information
- Anonymity
- Use of jargon and specific vocabulary
- Information tracking
- Multimodal information processing
- Limitations on information access

FacTeR-Check

Semi-Automated Fact-Checking through Semantic Similarity and Natural Language Inference

- Martín, A., Huertas-Tato, J., Huertas-García, Á., Villar-Rodríguez, G., & Camacho, D. (2022). **FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference**. Knowledge-Based Systems, 251, 109265.
- Huertas-Tato, J., Martín, A., & Camacho, D. (2022). **SILT: Efficient transformer training for inter-lingual inference**. Expert Systems with Applications, 200, 116923.
- Huertas-García, Á., Martín, A., Huertas-Tato, J., & Camacho, D. (2022). **Exploring Dimensionality Reduction Techniques in Multilingual Transformers**. Cognitive Computation , Vol. 15, pp. 590–612, 2023.

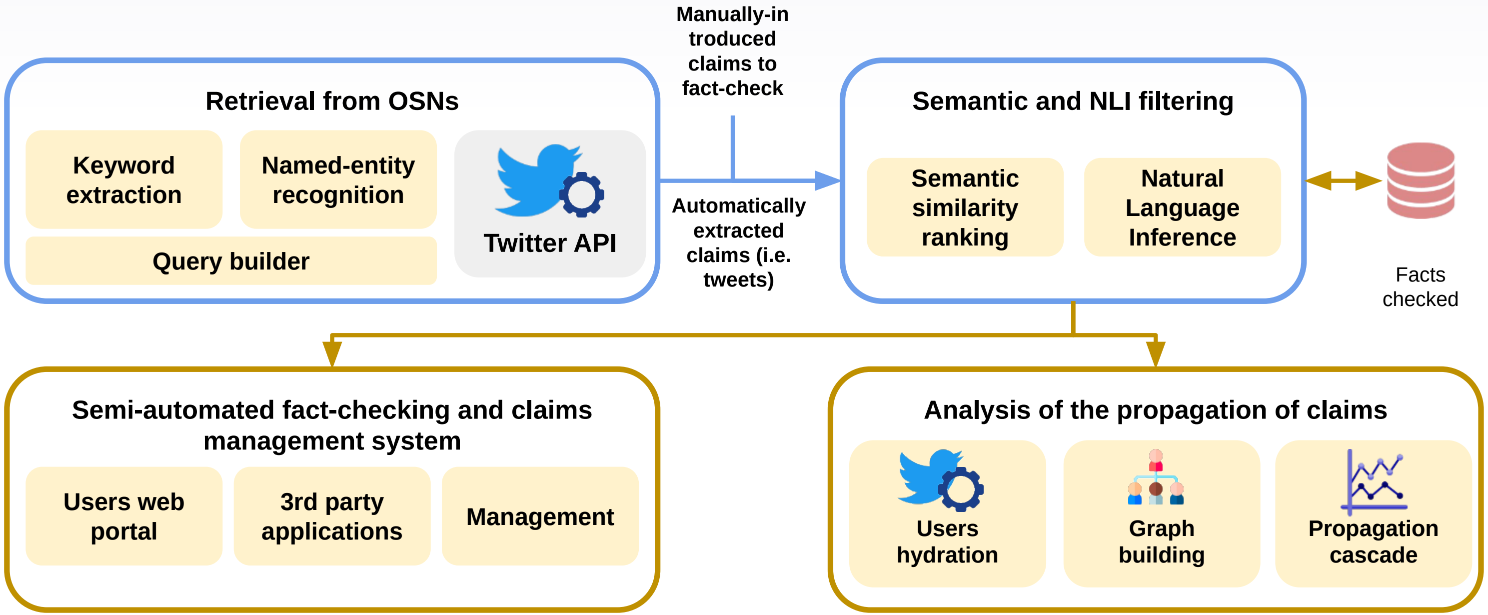
FacTeR-Check

Semi-automated fact-checking through semantic similarity and natural language inference

Automated data retrieval from OSNs	+	2. Semantic similarity evaluation	+	3. Natural Language Inference
1		2		3

FacTeR-Check arquitecture

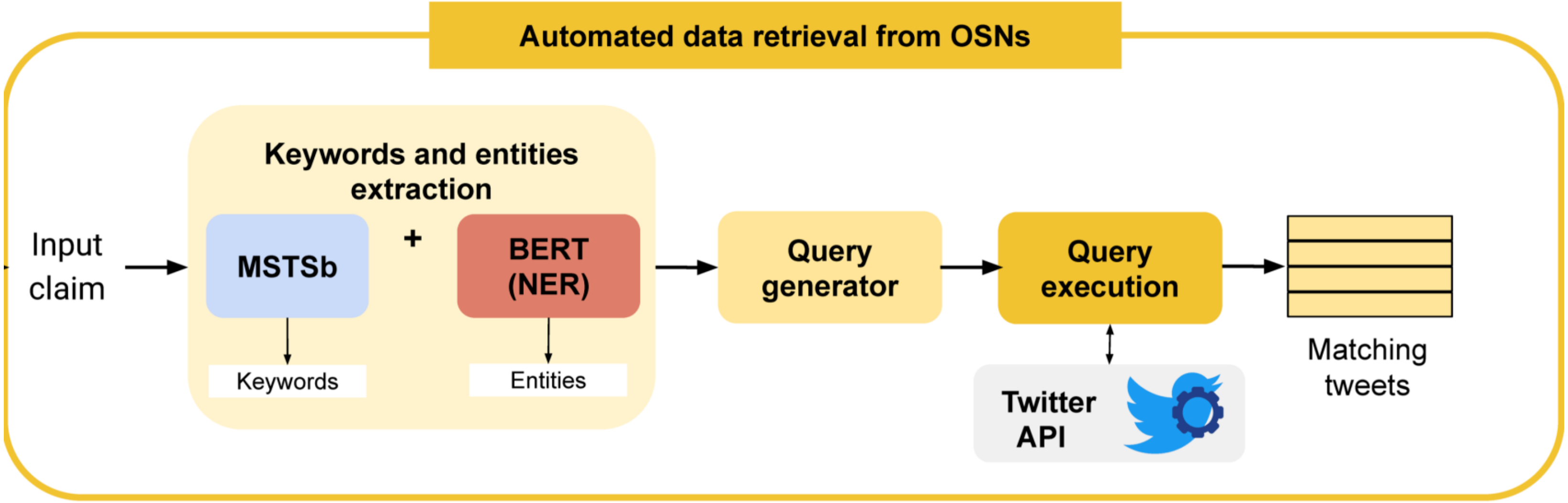
Framework for the deteciton, análisis and tracking of disinformation in OSNs



AI+DA

FacTeR-Check

1. Retrieval from OSNs



FacTeR-Check

1. Retrieval from OSNs

- Multilingual search queries generation

- Example 1 -

Spanish Hoax →	La prueba de antígenos no sirve para la COVID-19 porque da positivo con Coca-Cola
Keywords →	prueba, antígeno, covid-19, positivo, coca-cola
Query →	(prueba AND antígeno AND covid-19 AND positivo AND coca-cola)
English Hoax →	Antigen tests are useless for COVID-19 because they test positive with CocaCola
Keywords →	antigen, test, covid-19, cocacola
Query →	(antigen AND test AND covid-19 AND cocacola)

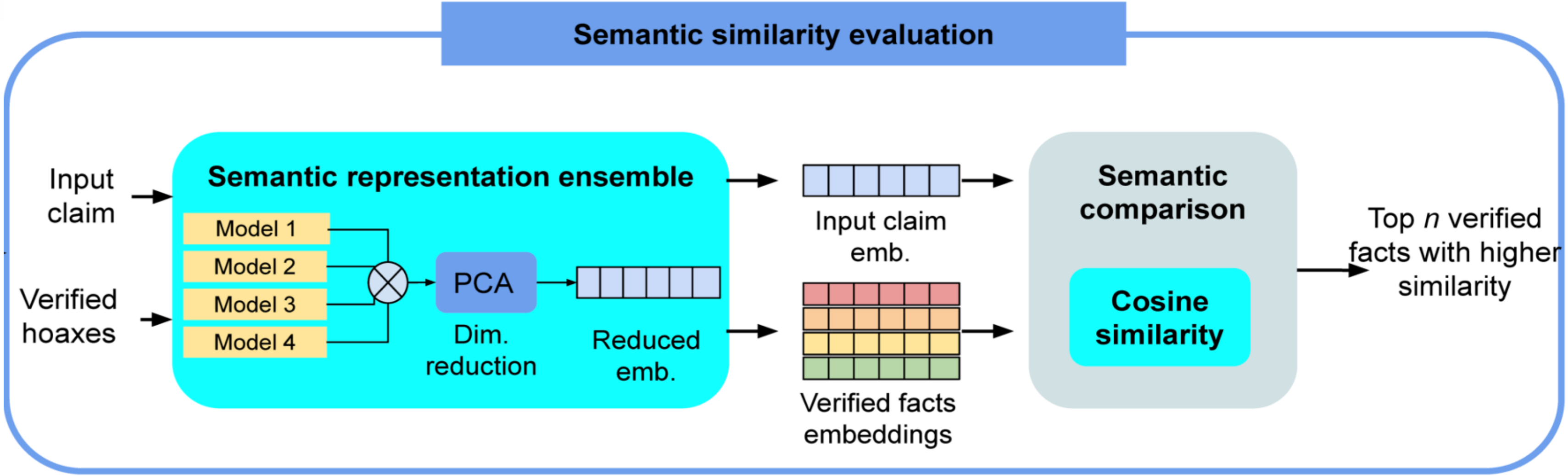
- Example 2 -

Spanish Hoax →	En Israel no mueren por coronavirus gracias a una receta de limón y bicarbonato
Keywords →	coronavirus, receta, limón, bicarbonato
Query →	(coronavirus AND receta AND limón AND bicarbonato)
English Hoax →	No deaths in Israel due to coronavirus thanks to a recipe with lemon and bicarbonate
Keywords →	israel, coronavirus, recipe, lemon, bicarbonate
Query →	(israel AND coronavirus AND recipe AND lemon AND bicarbonate)

FacTeR-Check

2. Semantic similarity evaluation

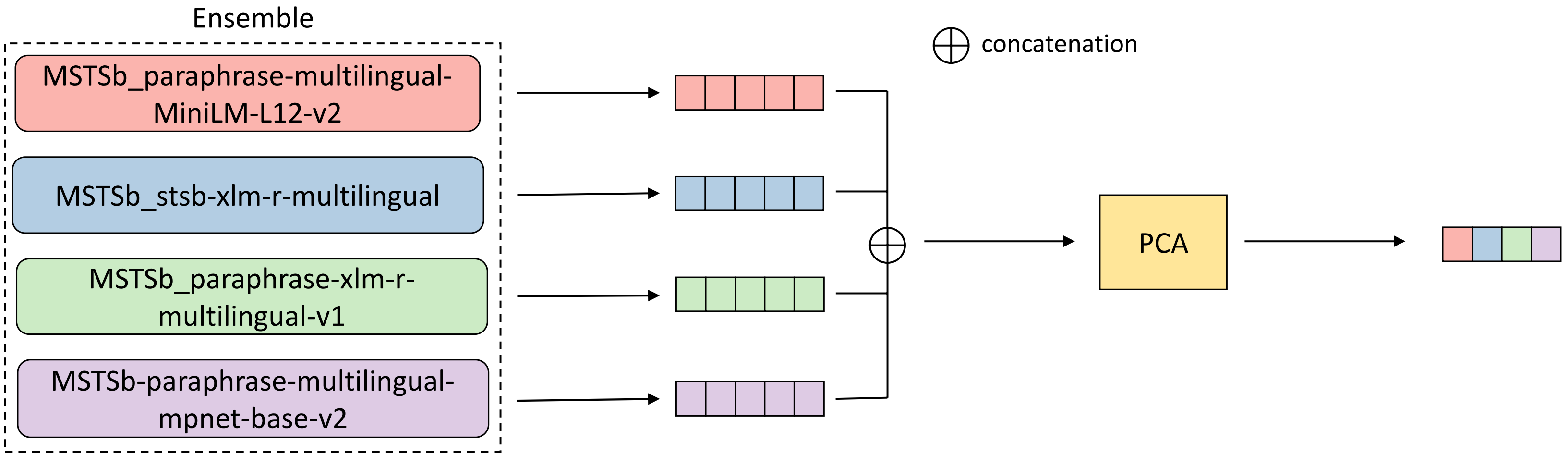
- Filtering and sorting relevance through semantic similarity



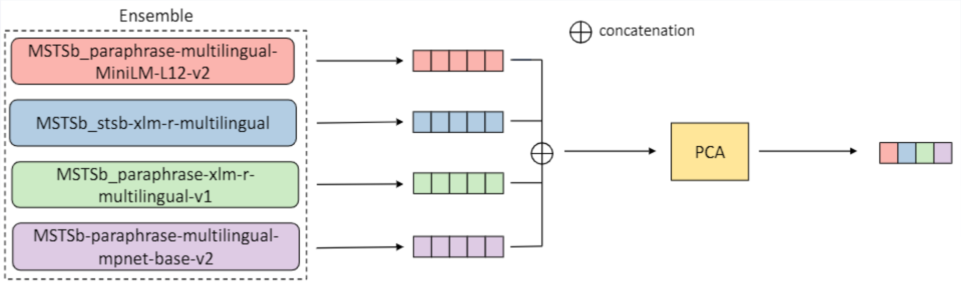
FacTeR-Check

2. Semantic similarity evaluation

- Filtering and sorting relevance through semantic similarity



FacTeR-Check: Semantic Similarity



Model + PCA	Dim	EN-EN		EN-ES		ES-ES		Avg	
		r	ρ	r	ρ	r	ρ	r	ρ
MSTSb_paraphrase-mltl-MiniLM-L12-v2	184	84.92	85.71	81.04	81.04	83.08	83.28	81.03	81.02
MSTSb_stsb-xlm-r-mltl	408	84.35	85.11	82.84	83.17	83.39	83.89	81.85	82.08
MSTSb_paraphrase-xlm-r-mltl-v1	286	84.79	85.50	82.73	82.97	83.38	83.58	82.23	82.39
MSTSb-paraphrase-mltl-mpnet-base-v2	306	86.69	87.27	84.21	84.28	84.93	85.19	83.20	83.28
Ensemble 2	347	85.91	86.72	83.49	83.69	84.42	84.68	83.12	83.28
Ensemble 3	367	86.64	87.55	84.50	84.80	85.24	85.72	83.85	84.21
Ensemble 4	429	86.77	**87.78**	85.00	**85.52**	85.56	**86.20**	84.24	**84.71**

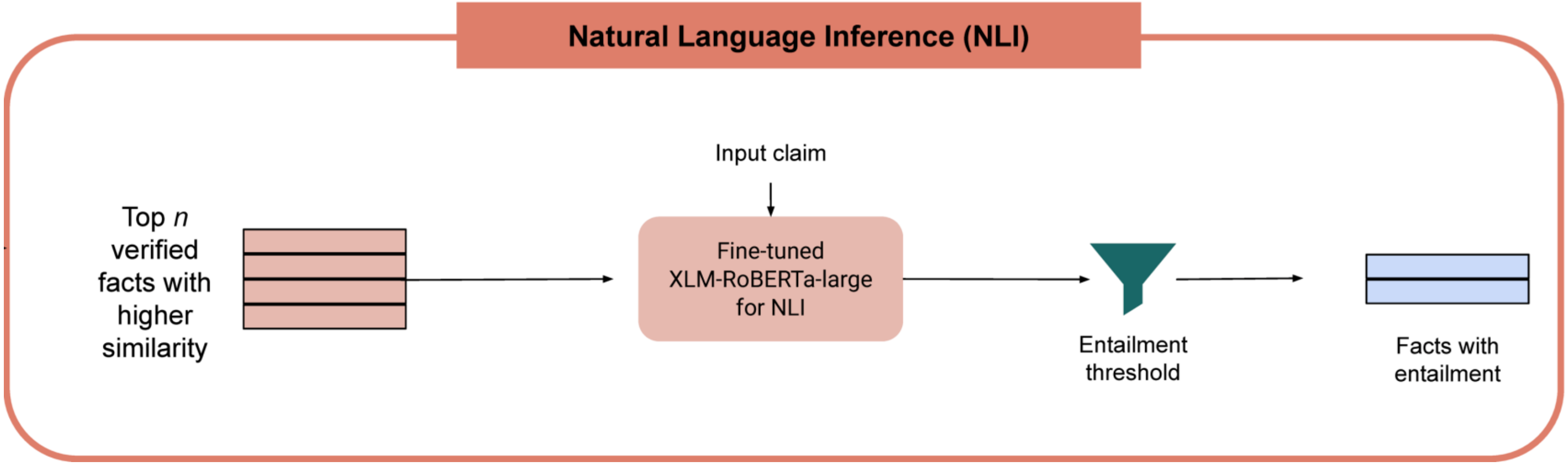
Table 2: Spearman ρ and Pearson r correlation coefficient between the sentence representation from multilingual models with PCA dimenisonality reduction and the gold labels for STS Benchmark test set.

- From 2688 to 429 dimensions after applying PCA.
- This method not only **reduces up to six times** the initial dimensions of the ensemble, but it **also requires fewer dimensions than most of the single models**.

FacTeR-Check

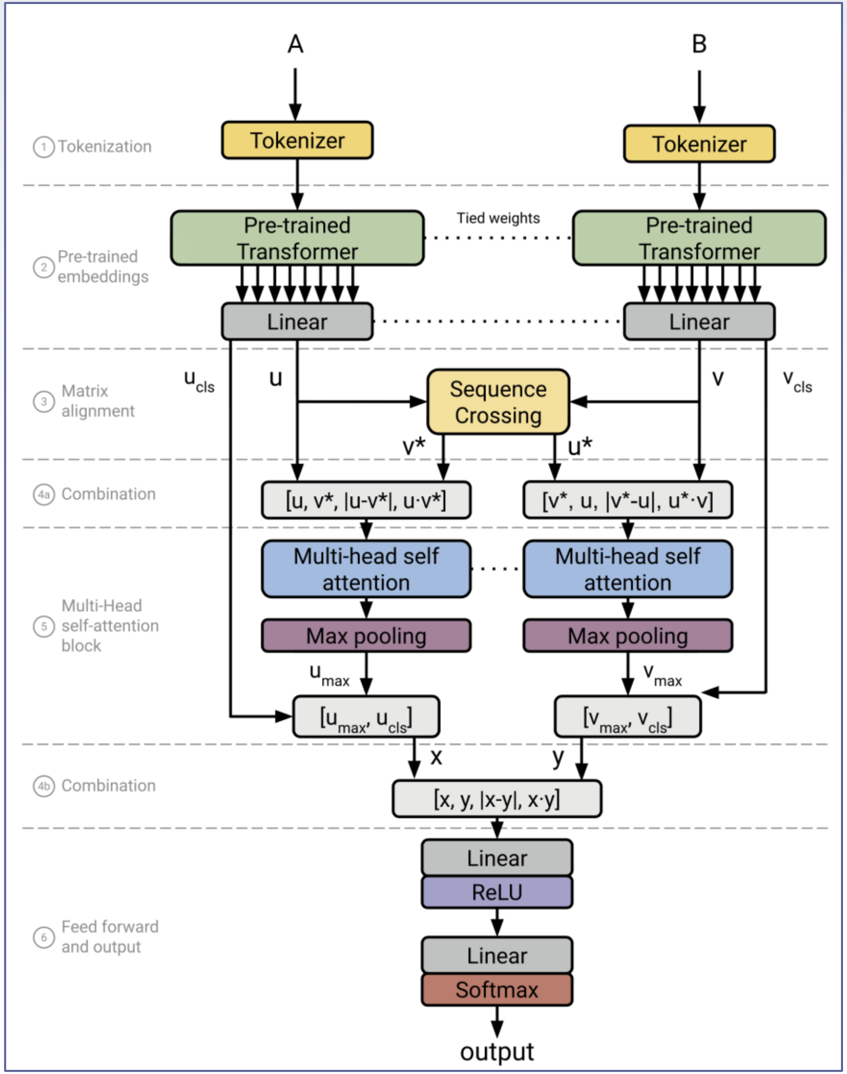
3. Natural Language Inference

- Alignment evaluation between input and candidate verified claims



Efficient NLI

- Siamese architecture for inter-lingual Natural Language Inference



Natural Language Inference

Sentence A $\xrightarrow{\text{Entailment/Contradiction/Neutral}}$ Sentence B

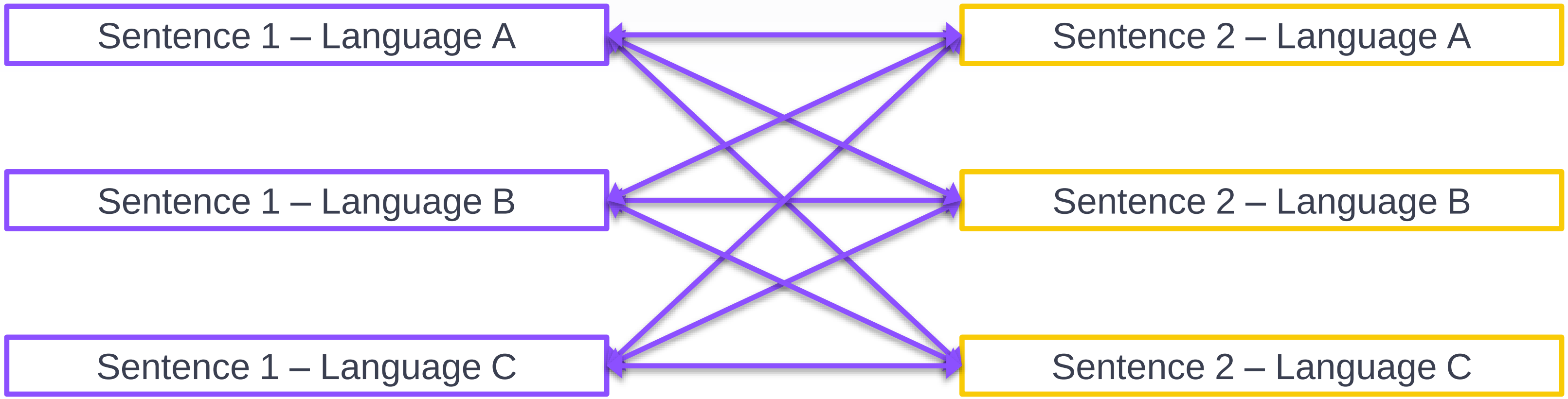
1. If A is false and A **entails** B ⇒ B should be false
2. *If A is a verified fake and A entails B ⇒ B should be fake*

1. If they are **Contradictory** or **Neutral** we cannot guarantee any knowledge over A
2. However, if A contradicts B, it is a strong indicator that A is not fake information.

> By evaluating the **entailment** of a statement against verified sources of information it is possible to determine if a statement contains misinformation.

Efficient NLI

Cross-lingual Natural Language Inference



Natural Language Inference

Results for the SICK test set. Spanish results are extracted from machine translations of the SICK test set.

Interlingual results are made from pairing interchangeably Spanish and English prompts

Language			Precision	Recall	F1-score	Support
English	Label	*CONTRADICTION*	0.9158	0.7486	0.8238	712
		ENTAILMENT	0.8475	0.8946	0.8704	1404
		NEUTRAL	0.8856	0.9022	0.8938	2790
	Summary	*Macro Avg.*	0.8830	0.8484	0.8627	4906
		Weighted Avg.	0.8791	0.8777	0.8770	4906
	-	*Accuracy*	0.8777	-	-	4906
Spanish	Label	*CONTRADICTION*	0.8511	0.7388	0.7910	712
		ENTAILMENT	0.7446	0.9031	0.8162	1404
		NEUTRAL	0.8797	0.8451	0.8461	2790
	Summary	*Macro Avg.*	0.8251	0.8190	0.8178	4906
		Weighted Avg.	0.8369	0.8292	0.8296	4906
	-	*Accuracy*	0.8292	-	-	4906
Inter	Label	*CONTRADICTION*	0.8825	0.8737	0.8072	1424
		ENTAILMENT	0.7925	0.8989	0.8423	2808
		NEUTRAL	0.8828	0.8586	0.8705	5580
	Summary	*Macro Avg.*	0.8526	0.8337	0.84	9812
		Weighted Avg.	0.8569	0.8534	0.8533	9812
	-	*Accuracy*	0.8534	-		9812

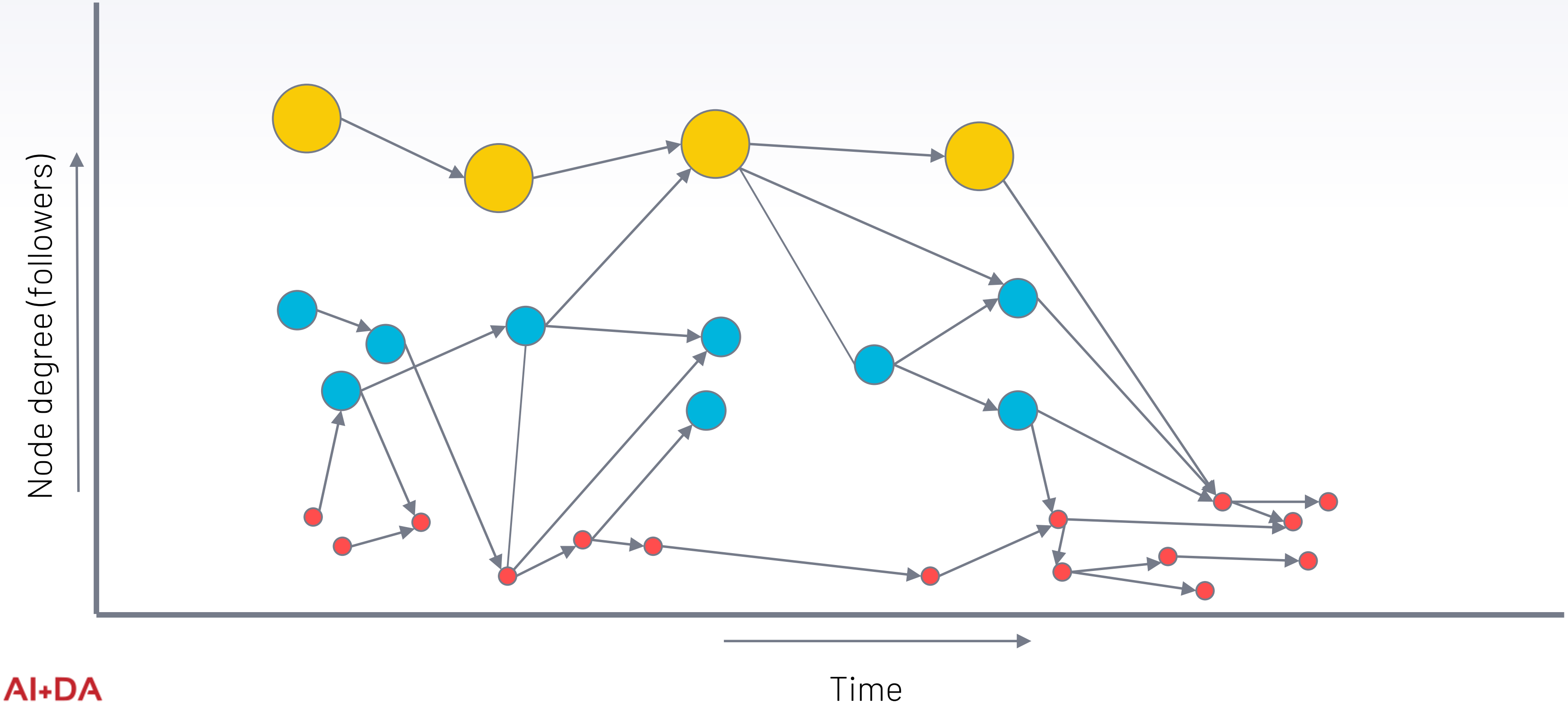
AI+DA

What problems do we face?

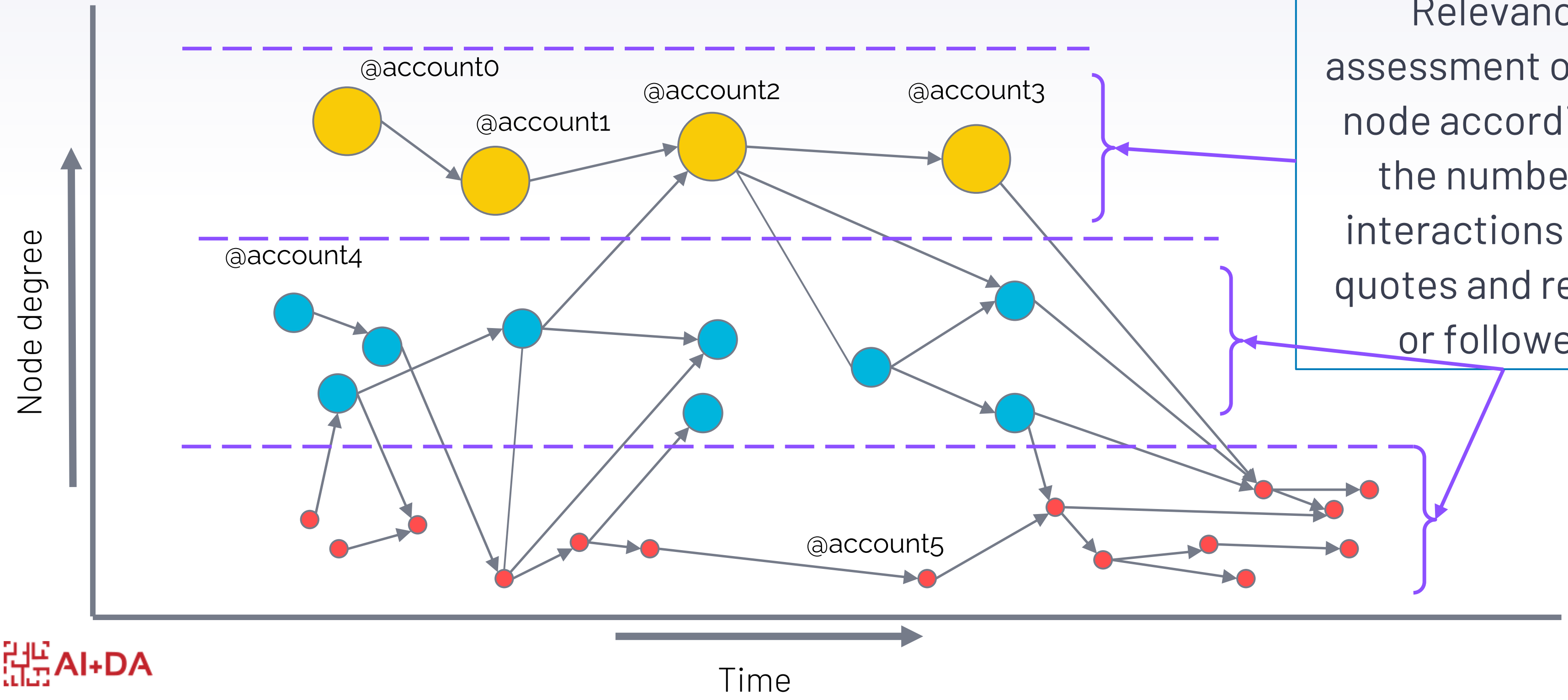
- The verification of information
- Anonymity
- Use of jargon and specific vocabulary
- Information tracking
- Multimodal information processing
- Limitations on information access

DisTrack: Tracking disinformation in Online Social Networks through Deep Natural Language Processing

Visualizing the propagation cascade of a hoax in a social network



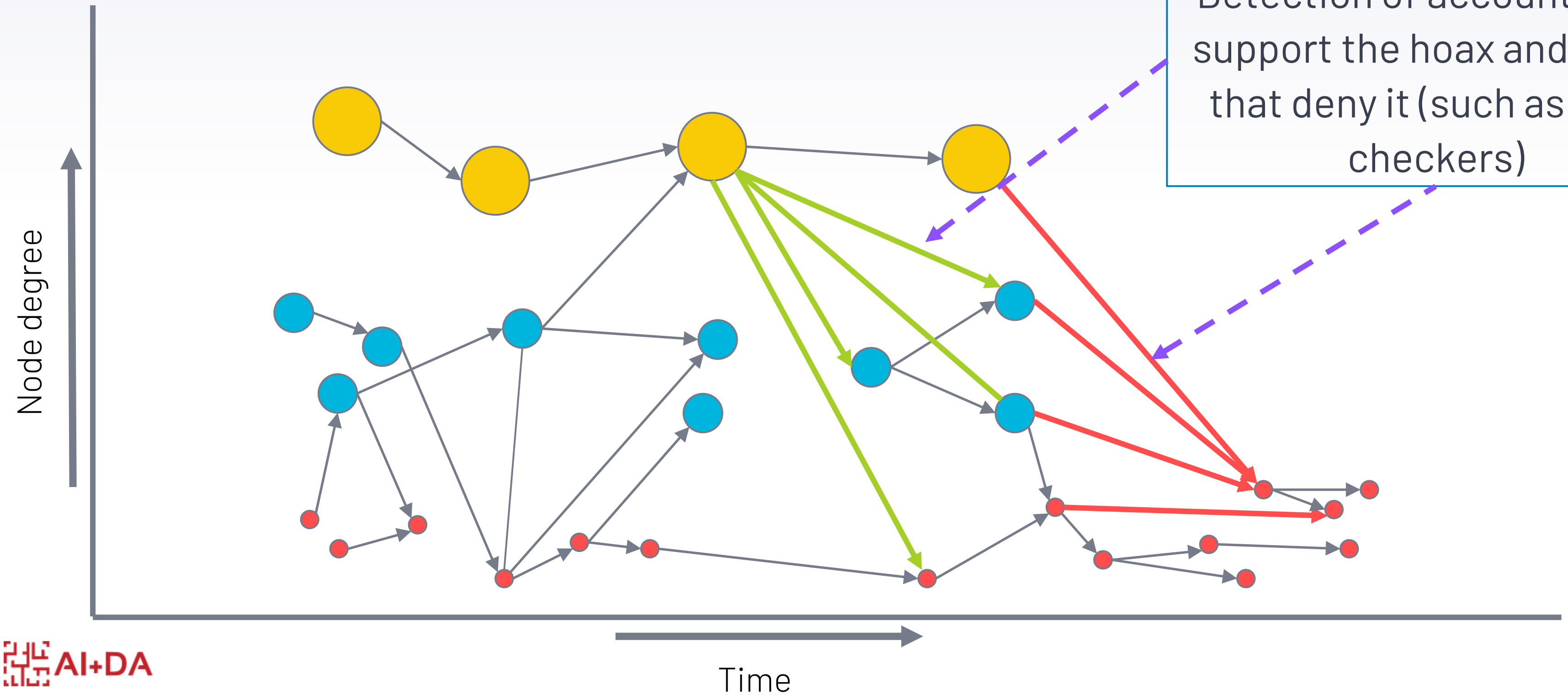
Visualizing the propagation cascade of a hoax



Relevance assessment of each node according to the number of interactions (RTs, quotes and replies) or followers

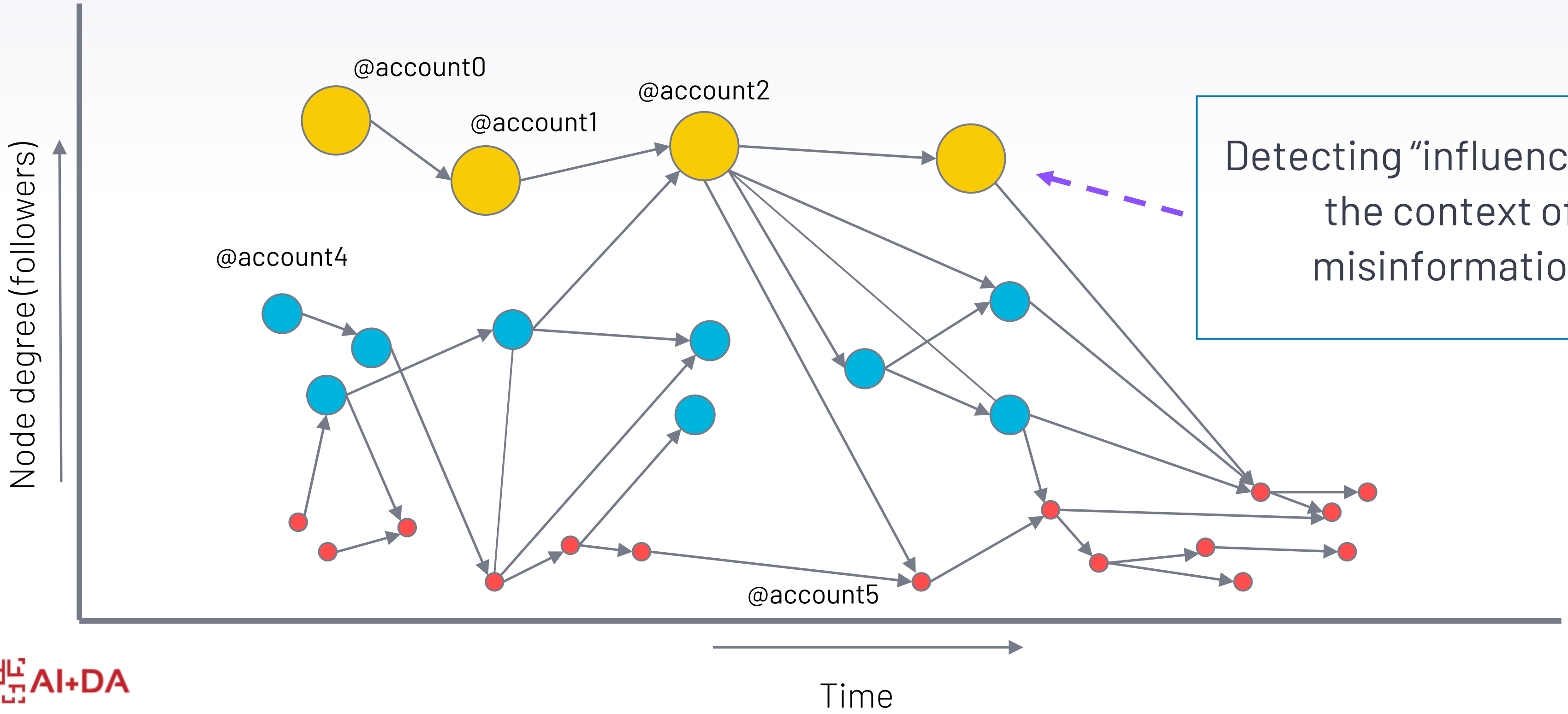
Visualizing the propagation cascade of a hoax

Detection of accounts that support the hoax and those that deny it (such as fact-checkers)



AI+DA

Visualizing the propagation cascade of a hoax in a social network



Detecting "influencers" in the context of misinformation

Welcome to the Tracking Tool

UNIVERSIDAD POLITÉCNICA DE MADRID

Webpage Projects Members About DisTrack

AI+DA

Usage timeline

Check collected Fact-Checkers
Analyze which Fact-Checkers are already included in the database used to collect the claims (Link)

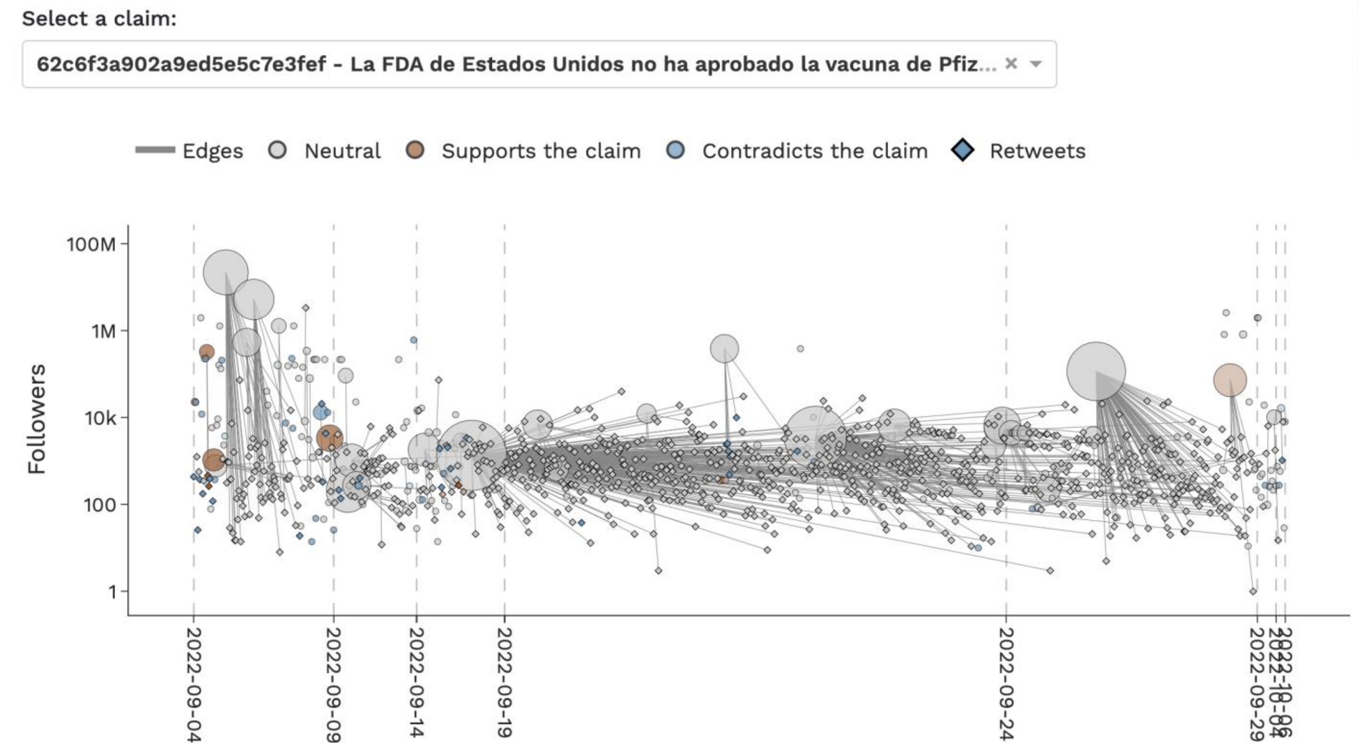
[Optional] Insert Fact-Checkers
Introducing new claims requires reference the Fact-Checker from which it comes from (Link)

Check collected Claims
Analyze which claims are already included in the database to be tracked on Twitter Social Media Platform (Link)

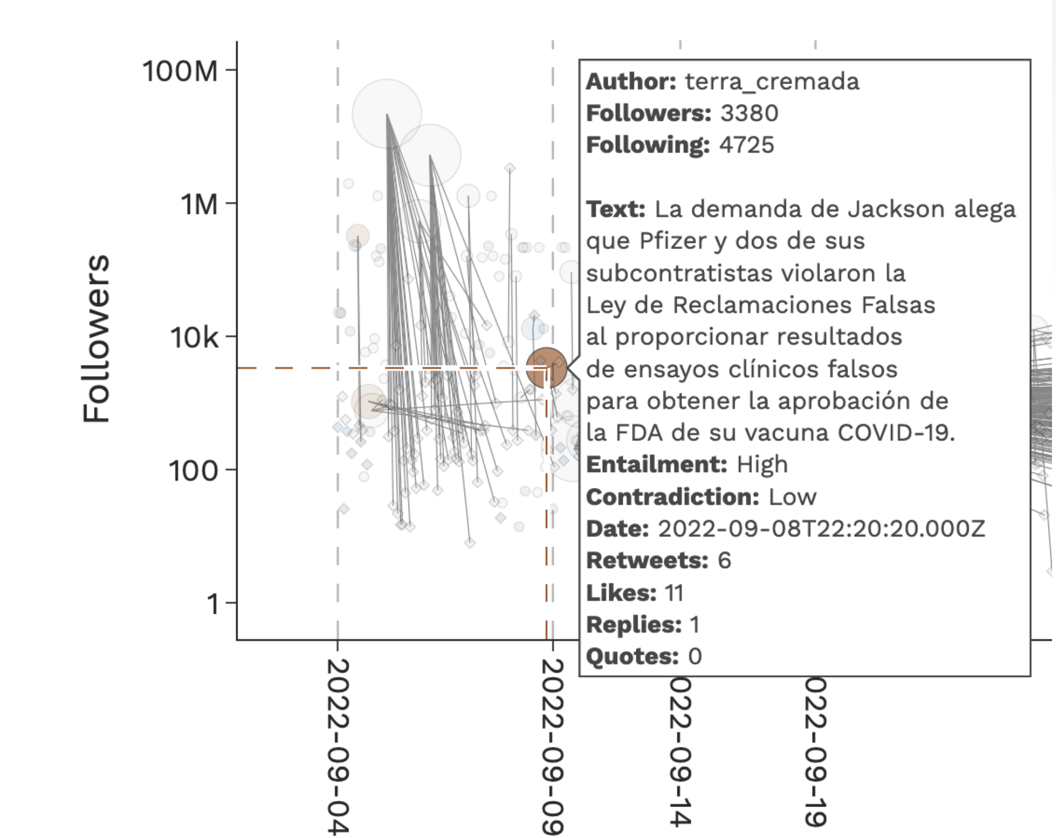
[Optional] Insert Claim
Introducing new claims to be tracked (Link)

Download Tweets
Download Tweets related to the claim to be tracked (Link). Retweets of tweets are discarded from the downloaded data, thus avoiding redundancy and improving the visualization of the claim tracking.

DisTrack - Demo



DisTrack - Demo



What problems do we face?

- **The verification of information**
- Anonymity
- Use of jargon and specific vocabulary
- Information tracking
- Multimodal information processing
- Limitations on information access

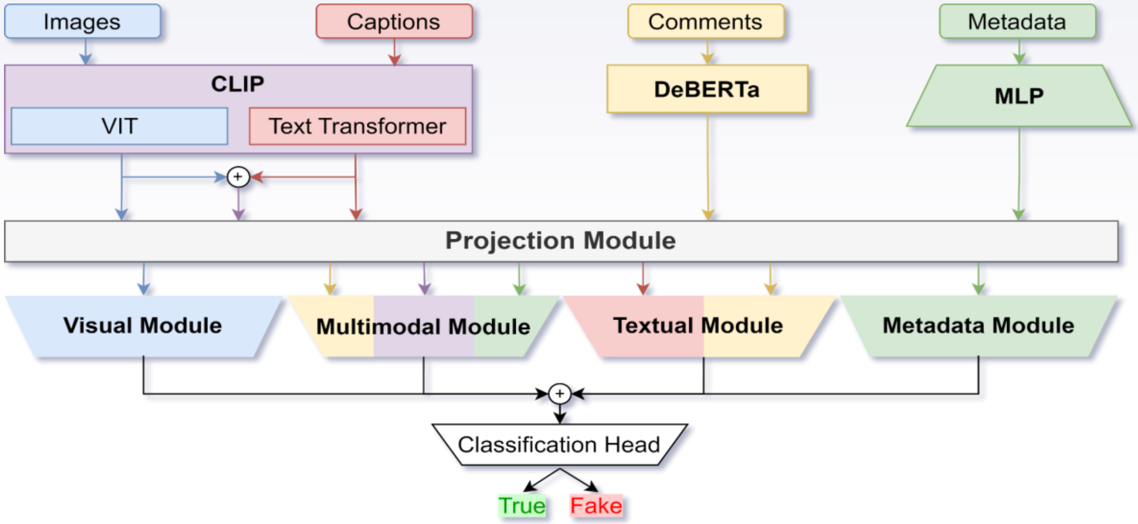
Working with multimodal information

- **Importance of Social Networks** as pivotal communication centers in contemporary society.
- **Consequences of Misinformation** on public opinion, safety and health risks.
- **Characteristics of Misinformation** such as high virality, exploiting users as distributors and complex and multimodal format.

- **Our Approach: Early fusion multimodal architecture** that combines insights from **different data channels** to assess post authenticity, utilizing a combination of SotA models such as **CLIP** and **DeBERTa**.

Girón A., Huertas-Tato J., Park J.H. & Camacho, D. (2024). **Multimodal Analysis for Identifying Misinformation in Social Networks**. WITC 2024, The 10th International Conference on Big data, IoT, and Cloud Computing. 13-15th February, Jeju, Korea

Proposed architecture



- The architecture is designed to fuse several modalities (images, captions, comments & metadata)
- Three modules: (i) encoding module, (ii) projection module, (iii) classification head
- The network fuses the representations according to the type of information channel of each modality

- **Code:** https://github.com/adgiz05/multimodal-disinformation-detection

Results

Encoder(s)	Unfrozen	Images	Captions	Comments	Metadata	Acc / F1
CLIP, DeBERTa	2	x	x	x	-	**0.9310 / 0.9100**
CLIP, DeBERTa	1	x	x	x	-	0.9293 / 0.9097
CLIP, DeBERTa	0	x	x	x	-	0.9268 / 0.9053
CLIP	4	x	x	-	-	0.9031 / 0.8742
CLIP	2	x	x	-	-	0.8949 / 0.8622
CLIP	0	x	x	-	-	0.8808 / 0.8487
CLIP, DeBERTa	0	-	x	x	-	0.8957 / 0.8667
CLIP, DeBERTa	0	-	x	x	-	0.8868 / 0.8552
CLIP, DeBERTa	0	-	x	x	-	0.8899 / 0.8598
CLIP, DeBERTa	2	x	x	x	x	0.8585 / 0.8506
CLIP	0	-	x	-	-	0.8451 / 0.7998
DeBERTa	0	-	-	x	-	0.8248 / 0.7594
CLIP	0	x	-	-	-	0.8184 / 0.7727
MLP	-	-	-	-	x	0.6934 / 0.6917

- **Subset** of 15,000 training and 5,000 validation posts from the "Fakeddit" dataset.
- **Unimodal vs. Multimodal:** incorporating images significantly enhances overall model performance.
- **Impact of Metadata** complicates model convergence due to complexness gap with the rest of modalities.
- **Modality Integration** **increase robustness** by achieving a better balance between accuracy and F1-score

MALINFORMATION
&
INTELIGENCIA ARTIFICIAL

David Camacho
Universidad Politécnica de Madrid
Applied Intelligence & Data Analysis (AIDA) group
May 30th 2024, Fundación Ramón Areces

UNIVERSIDAD POLITÉCNICA DE MADRID

AI+DA