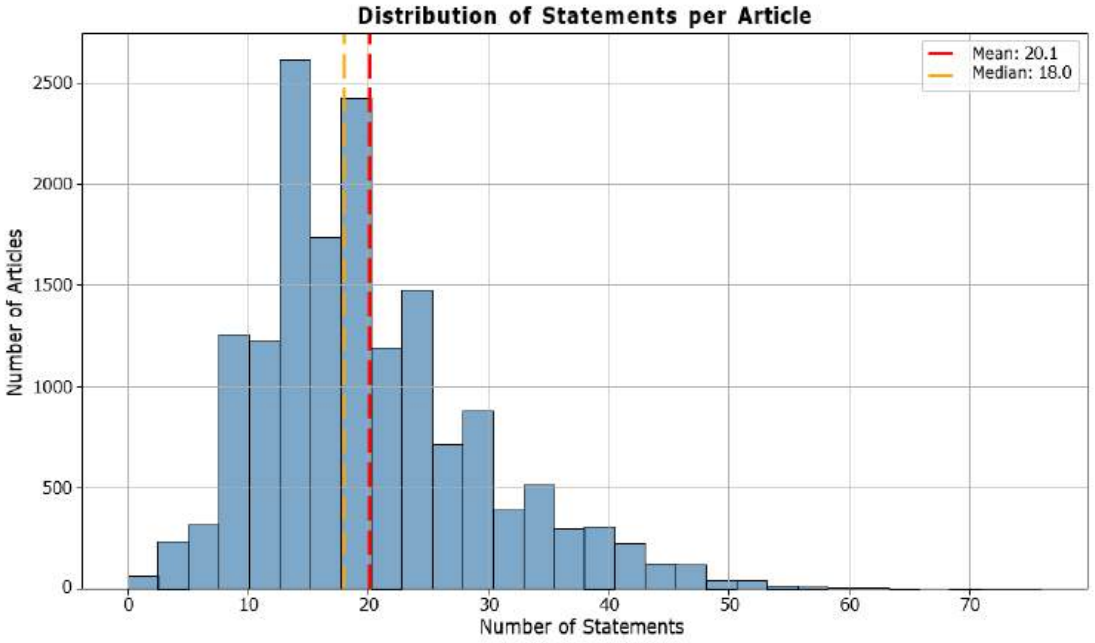
Change rate	17.9%	18.2%

4.3.2 Analysis of stance detection

Of the 326,671 total extracted statements, 267,712 (82%) successfully matched with relevant fact-checks and received stance classifications.

The extraction of claim statements from news articles shows that on average an article contains 20.1 statements (all sources), 21.6 statements (Iberifier digital media list), see Figure 18.

Figure 18: Distribution of statements per article (all sources, n=16,249 articles).

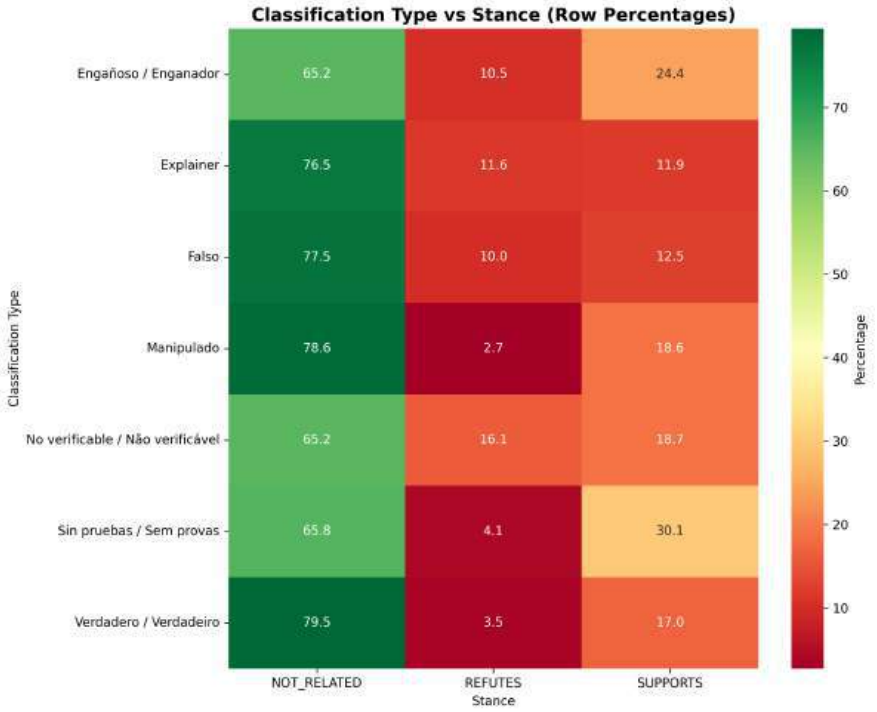


The type of fact-checks retrieved were primarily False (49%), Explainer (26%) and Misleading (15%), see Figure.

Across all sources, 15.8% of statements SUPPORT fact-checks classified as “Falso” (see Figure 19) in legitimate media channels. To examine whether supporting and refuting stances vary in their placement throughout news articles, we analyzed the position of each stance-bearing statement within its article (0% = beginning, 100% = end). This analysis investigates whether media outlets strategically position fact-check-related content in prominent locations (headlines, leads) or bury it in later paragraphs.

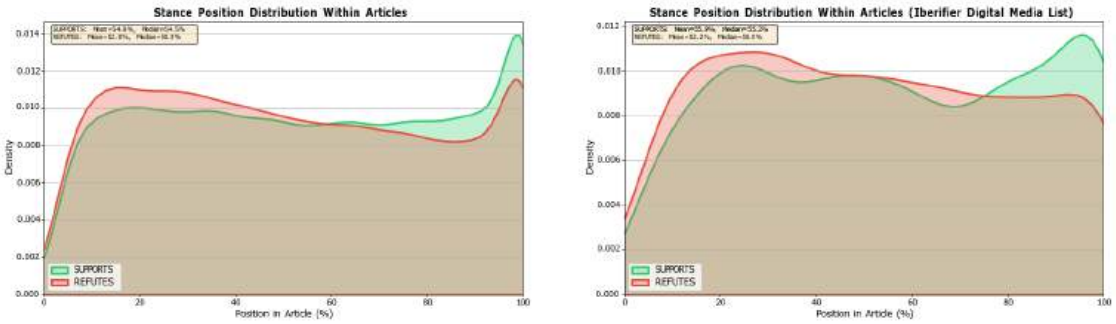
Figure 19: Stance distributions by fact-check qualification type (all sources). Critical finding: 15.8% of statements SUPPORT fact-checks classified as “Falso” (debunked claims), indicating problematic misinformation propagation through

media coverage. Conversely, "Verdadero" (true) claims receive 33.7% support, demonstrating truth amplification.



Analysis of stance positioning reveals uniform distribution throughout article structure, with SUPPORTS stances appearing at a mean position of 53.4% (median 50.0%, SD 31.1%) and REFUTES stances at 52.8% (median 50.0%, SD 30.6%). Both stance types distribute evenly across all article positions, from opening paragraphs through closing sections. This uniform distribution indicates no headline bias: supporting and refuting statements are not preferentially placed in prominent article positions such as headlines, leads, or conclusions. Rather, stances appear integrated throughout article structure, suggesting content integration rather than strategic framing see Figure 20.

Figure 20: Distribution of stance positions within articles. Kernel density estimation shows positioning for SUPPORTS and REFUTES stances, revealing uniform distribution across article structure (mean 53%, median 50%) with no concentration in headlines or conclusions.



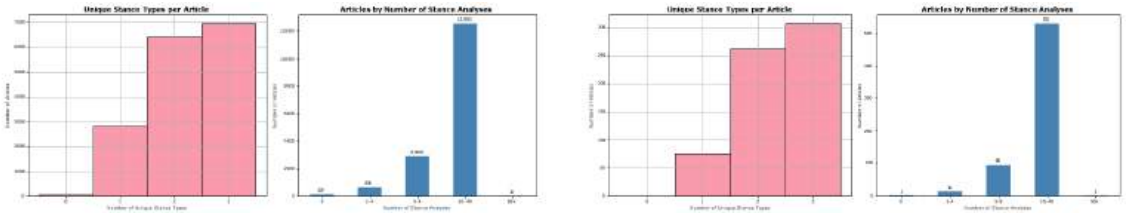
(a) All sources (b) Iberifier digital media list

4.3.3 Article-Level Stance Coverage

To understand how stances are distributed across articles, we examined the number of unique stance types present in each article and the number of stance analyses performed per article. This analysis investigates whether articles contain mixed stances or predominantly exhibit single stance types.

Analysis of article-level stance coverage reveals that most articles contain multiple statements with stance classifications but exhibit limited diversity in stance types. For all sources, the majority of articles contain between 1-4 stance analyses per article, though a substantial number contain 10 or more analyses reflecting longer articles with more claim statements. Regarding unique stance types, most articles exhibit 1-2 different stance types rather than all three possible stances (SUPPORTS, REFUTES, NOT_RELATED). This pattern indicates that while articles may contain multiple analyzed statements, they tend to maintain relatively consistent positioning rather than presenting highly mixed or contradictory stances within the same article. The Iberifier digital media list shows similar patterns, with comparable distributions of stance coverage and unique stance types, suggesting that these article-level structural characteristics are consistent across different media outlet types (see Figure 21).

Figure 21: Article-level stance coverage showing (left) distribution of unique stance types per article and (right) distribution of total stance analyses per article. Most articles contain multiple stance analyses, with the majority exhibiting 1-2 unique stance types, indicating mixed content rather than uniform positioning.



(a) All sources (n=214,062) (b) Iberifier digital media list (n=9,034)

4.3.4 Article-Level Aggregated Patterns

To complement statement-level analysis, we examined article-level aggregate patterns by computing continuous positioning scores for each article. Articles receive scores ranging from -1 (complete refutation) to +1 (complete support) based on the mean of all detected stances (SUPPORTS = +1, NOT_RELATED = 0, REFUTES = -1). This approach captures nuanced article positioning beyond binary classification.

Analysis of 16,249 articles reveals predominantly neutral positioning across the corpus. The mean aggregate score of 0.045 (after consistency validation) indicates slight overall supportive bias, though the magnitude suggests minimal practical

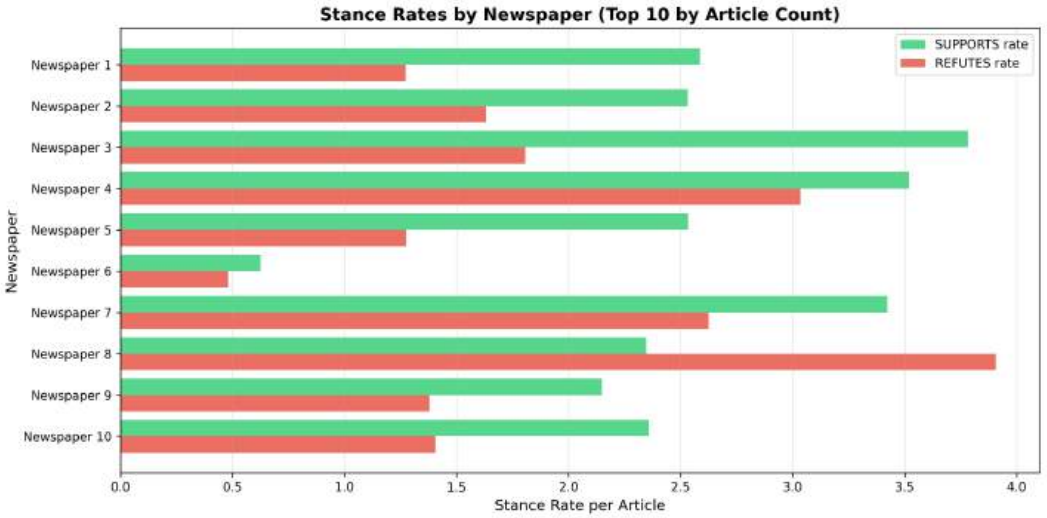
significance. Approximately 88.4% of articles showed mixed or neutral positioning (scores between -0.5 and +0.5), with only 6.7% predominantly supportive (scores > 0.5) and 4.9% predominantly refutative (scores < -0.5). The consistency validation process reduced supportive positioning, with 54.0% of articles becoming more refutative or less supportive through contextual re-evaluation, demonstrating the value of article-context validation.

Substantial newspaper variation emerges at the article aggregate level, with mean scores ranging from -0.347 to +0.385 across newspapers with sufficient article counts. However, high within-newspaper standard deviations (typically 0.24-0.31) indicate that individual article variability exceeds outlet-level effects. Topic-dependent positioning patterns are evident, with climate denial and renewable energy opposition queries showing supportive positioning (mean scores 0.124 and 0.089), while climate action advocacy and fact-checking promotion queries showed refutative positioning (mean scores -0.038 and -0.094). Monthly trends remained stable throughout the study period (January-April 2025), with mean scores ranging only from 0.043 to 0.061, indicating consistent patterns rather than temporal shifts in media positioning.

To understand outlet-specific behaviors, we calculated stance rates per article for each newspaper with sufficient article coverage.

For each newspaper, we computed the SUPPORTS rate per article, Total SUPPORTS stances (using FINAL stance) divided by total articles and the REFUTES rate (same formula but with the REFUTES counts. These metrics represent the average number of supporting or refuting stances per article for each outlet. Figure 22 shows the top 10 newspapers by article count..

Figure 22: Stance rates by newspaper (top 10 by article count). The figure shows the average number of SUPPORTS and REFUTES stances per article for each outlet. Substantial variation reveals heterogeneity in positioning patterns across different newspapers. The name of the outlet is anonymised as the data collection is not representative.



Analysis reveals substantial heterogeneity in stance patterns across individual newspapers, with widely varying rates of supporting and refuting stances per

article. Some newspapers consistently show higher engagement with fact-check-related content, demonstrating outlet-specific patterns that reflect editorial priorities and content strategies. Importantly, database-level aggregation obscures these important outlet- specific differences, indicating no systematic bias at the aggregate level. This heterogeneity underscores that editorial policy, topic specialization, and ownership structure drive positioning patterns more strongly than database membership.

4.3.5 Iberifier Digital media list

Concentrating the analysis on the Iberifier digital media list, the results show a broad media landscape, internal variation within the Iberifier digital media list group is substantial, with individual outlets ranging widely in their positioning patterns. Analysis indicates that editorial policy and topic specialization appear to influence positioning more strongly than database membership, and no systematic bias toward supporting or refuting fact-checked claims emerges at the database level.

The 15.8% support rate for debunked claims ("Falso") is noteworthy, as it suggests potential pathways for misinformation propagation through legitimate media channels. However, the 9.9% overall refutation rate (FINAL stance) and uniform stance positioning suggest active debunking efforts without strategic framing biases.

4.3.6 Discussion

The stance detection pipeline employs a systematic approach to identifying and analyzing the spread of misinformation within the Iberian digital media ecosystem. By combining verified fact-checks from established organizations with advanced computational techniques, the system enables large-scale analysis of how claims are mentioned, framed, and circulated within digital media. The hybrid retrieval strategy, statement-level granularity, and distributed processing architecture provide a replicable framework for examining misinformation patterns across the multilingual Iberian context.

The system is designed to be open source and modular, allowing researchers and practitioners to extend its capabilities beyond the current focus and ingest other types of structured and unstructured data. This adaptability ensures that the framework can evolve with the needs of the research and policy community while supporting comparative studies across domains. Moreover, the pipeline is implemented as a full retrieval-augmented generation (RAG) system built upon the existing Iberifier database, enabling fact-checkers to integrate large language models (LLMs) into their established workflows. This facilitates not only more efficient retrieval and contextualization of relevant claims but also supports evidence-based reasoning aligned with fact-checking practices.

Several directions are identified to further strengthen the pipeline's impact and applicability:

- **Automating data ingestion**: Streamlining the integration of new fact-checks from the Iberifier database into the RAG system to ensure continuous updates and minimize manual intervention.
- **Enhancing stance detection**: Improving the stance detection models, with particular focus on the internal consistency of outputs and robustness across diverse media contexts.
- **Cross-lingual verification**: Developing methods to identify and analyze whether claims and fact- checks propagate across languages, thereby capturing misinformation dynamics in the broader Iberian con- text.
- **Portuguese media integration**: Expanding the pipeline to systematically test and analyze the Portuguese media landscape, complementing the current emphasis on Spanish and Catalan sources.
- **Geographic and origin analysis**: Incorporating more refined tools to examine the provenance of fact- checks, their original sources, and the geographic distribution of the digital media in which they circulate.
- **Enhancing the IBERIFIER database**: While the information contained in the IBERIFIER database is well defined, there is room to improve the different consistency of fields to ensure a common understanding of the different fields. The use of the field debunked, should become systematic and offer a clear distinction between the claim, its context and the information added by the fact-checkers.

These developments aim to extend the system's scientific value, ensuring broader coverage, higher accuracy, and deeper insights into the mechanisms of misinformation in multilingual and cross-national contexts.

5. The spread of misinformation on Twitter

5.1 Data Collection and Misinformation Identification

As mentioned, we used the GlobalClaims[10] dataset (Vranic et. al. 2025), collected and curated by the Social Physics and Complexity (SPAC) team at LIP. This is a collection of 67,000 fact-checked claims published between 2015 and 2023 by over 100 fact-checking organisations worldwide. This dataset provides the original verdict for each claim, as well as a normalised veracity category mapped to a unique scale across different fact-checker categories (true, false, other, and satire). In addition to veracity labels, the dataset also included thematic classification across 22 subtopics, grouped under six broader domains: Politics, Health, Environment, Science and Society. Each claim is also associated with its source URL, enabling further analysis of the spread of these claims on social media platforms.

For the purpose of this study, the analysis focused on *false* claims fact-checked by organisations based in Spain, Brazil, and several Spanish-speaking countries, including Chile, Mexico, and Venezuela. These countries were selected due to their media relevance within the Iberian and Latin American information ecosystem. The thematic scope was limited to claims related to *Covid-19 and Environmental* issues, two domains that generated large amounts of misinformation activity across social platforms in the past decade. The distribution of claims across fact-checkers, countries, and thematic categories is summarised in Table 12 below.

Table 12: Number of false Covid-19 and Environmental claims analysed by fact-checkers in Spanish- and Portuguese-speaking countries.

Fact-Checker	Country	N Covid claims	N Environment claims
Maldita.es	Spain	13	0
AFP Checamos (Brazil)	Brazil	164	45
Comprova	Brazil	90	35
Agência Lupa	Brazil	3	4
AFP Factual	Spanish-speaking Countries	302	237
Mala Espina Check	Chile	23	26
FastCheckCL	Chile	23	12
Telemundo	Florida (Latin community)	19	2

In total, the selected subset includes over 1000 claims across both topics, with a predominance of Covid-19–related misinformation verified by Brazilian and regional Latin American fact-checkers. Environmental misinformation, while less frequent, is more analysed by Spanish-language outlets.

[10] https://doi.org/10.1145/3746275.3762201

5.2 Spread of fact-checked claims on Twitter

To understand how verified misinformation circulates on social networks, we analysed the diffusion of fact-checked claims on Twitter. Using the Academic Twitter API, we collected all publicly available posts containing URLs corresponding to fact-checked claims verified by fact-checking organisations in Portuguese- and Spanish-speaking countries. The dataset collection lasted until June 2023, when the Academic API was discontinued.

In total, we collected approximately 20,000 tweets, including original posts and their associated retweets. Each tweet was categorised according to the claim's topic (*Covid-19* or *Environment*) and whether a Portuguese or Spanish fact-checking organisation analysed it. The tables below display the false claims mostly shared on Twitter by language (Spanish and Portuguese) and topic (Covid-19 and Environment) (Tables 13A-13D).

Table 13A: *Collected false Covid-19 claims in Spanish*

N	Claim	N tweets
1	El Gobierno paraliza en Zaragoza 5.000 kilos de mascarillas para Madrid porque «Aduanas cierra a las 15h»	4069
2	Celáa y Valerio, Ministras y ex ministras de Sánchez se pusieron guantes de látex el 8M por miedo al coronavirus	676
3	Whatspp cede a las presiones del Gobierno y limita el reenvío de mensajes durante la cuarentena	416
4	Las vacunas del COVID-19 magnetizan a las personas.	361
5	"LA BANCA ROTHSCHILD REGISTRÓ HACE CINCO AÑOS UN MÉTODO PARA TESTEAR EL COVID-19"	308
6	Anuncia Cuba que fabricó vacuna contra el coronavirus	200
7	La FDA acaba de actualizar su postura sobre las PCR y no volverán a ser usadas en Estados Unidos EEUU	184
8	Murió más gente por las vacunas contra la COVID-19 en estos tres meses en EEUU que en la última década	89
9	Nuevo estudio evidencia relación covid-19 y la radiación 5G	55
10	La crisis del coronavirus rebaja en Italia el número de políticos	31
11	No es sano respirar CO2 dióxido de carbono. "USO RACIONAL DE LAS MASCARILLAS"	21
12	"(Las recomendaciones para prevenir el COVID-19) son profundamente anticientíficas que se enfocan en un control absoluto del Gobierno sobre cada aspecto de nuestras vidas".	17
13	Los vacunados tienen una carga viral 251 veces superior de coronavirus que nos lo no vacunados contra el covid-19	16
14	Una exempleada de Pfizer confirma que la vacuna contiene nanopartículas de óxido de grafeno	12
15	El gobierno británico reconoce en un documento oficial que el resurgimiento tanto de las hospitalizaciones como de las muertes está dominado por los que han recibido dos dosis de la vacuna.	12
16	La ivermectina sirve para tratar el COVID-19.	11
17	"Afortunadamente, mientras el virus comenzó a extenderse, el presidente actuó rápido para asegurar que los ventiladores fueran a los hospitales que más los necesitaban"	11

Table 13B: *Collected Environment claims in Spanish*

N	Claim	N tweets
1	CAYÓ UN METEORITO EN LAS COSTAS DE ANTOFAGASTA DURANTE LOS SISMOS Y FALSA ALERTA DE LA ONEMI	1274
2	"ESTADO DE CHILE PAGARÁ SUELDO MENSUAL A LA EMPRESA AES GENER DURANTE 5 AÑOS PARA QUE CIERRE SU TERMOELÉCTRICA"	37
3	"Puerto Rico ha recibido 91.000 millones de dólares por el huracán, más dinero del que nunca se ha obtenido por un huracán antes"	15

Table 13C: *Collected Covid-19 claims in Portuguese*

N	Claim	N tweets
1	Senador Randolfe promete comendas a médicos do Amapá pelo que chamou de heróico combate a Covid-19. Detalhes: eles trataram com hidroxcloroquina, ivermectina, vitamina D e mais. Vai condecorar?	3170
2	Homem de 33 anos sofre grave acidente de moto. Houve traumatismo cranioencefálico. O óbito foi por Covid-19	2002
3	Maior estudo retrospectivo em pacientes hospitalizados mostra que a hidroxicloroquina reduz significativamente o risco de mortalidade por covid-19	1423
4	O uso de máscaras não reduz o risco de contrair o coronavírus e é prejudicial à saúde	477
5	Hospital Unimed-Rio cura covid-19 do paciente Antônio Carlos,67, com cloroquina	314
6	Cientistas afirmaram que sol do meio-dia é extremamente eficaz na erradicação do vírus. Segundo o estudo, o sol mata o coronavírus em 34 minutos.	101
7	Bolsonaro recusou um contrato de 70 milhões de doses de vacina até dezembro/21 em prol de um contrato de 100 milhões de doses até SETEMBRO/21. O que a mídia faz? Divulga só a primeira metade da notícia.	95
8	Pessoas com lúpus não desenvolvem Covid-19	63
9	Usar máscaras é destruir mais ainda com a sua imunidade	62
10	OMS alerta sobre máscaras infectadas que chegam ao Brasil	53
11	Médico do Hospital Municipal Ronaldo Gazzola, no Rio de Janeiro, defende o fim da quarentena	42
12	Dez artistas morreram de covid mesmo após tomar Coronavac	34
13	Variante Delta é menos agressiva, vacinas não protegem contra ela e crianças não transmitem o vírus	28
14	Médica afirma em vídeo que a covid-19 pode ser curada com a ivermectina, remédio usado para tratar parasitas	22

Table 13D: *Collected Environmental claims in Portuguese*

N	Claim	N tweets
1	Madeira ilegal apreendida no Pará pelo Exército seria de fundador de uma ONG ligada à proteção ambiental na Amazônia, que também teria envolvimento com o MST no estado.	599
2	Arrozeiros de Roraima viram as áreas plantadas encolherem pela metade depois da demarcação da Terra Indígena Raposa Serra do Sol	590
3	Amazônia não está queimando, Brasil é o país que mais preserva áreas nativas do mundo e alimentos brasileiros são os mais sustentáveis do planeta	246

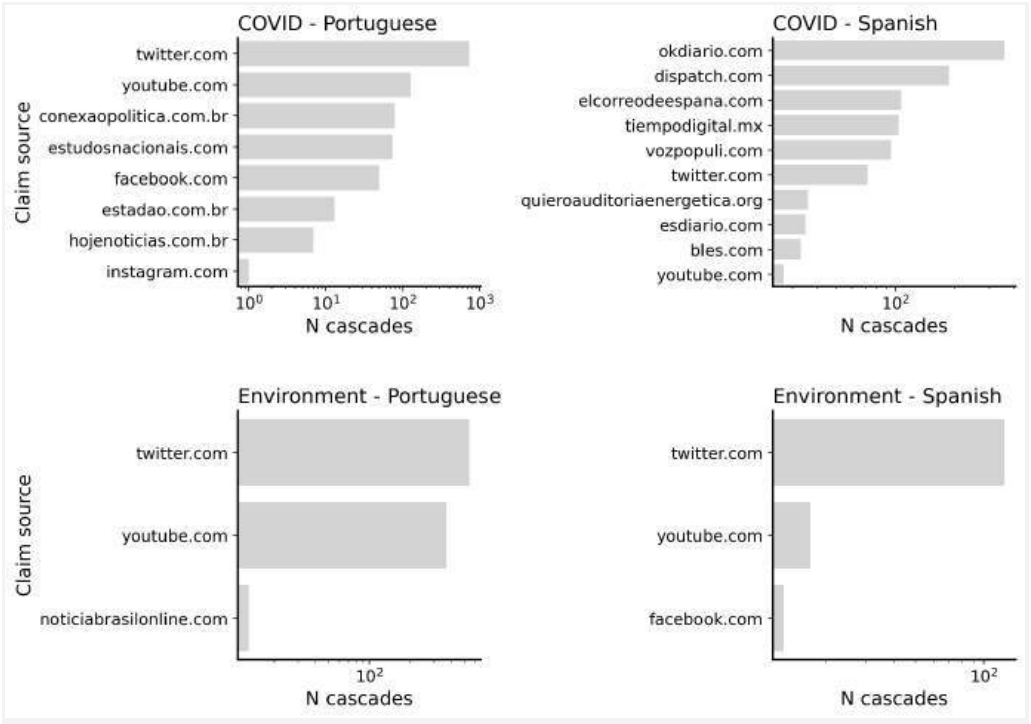
Table 13D: *Collected Environmental claims in Portuguese*		
N	Claim	N tweets
4	o maquinário que os chineses estão usando para fazer a estrada de ferro que vai da região de Lucas do Rio Verde	202
5	Sérgio Moro tem o aval legal para mandar a Polícia Federal investigar o Greenpeace	107
6	STF vai "constitucionalmente" cuidar do meio ambiente	69
7	Técnicos da Anvisa liberaram agrotóxicos banidos na Europa mas não a importação da vacina Sputnik V. Não é politização, é coincidência...	21

From these tables, several patterns emerge. First, Covid-19 misinformation dominates in both the number of unique claims and their overall diffusion. Spanish and Portuguese claims share similar narratives, vaccine scepticism, and distrust in health authorities. The most viral claim in Portuguese promoted treatments such as ivermectin or hydroxychloroquine, whereas in Spanish, it focused on the use of masks and government actions in deploying them. In contrast, environmental misinformation is less frequent. Spanish-language examples often exploit local events, such as meteorite sightings or regional energy policies, while Portuguese claims focus on Amazon deforestation and governmental regulation.

5.2 Properties of Twitter Cascades

A **Twitter cascade** is defined as a chain of retweets originating from an initial post that shared a given claim's URL. By analysing these cascades, we can quantify the replication and visibility of claims through user engagement. Figure 23 shows the number of cascades per URL source shared on Twitter.

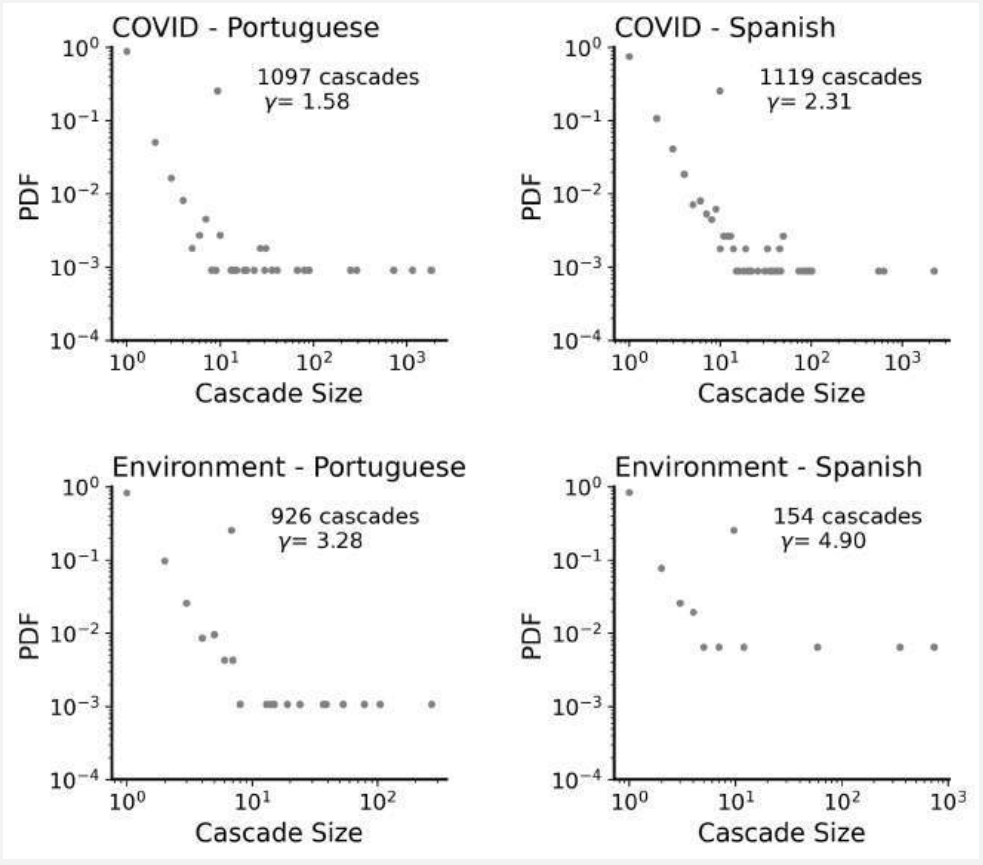
Figure 23: Number of cascades per source which were shared on Twitter. Please note the difference in scale (log, x-axis).



The structural properties of misinformation diffusion on Twitter were analysed by examining the cascade size distributions (i.e., the total number of tweets in one cascade) for both topics (*Covid-19* and *Environment*) in Portuguese and Spanish. The resulting distributions, shown in Figure 24, were fitted to a power-law function of the form $P(k) \sim k^{-\gamma}$, where $P(k)$ denotes the probability of observing a cascade of size k, and γ is the scaling exponent. This model is often used to describe heavy-tailed diffusion processes in social media environments, where a small number of events achieve disproportionate amplification while most receive minimal engagement.

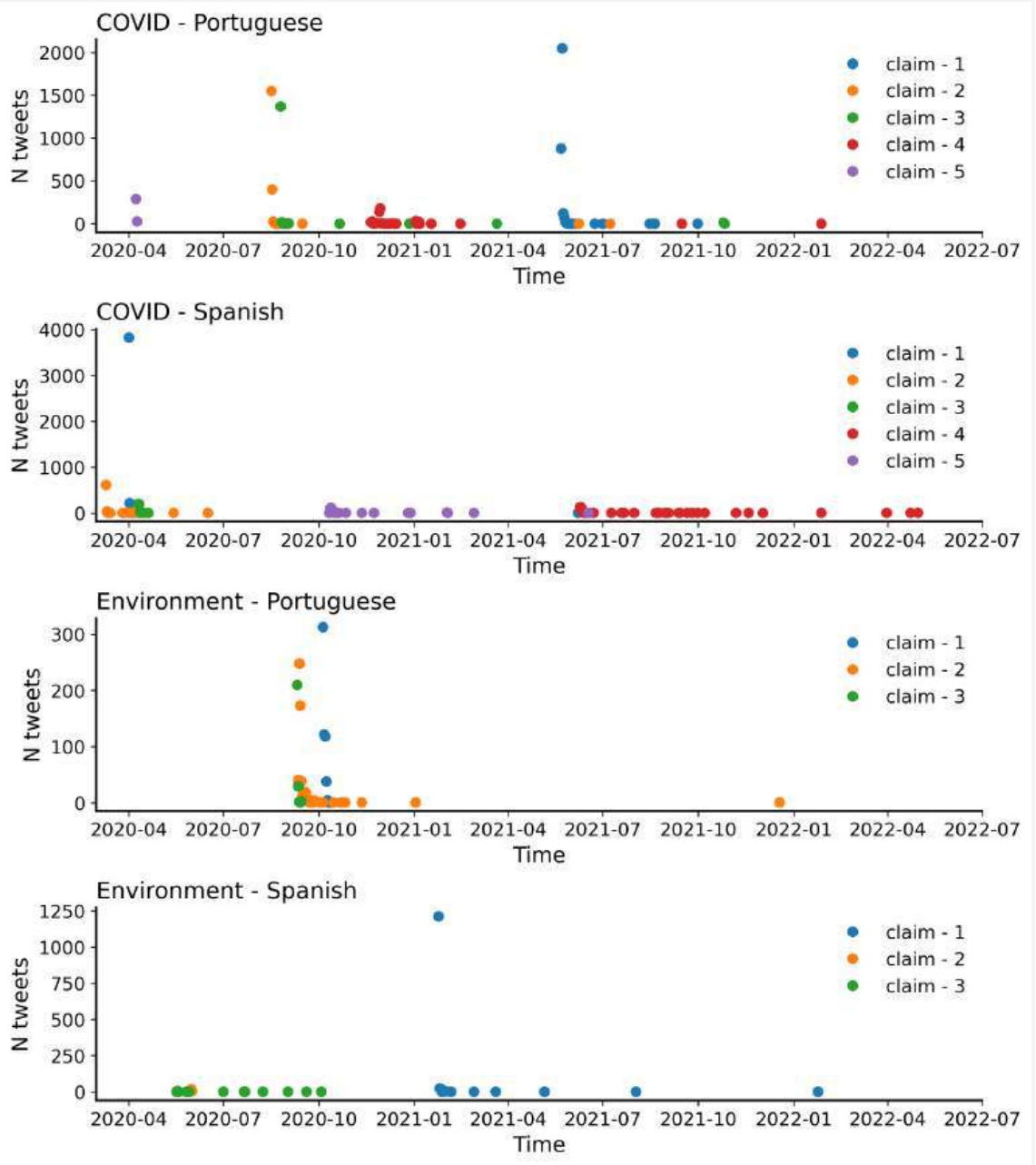
As expected, the distributions exhibit a heavy-tailed pattern across all languages and topics, suggesting that most misinformation claims generated small cascades, while a limited subset achieved widespread diffusion. For COVID-19 misinformation, both Portuguese and Spanish datasets contained substantially more cascades than their environmental counterparts—1,097 cascades in Portuguese ($\gamma = 1.58$) and 1,119 cascades in Spanish ($\gamma = 2.31$). Environmental misinformation produced fewer cascades overall—926 in Portuguese ($\gamma = 3.28$) and 154 in Spanish ($\gamma = 4.9$).

Figure 24: Probability distribution (PDF) of cascade sizes for claims classified as false within the Covid-19 and Environment topics, in Spanish and Portuguese. Distributions are fitted with a power-law $P \sim k^{\gamma}$ distribution, labelled with the power-law exponent and the number of cascades.



Naturally, these tweets were not uniformly distributed over time. Figure 25 illustrates the frequency of misinformation dissemination identified. As can be seen for both topics, we observe bursts, followed by periods of reduced activity.

Figure 25: Number of tweets that shared Covid-19 and Environmental claims identified as being false in Portuguese and Spanish, over time. Each dot represents the number of tweets shared per day for the top shared claims from tables 13A to 13D.



The Portuguese Covid-19 misinformation claim N1 from Table 13C, *Senador Randolfe promete comendas a médicos do Amapá pelo que chamou de heróico combate à Covid-19. Detalhes: eles trataram com hidroxcloroquina, ivermectina, vitamina D e mais. Vai condecorar?* , recorded over 2,000 shares on its first day of circulation (22 May 2021). In contrast, claims N2 and N3 achieved peak engagement earlier, between 16 and 25 August 2020.

A similar temporal pattern was observed for the Spanish Covid-19 misinformation claim N1, Table 13A, *El Gobierno paraliza en Zaragoza 5.000 kilos de mascarillas para Madrid*

porque «Aduanas cierra a las 15h» This claim was shared more than 4,000 times on 1 April 2020, after which its popularity declined markedly in subsequent days.

Comparable dissemination dynamics emerged among environmental misinformation claims. We observe that collected Portuguese environmental claims (listed in Table 13D) circulated predominantly between 11 September 2020 and 18 October 2021, while the Spanish environmental claim (listed in Table 13B)—which concerned a meteorite event—began circulating on 24 January 2021, reaching its highest level of engagement on that date.

5.3 Comparative Analysis

It is common that rumours that spread on one social network also spread on others. Therefore, we compared the fake news identified through Global Claims and the hoaxes described in Section 2 of this report. Using a Large Language Model with a prompt to compare the four lists in 1 one to 1 pairs (Spanish COVID, Portuguese COVID, Spanish Environment and Portuguese Environment), the model found no matches. We then tried a fuzzy matching approach, using Levenshtein Distance, and no claims had a distance under 50.

In summary, the study reveals that misinformation in Portuguese- and Spanish-speaking contexts—particularly around Covid-19—spread widely on Twitter/X, exhibiting distinct temporal spikes, with a small number of false claims achieving substantial reach and engagement (hundreds of shares). There is sub-topic overlap between Twitter/X and Telegram (Ex. risks of vaccines), but the wording of the claims varies significantly. These findings underscore the uneven yet impactful dynamics of misinformation circulation across languages, topics, platforms, and time, emphasising the importance of cross-linguistic, thematic and multi-platform analyses in understanding how false information propagates within digital ecosystems.

6. Technical report

6.1. Data gathering from Telegram

Telegram is a communications app where several people can form “Channels” to talk through. Telegram has an interface point (API) that lets a phone number connect to a known chat. There is no direct method to crawl the channels in the app as access to unknown channels is strictly forbidden.

Telegram is a communication platform that allows users to create “channels” for group discussions. Access to these channels is provided through an application programming interface (API), which requires a phone number to connect to a known chat. However, Telegram does not offer a direct method for large-scale crawling, as the channel name is required to retrieve its messages. Consequently, for this study we relied on a curated set of 491 channels, selected by our research team in collaboration with colleagues, which were periodically crawled to extract newly available messages.

6.2. Hoax identification in the wild

We focus on three main topics and aim to search our Telegram database for related content. Since the cascade algorithm requires well-defined queries, we first need a reliable source of hoaxes to guide the search. For this purpose, we use the *Fact-checkers* section of the Iberifier website, which compiles hoaxes debunked by various fact-checking organizations. From this source, we collect both the date when each hoax was debunked and its title.

In total, we retrieved 8,256 hoaxes spanning from April 2019 to June 2025. To align with the period covered by our Telegram dataset, we then selected only the hoaxes published between 2020 and 2023 (inclusive), resulting in a final set of 4,252 hoaxes. These hoaxes cover a wide range of topics, wider than the three main topics we want to research, namely: covid, climate change and gender related issues. Therefore, the hoaxes dataset was further refined to find hoaxes of the selected themes.

In our case, we converted all 4,252 hoaxes into embeddings using Transformer-based models and stored them in a vector database. We then formulated 24 sentences or concepts representing the target theme and used them as queries to locate hoaxes near the queries in the vector space. For example, for the climate change theme, we included concepts such as *“climate change”* and *“Rising global temperatures and greenhouse gas emissions.”* The complete list of concepts used for each topic is provided in Table 14.

Table 14. Queries used to find related hoaxes and messages.

Climate Change	Gender related issues
"Climate change"	"LGTBIQ+"
"United Nations and climate change action"	"Transgender"
"IPCC reports and global warming"	"feminism",
"Greta Thunberg and youth climate activism"	"Gloria Steinem",
"Paris Agreement and international climate policy"	"Me Too movement",
"NASA climate research and satellite data"	"sexual harassment",
"Shell and fossil fuel emissions"	"International Women's Day",

"Greenpeace campaigns against deforestation"	"Malala Yousafzai",
"EU climate goals and emissions targets"	"NOW and feminist advocacy in the U.S.",
"ExxonMobil and climate change controversy"	"Simone de Beauvoir",
"Bill Gates and climate innovation funding"	"CEDAW and women's human rights",
"Extreme weather events"	"Emma Watson and HeForShe campaign",
"Carbon footprint reduction"	"Planned Parenthood",
"Rising global temperatures and greenhouse gas emissions"	"women's reproductive rights",
	"Pride parades",
	"Laverne Cox and transgender representation",
	"Queer culture",
	"Elliot Page and transgender visibility",
	"Same-sex marriage",
	“Conversion therapy",
	"Religious views on homosexuality",
	"Legal protections for homosexual couples",
	"Discrimination against gay men in the workplace",
	"Harvey Milk and gay rights activism",

Through this approach, we identified 157 hoaxes related to climate change and 149 hoaxes concerning gender-related issues.

6.3. Message identification in the wild

As noted above, determining whether a given message conveys the same content as a hoax is not straightforward. Processing more than six million messages individually for each hoax is unfeasible. To address this challenge, we filtered the dataset to identify messages related to the target themes.

First, each channel was segmented into a set of topics, where a *topic* is defined as a group of messages discussing the same subject. This process yielded a total of 21,614 topics. For each topic, we computed an embedding and stored it in a vector database. We then applied the same set of queries used in the hoax analysis to identify topics aligned with the desired themes. Finally, we retrieved all messages belonging to the selected topics from the database.

Using this approach, we identified 40,860 messages related to climate change published by 97 channels and 52,289 messages related to gender issues published by 102 channels.

6.4 The Covid-19 case study

The procedure used to analyze the Covid-19 theme differed from that applied to climate change and gender-related issues, as we already possessed a subset of 21 channels known to contain discussions about vaccines and Covid-19. Consequently, the process was more straightforward.

Instead of employing a vector database to search for relevant messages and hoaxes, we directly compared the embedding of each topic with the embeddings of each hoax. If the distance between a topic and a hoax was smaller than the distance

between that hoax and the mean of all embeddings, we inferred that the topic addressed the hoax. Using this method, we identified 40 distinct hoaxes.

In this case, we did not filter messages, as the dataset was considerably smaller—approximately 280,000 messages.

6.5. From messages to cascades: Introducing SLICs

SLIC stands for *Semantically-linked information cascades*, which is the end result we want to achieve. We describe them as messages linked by a combination of semantic meaning and other factors. To construct them, there are several steps to be taken: 1) choosing a hoax, 2) retrieving messages relating to a hoax, 3) reranking top-k messages, 4) triangular self-similarity adjacency, 5) Pruning the triangular self-similarity graph into a cascade. We examine this process step by step.

1. Choosing a hoax / Building the query: The first step is knowing the target hoax to be tracked. The SLIC generation algorithm tracks the cascade of a single misinformation piece. It can be used for information tracking in general, but that use-case has gone untested so far. Granularity may be as coarse or wide as needed, but it is recommended to be very specific. For example: “Anthropogenic climate change is a hoax generated by big government.” is a much more effective query than, “Climate change”, “Climate change is a hoax generated by big government.” and “Anthropogenic climate change doesn’t exist”. Regular rules applied to LLMs usually extend to this method on principle, therefore negatives like “Don’t” or “Not” are usually disregarded by the algorithm. Constructing an effective query hoax requires adjusting the level of specificity desired and running through the entire pipeline.

2. Coarse grain message retrieval: First, we need the entire message database encoded as numerical representations (embeddings) to proceed with the **semantic similarity** procedure (Consult Annex 2 for details). The embeddings are used as keys, as in an archiving system, they serve as short-hand to search for any given item; therefore we need a query to compare against these keys. The query was obtained in step 1, but also needs to be transformed into an embedding.

We perform cosine similarity on each key-query pair, obtaining one score number for each pair. We desire to extract a group of messages to link them semantically, therefore the *K* keys with the highest scores are selected from the database and sent forward for further analysis. This is a very common step on **Information Retrieval** problems.

Technical details: The encoder used for this purpose is Qwen3-Embedding-4B[11], quantized to 4-bit for speed-performance tradeoff. The system prompt was: “Given a query, retrieve relevant passages that are thematically related to the query: “. No thinking tokens were used.

3. Fine grain message retrieval: Similarity search is known to be imprecise. Common practice uses a **reranking** step. Reranking is very simple, given a **small** set of keys,

[11] Model checkpoint: Qwen/Qwen3-Embedding-4B

extract the most relevant ones. In essence, this is almost identical to the previous step, however, the tools applied to extract reranking scores are much more computationally demanding, therefore it is often used on reduced datasets.

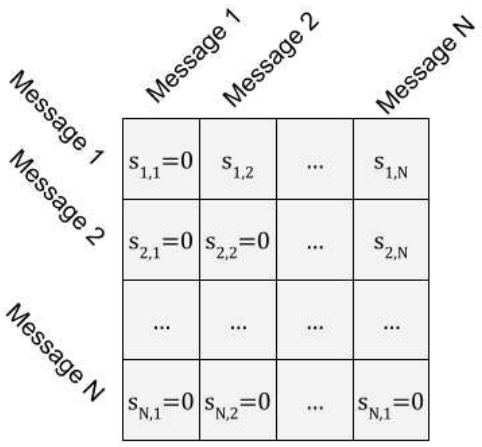
Reranking gets all key-query pairs and inputs them to the reranker. The reranker computes a score for each pair. Coarse grain IR required 1 execution of the encoder. Fine grain IR requires N executions of the reranker, one for each pair. In the end, the results are a group of scores for each key, where we extract the *Q* highest scores.

Technical details: The reranker used for this purpose is Qwen3-Reranker-4B[12], quantized to 4-bit for speed-performance tradeoff. The system prompt was: "Given a query, decide whether the query agrees with the document mutually (bidirectional): ". No thinking tokens were used.

4. Triangular self-similarity adjacency matrix: We have a group of messages that we know relate to the target hoax. Our goal is to build a cascade of information by linking these messages. This is performed in two steps. The first phase of this procedure is computing a score for each pair of messages. The result of this step will be a directed graph (a message influences another in the future) that is fully connected forward. A similarity of 0 means there is no link.

Figure 26 is a simplification of the process. First, all messages are sorted by ascending date. All pairs of messages are evaluated through the reranker algorithm (same as step 3) and their similarity scores are annotated in matrix format.

Figure 26. Triangular self-similarity adjacency matrix explanation



Source: Original

Furthermore, as we discussed before, similarity by itself is not enough to link messages. A simple and effective approach is to diminish similarity by a function of time. Say, for example, messages 1 and 2 are distanced by 3 days, while messages 1 is distanced from 3 by 3 years. It stands to reason that, independently of similarity, message 1 is much more influential to message 2 than message 3, as the discourse

[12] Model checkpoint: Qwen/Qwen3-Reranker-4B

has varied wildly in that time frame. This reasoning has some flaws, as messages may really have long-term influences and be re-visited, however it is robust for common day-to-day discourse, especially in the highly mutable ecosystem of social media where trends and ideas are quickly deemed obsolete.

We perform the penalty using (years + 1)^2. The square term is added due to the previous consideration, where social media heavily rewards immediate interactions. The choice of time-frame (years) was arbitrary, as any time-frame would work, but years provided good representations as most similarity scores by the reranker ranged between 10 and -10.

Technical details: Same configuration as step 3. This time all messages are related pairwise and obtained their similarities. For efficiency, only the upper triangle of the matrix was computed. The only computation performed was the following:

$$s_{i,j} = f_\phi(m_i, m_j) \,/\, (y_{i,j}+1)^2 \;\forall\; i>0, j>1, j>i$$

F_ϕ is the parametrized reranker, m_x is the selected message, y is the number of years between messages x and y.

5. Pruning the adjacency matrix to create SLICs: We already have a graph detailing the influence from one message to another. However, a fully connected graph is impossible to interpret beyond 10 or more nodes. Instead, we perform a greedy algorithm that prunes the adjacency matrix to develop a tree formed from the relationships in the graph. Thus, we develop an original algorithm to build a SLIC.

To summarize the behaviour in layman terms. We evaluate every pair of relations between messages. We begin with the first message, which we call the 'Root'. We find all the edges that begin on this root and end outside the new graph we are building. We find the edge with the highest score, in other words, the pair of messages that have the highest similarity between them. Each step of the algorithm adds a new message to the graph until the information cascade is completed. What we optimize is this algorithm is the sum of all relationships, which we aim to maximize. This algorithm provides a fair estimation of the best cascade, i.e. the cascade that better propagates information forward with the highest possible similarity scores.

Some additional considerations. The graph visualization and the SLIC are not identical. This means that some conveniences have been taken for the sake of readability. For example, in the visualization, messages are aligned by their channel of origin, a detail we never contemplate in the SLIC construction. This is by design, as we meant to influence the algorithm with as little information as possible, and this has been similarity and time difference.

Technical details: We detail the novel algorithm here in Algorithm 1. It has $O(N^2)$, which is comparable to the self-similarity matrix creation, thus offering no more compute overhead to the process.

Algorithm 1. SLIC building algorithm

LET G be a triangular self-similarity adjacency matrix, our fully-connected graph
LET T be a graph with a root (the first message), representing the final SLIC
LET N be the side size of the matrix
REPEAT N times // Each Edge E in Graph G
    $B \leftarrow$ Empty // Will use to represent the best edge
    FOR EACH E in M **DO** // Each Edge E in Graph G
        $U \leftarrow$ Edge E_{origin} | V $\leftarrow$ Edge $E_{destination}$ // A Vertex
        IF ($U \in T_{Vertices}$ **AND** $V \notin T_{Vertices}$) **AND** (B is Empty **OR** $B_{weight} < E_{weight}$)
        // Add the most similar edge that points to an unvisited vertex.
        THEN $B \leftarrow E$
    Remove B from G_{Edges}
    Add $E_{destination}$ to T
    Add B to T
RETURN T

Concluding remarks: The developed algorithm for SLIC construction is a multi-step process with high computational demand, especially the reranking and self-similarity phases. However, they provide high-quality estimations of semantically linked cascades, which was our initial objective.

6.6. Grouping SLICs by case

An SLIC illustrates how a single hoax is distributed across channels. By analyzing a single SLIC, we can observe patterns in which some channels initiate a hoax while others act as hubs that propagate it. Using pattern mining techniques, we can analyze multiple SLICs to model how channels influence one another more generally. This is typically represented with a who-copies-who graph—a directed graph where nodes represent channels and edges indicate influence between them. Once the graph is constructed, we can apply Social Network Analysis (SNA) techniques to identify the most important nodes in the distribution process and to classify each node as either a hub or an authority (with hubs tending to aggregate hoaxes and authorities tending to initiate their distribution), among other things. This process involves two main steps: 1) Generating a who-copies-who graph by analyzing several SLICs. 2) Applying SNA techniques to analyze the graph, determine each channel's role, and assess their importance within the network.

1. Generating a who-copies-who graph by analyzing several SLICs: Modeling a who-copies-who graph remains an open problem, and the state of the art provides several approximations. For this study, we use the **NETINF algorithm**, which frames the task as a probability maximization problem and applies a greedy method capable of scaling to networks with hundreds of thousands of nodes.

NETINF is based on the **Independent Cascade Model**, where each node in the network has a single chance to influence its neighbors with a fixed probability. To represent this probability, NETINF employs a configurable distribution that depends on the time elapsed between channels publishing a hoax—commonly modeled with a **Power-law** or **Rayleigh** distribution.

Additionally, NETINF incorporates a **default influence probability** to account for influences that occur outside the observed data (for example, when two channels pick up a hoax from an internet forum). The algorithm begins with a graph where all nodes are connected via default influences. It then iteratively replaces one default influence edge with a real influence edge. At each step, the edge selected is the one that maximizes the likelihood of observing all the SLICs in the dataset. The process finishes when a certain number of edges is added. For this study we have added edges until all the channels in the case study appear in the graph.

2. Applying SNA techniques to analyze the graph, determine each channel's role, and assess their importance within the network: A who-copies-who graph on its own is of limited use. As the number of nodes and edges grows, it becomes increasingly difficult to extract meaningful insights directly from the graph. Therefore, we rely on Social Network Analysis (SNA) techniques to identify key information. In particular, we need methods to determine the most influential accounts in the distribution process and to reduce redundant edges, making the graph more interpretable.

Centrality metrics are a core component of the SNA toolbox, providing ways to measure the importance of both nodes and edges. Since each metric emphasizes different aspects of the network, for this study we selected the **HITS algorithm** (Kleinberg 1999) for node centrality. HITS assigns both an *authority score* and a *hub score* to each node. Hubs are nodes that point to many authorities, while authorities are nodes that are pointed by many hubs. In our who-copies-who graph, authorities represent channels that tend to initiate hoax propagation, whereas hubs represent channels that distribute hoaxes across the network.

For edges, we selected **betweenness centrality**, which measures how many shortest paths pass through a given edge. Edges with higher betweenness are more critical for holding the network together.

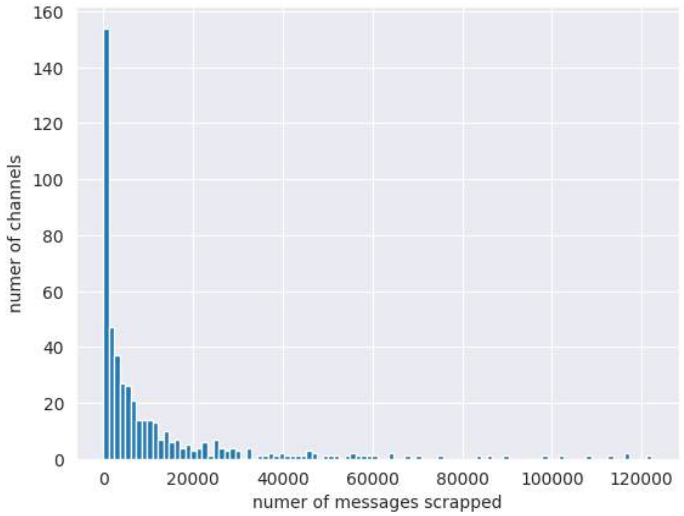
Using these two metrics, we designed the visualization presented in Section 3. Node color distinguishes hubs from authorities, while node size and saturation are proportional to the HITS score—larger, more saturated nodes play a more significant role in the diffusion process. Additionally, the names of the three most important hubs and the three most important authorities are displayed next to their nodes. For edges, saturation corresponds to betweenness centrality—darker arrows indicate edges that are more essential for maintaining network connectivity, which further enhances readability by effectively downplaying redundant edges.

6.7. Visualization data source metrics

The dataset used in this study comprises **491 channels** and **6,805,626 messages** collected between 2020 and 2023 (inclusive). The median number of messages scraped from each channel is **4,043**. However, the distribution is not homogeneous; instead, it follows a **power-law pattern**. The 50% of channels with the fewest messages account for only **4%** of the total messages, while the top 5% of channels account for **45%** of all messages. Notably, the maximum number of messages

scraped from a single channel is **623,344**, representing approximately **10%** of the dataset—nearly twice the total contributed by the 50% of channels with the fewest messages. A histogram of the message distribution across channels is shown in Figure 27.

Figure 27. Histogram of number of messages published by channel



Regarding the temporal characteristics of the dataset, the number of messages scraped increased each year. Specifically, **419,351 messages** were collected in 2020; **1,576,924** in 2021; **1,966,060** in 2022; and **2,571,944** in 2023. Older messages were also scraped; however, since the quantity was very small, they were discarded from the analysis.

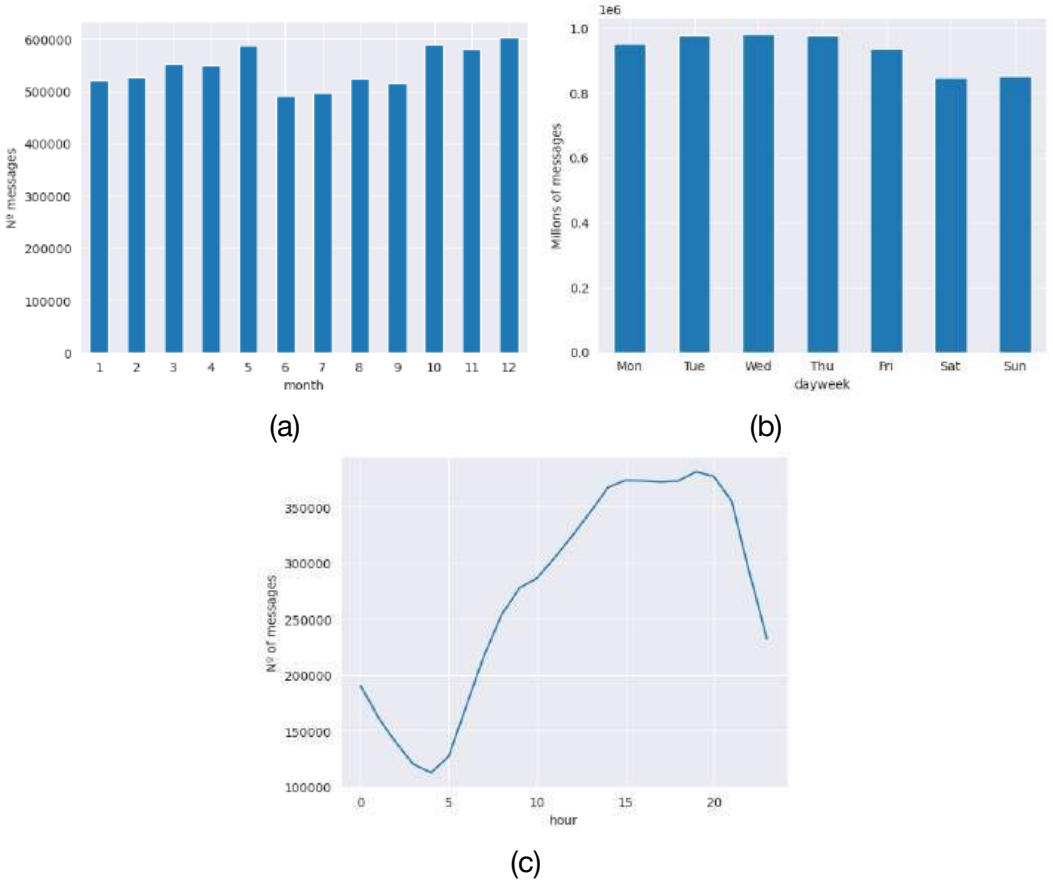
When examining the months of publication, we observe a slight increase in message volume during spring and at the end of the year (see Figure 28A). At the weekly level, fewer messages are published during weekends compared to weekdays (see Figure 28B). Finally, analyzing the time of day, peak activity occurs in the evenings, approximately between **15:00 and 20:00** (see Figure 28C).

In addition to text, many messages also included multimedia content. The most common type was **video**, shared in **771,617 messages (≈11%)**, followed by **files** (e.g., PDF, ZIP) with **8,501 messages (>1%)**, **images** with **5,414 messages (>1%)**, and **audio** with **3,049 messages (>1%)**.

Moreover, **3,739,279 messages (≈55%)** contained a URL linking to external websites. Among these, **187,930 (≈3%)** pointed to YouTube videos, **418,525 (≈6%)** to other Telegram channels, **319,517 (≈5%)** to X (formerly Twitter), **74,210 (≈1%)** to Rumble—an online video platform positioned as an alternative to YouTube—**60,304 (>1%)** to Reddit, **24,094 (>1%)** to Facebook, **20,852 (>1%)** to Instagram, and **11,946 (>1%)** to TikTok content.

In addition to other social media platforms, some links directed to mainstream international outlets such as *The Guardian*, *BBC*, *Reuters*, *NBC News*, *ABC* (Spain), *El Mundo*, *20 Minutos*, and *MSN*, although these appeared relatively infrequently compared to other sources. By contrast, significantly larger volumes of links originated from alternative or partisan media outlets, including *The Gateway Pundit*, *Daily Mail*, *Fox News*, *Breitbart*, *RT*, *Epoch Times*, *ZeroHedge*, *Natural News*, and *Rebel News*.

Figure 28. Volume of messages published by month/day of week/hour.

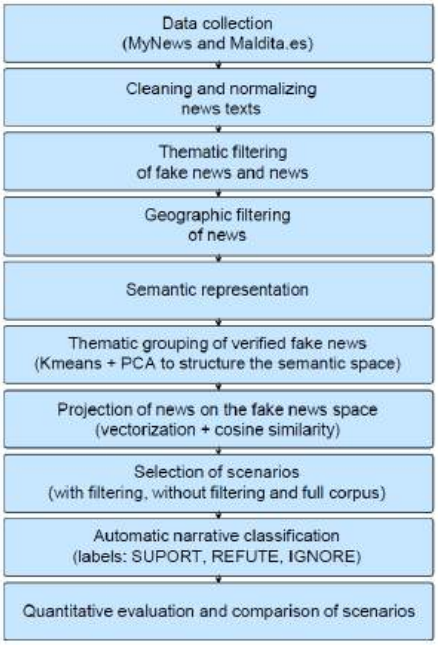


6.8. Thematic clustering methodology

This approach explores the impact of thematic clustering before classification in the automated detection of narrative positions about climate disinformation in digital media. Research on disinformation detection often uses models like BERT for supervised classification, with studies with French texts (Mebdeb et. al. 2022) ; adding explainability tools (Szczepanski et. al. 2021) and using multimodal analysis of social media (Arcos et. al. 2025). This study contributes with an unsupervised clustering phase before narrative inference, using the multilingual model distiluse-base-multilingual-cased-v2 (Reimers & Gurevych 2019), which produces comparable vector representations. The study proposes a hybrid method that combines natural language processing and thematic clustering before narrative classification to improve the accuracy of the system.

The study was structured in several phases that combined collection, filtering, vectorization, thematic clustering of contents, and later narrative analysis with pretrained models. This design made it possible to explore if the use of clustering before classification improves the performance of language models in narrative detection. The methodological workflow integrates unsupervised thematic analysis with supervised semantic classification, see Figure 29.

Figure 29: Methodological scheme of the study, which combines unsupervised thematic analysis with super- vised classification based on language models.



The corpus was built from two sources. First, 7,286 fake news verified by Maldita.es maldita between 2017 and 2025 were collected. From these, 50 climate fakes were selected after applying a thematic filter centered on climate topics. Second, 23,441 news articles were downloaded from MyNews news during the first four months of 2025. Using relevant keywords from the fakes, a thematic filter reduced this corpus to 3,006, and finally 1,711 were from Spanish media. The purpose was not to create an exhaustive corpus for supervised training, but to test if prior clustering can help automatic detection of narrative positions without fine-tuning.

The selected fakes were organized into two semantic groups using KMeans (k=2). Then the news were vectorized and projected onto the same semantic space, with extra filtering by theme and geography. Cosine similarity was used to measure closeness to the fake narratives.

For narrative classification, the model mDeBERTa-v3-base-mnli-xnli (Laurer et. al. 2022) was used. Each news item was compared with representative statements of the fake, such as “Climate change has no relation with human activity”. The categories were SUPPORT, REFUTE and IGNORE, equivalent to entailment, contradiction and neutral.

Experimental Design: The evaluation included three experimental scenarios to assess the impact of clustering and semantic filtering on classification performance:

- **Scenario A**: news with similarity ≥ 0.5 to fake narratives
- **Scenario B**: all news in the climate cluster
- **Scenario C**: the complete corpus without clustering

Manual labeling of 40 random news per scenario was used for validation. This approach enabled systematic comparison of classification accuracy across different levels of thematic organization and semantic filtering.

6.9. Stance detection methodology and data

This study evaluates in-context learning for multilingual stance detection with generative LLMs, minimizing fine-tuning. We examine how prompt design, model family, and model size affect performance across two datasets and three languages, using five prompting schemes and four instruction-tuned LLMs. Unless noted, experiments run in zero-shot settings.

All experiments were executed on a single MareNostrum 5 Accelerated node (4×NVIDIA H100 64 GB, Intel Sapphire Rapids 8460Y+, 512 GB RAM), enabling large-scale inference without fine-tuning.

Datasets: We use SemEval-2016 Task A (English) (Mohammad et. al. 2016) and the Catalonia Indepen- dence Corpus (CIC) in Catalan (CIC-CAT) and Spanish (CIC-ES) (Zotova & Agerri 2021). Both datasets contain political tweets with target topics and ground truth labels. SemEval covers five targets (Atheism - AT, Climate Change is Real - CC, the Feminist Movement - FM, Hillary Clinton - HC, and the Legalization of Abortion - LA); we evaluate on the official test split (1,249 tweets). The CIC dataset contains tweets regarding the Catalonian Independence Movement. CIC-CAT and CIC-ES each contain ~10k tweets labeled favor /against /none; we use only the test partitions (~2k tweets per language).

LLM Models Tested: We evaluate four open-source multilingual, instruction-tuned LLMs1[13] via Hugging Face[14]: *Mistral-7B-Instruct-v0.3, Mistral-Small-24B-Instruct-2501, Qwen2.5-7B-Instruct, and Qwen2.5-72B-Instruct.* Inference uses max sequence length 250 and greedy decoding. Prompts follow each model's chat template; batches are left-padded, and decoded after removing the prompt prefix.

Prompting Schemes. We test five schemes: zero-shot (Kojima et. al. 2022), few-shot (Brown et. al. 2020), chain-of-thought (CoT) (Wei et. al. 2022), COLA multi-agent prompting (Lan et. al. 2024), and a modified Chain-of-Stance (mCoS) (Ma et. al. 2024). The original Chain-of-Stance method proved to be unreproducible, therefore an inspired version was created that simplified the prompts and kept the core idea of each step. Prompts are written in the dataset language (EN for SemEval; ES/CAT for CIC). We additionally run a cross-lingual condition where all prompts (and in-context examples) are translated to English.

Baselines: To contextualize LLM performance, we compare against supervised and in-context baselines. Supervised baselines include TF-IDF+SVM (Sánchez & Martín n.d.) with feature selection via Information Gain (Cover & Thomas 2005), plus fine-tuned transformer encoders (mBERT (Devlin et. al. 2019) and XLM- R (Conneau et. al. 2020)). In-context baselines include a plain zero-shot instruction (Kojima et. al. 2022) and Stance Reasoner (Taranunkhin et. al. 2024), which augments few-shot

[13] For brevity, we refer to Mistral-7B-Instruct-v0.3 as *Mistral-7B* and Mistral-Small-24B-Instruct-2501 as *Mistral-24B*. Like- wise, Qwen2.5-7B-Instruct is denotes as *Qwen-7B* and Qwen-72B-Instruct as *Qwen-72B* throughout the article.
[14] https://huggingface.co

prompts with premise→conclusion reasoning; for comparability we cite results from (de Landa & Agerri 2025) on SemEval-2016 and CIC.

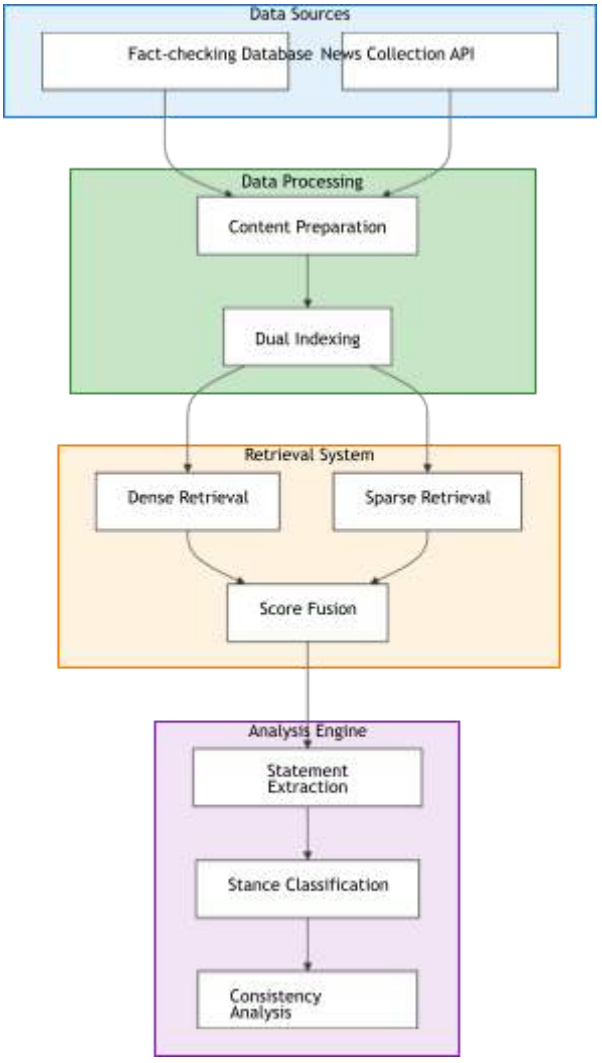
Evaluation Metrics: Following SemEval-2016, we report the average F1 over favor and against, excluding none: $F_{avg} = (F_{favor} + F_{against})/2$. This emphasizes explicit stance expression and avoids ambiguity in none since it is rarely a truly neutral stance.

6.10. Stance Detection Pipeline with Hybrid Retrieval architecture

The stance detection pipeline follows a modular architecture that integrates multiple data sources and processing components. The system processes data through four main stages:

(1) Data Sources collect fact-checks and news articles from APIs, (2) Data Processing performs content preparation and dual indexing, (3) Retrieval System combines semantic and keyword-based search with score fusion, and (4) Analysis Engine extracts statements, classifies stance, and performs consistency analysis to contextualize results within the overall article. This process is illustrated in Figure 30.

Figure 30: Stance Detection Pipeline Architecture.



The architecture demonstrates a systematic approach where fact-checking data from the IBERIFIER consortium and news articles from MYNEWS are first processed and indexed using both vector embeddings and traditional text search methods. This dual-indexing strategy enables the hybrid retrieval system to leverage both semantic similarity and exact keyword matching. The analysis engine then processes news articles by extracting individual statements, determining their fact-checkability, retrieving relevant fact-checks, and performing stance classification. The final consistency analysis step addresses potential context loss by evaluating statement-level results within the broader article context, ensuring that articles reporting disinformation for refutation purposes are not incorrectly classified as supporting false claims.

6.11. Stance Detection Pipeline Methodology

The system is built on three core principles that address the practical challenges of automated misinformation detection in multilingual European media environments.

- **Evidence-based verification**: The pipeline compares media content directly against a database of verified fact-checks rather than relying solely on machine learning patterns. The system leverages fact- checking work completed by established fact-checking organizations that have systematically documented and verified misinformation claims. This approach ensures that assessments are based on expert human judgment rather than algorithmic pattern recognition alone. Each analysis can be traced back to specific fact-checks, providing transparency about how conclusions were reached.
- **Statement-level analysis**: To classify the article, the system breaks down articles into individual factual statements for separate analysis. This approach recognizes that news articles often contain both accurate information and problematic claims within the same text. By analyzing statements individually, the system can identify specific misleading content while preserving the context of surrounding accurate reporting. This granular approach provides more useful information for both researchers studying misinformation patterns and policymakers developing targeted responses.
- **Consistency-level analysis**: An issue raised by the statement-level analysis is context lost. For in- stance, sometimes articles report a disinformation piece to be able to refute it. In this scenario, if the analysis is solely statement-level, it will be flagged as Supporting disinformation. The Consistency-level analysis allows to recontextualise the stance detection result within the overall article and give a result that is representing the overall consistency of the article rather than just the included statements.

Data processing: The pipeline handles fact-checking data from the IBERIFIER consortium fact-checking organizations[15]: EFE Verifica, Maldita.es, Newtral, Polígrafo, Verificat, Chequeado, and Infoveritas. Each fact-check document undergoes text extraction by concatenating title and content fields, followed by metadata extraction including keywords, formats, sources, categories, type, link, organization, and organization qualification with explanation. Documents are then split into manageable segments using RecursiveCharacterTextSplit- ter with configurable chunk size (1000 characters) and chunk overlap (500 characters) to preserve context across boundaries.

For enhanced analysis, each chunk is processed through an LLM (Qwen2.5-32B-Instruct) to extract structured metadata including main topic and named entities such as people, places, and organizations. The final step merges document-level metadata, chunk-level metadata (chunk index and character count), and LLM-extracted metadata into a unified structure that is then serialized for storage.

The storage process employs a dual-indexing strategy that enables both semantic and keyword-based retrieval capabilities. Text content undergoes careful segmentation into manageable chunks of approximately 1000 characters with 250-character overlaps between segments to preserve contextual relationships across boundaries. This segmentation process respects natural text structure by avoiding breaks within sentences or paragraphs, ensuring that semantic coherence is maintained throughout the chunking process.

The dual indexing architecture consists of two complementary storage systems. For semantic retrieval, chunks are processed through the Jina-v4 multilingual embedding model, which generates high-dimensional vector representations of the text content. These embeddings are normalized to unit length to ensure consistent similarity calculations and stored within a ChromaDB persistent client.

Parallel to the semantic indexing, the system creates a BM25 full-text search index using the Whoosh library[16] for precise keyword-based retrieval. This index employs a straightforward schema containing unique document identifiers and stored text content, enabling traditional information retrieval methods that complement the semantic search capabilities. The Whoosh index is persisted to a configurable directory structure, allowing for efficient keyword matching and Boolean query operations.

The system incorporates sophisticated deduplication mechanisms that track previously processed docu- ment identifiers, preventing reprocessing of existing fact-checks and enabling efficient incremental updates as new fact-checking content becomes available. Comprehensive quality control measures include error handling for failed batch operations, API connection issues, JSON parsing failures, and validation errors during the metadata extraction process, ensuring robust operation across varying data quality conditions.

[15](In order of appearance): verifica.efe.com/, maldita.es/, www.newtral.es/, poligrafo.sapo.pt/, www.verificat.cat/, chequeado.com/, info-veritas.com/
[16] github.com/mchaput/whoosh

Retrieval Methods: The system combines multiple computational approaches to improve accuracy and coverage across different types of queries and content. The dense retrieval component employs the Jina-v4 embedding model to generate multilingual vector representations that capture semantic relationships across Spanish, Portuguese, and Catalan content. ChromaDB[17] serves as the vector storage system, utilizing cosine similarity calculations for retrieval of semantically related fact-checks. The system applies configurable similarity thresholds and score adjustments to ensure consistent ranking across different query types and languages.

A BM25-based retrieval layer implemented through the Whoosh library complements the dense search capabilities by providing precise keyword matching. This component serves to match exact terms. The sparse retrieval particularly benefits queries involving specific names, dates, or technical terms that require exact matching.

Dense retrieval alone faces significant limitations in disinformation detection due to the highly specific nature of false claims. Disinformation typically involves precise details about people, locations, organizations, dates, and numerical data that are crucial for accurate fact-checking. Semantic embeddings may capture general topical similarity but can miss the specific contextual elements that distinguish between similar but factually different claims. For instance, false claims about a specific politician's actions on a particular date require exact matching of these named entities, which sparse retrieval handles more effectively than semantic similarity alone. The hybrid approach ensures that both the conceptual context and the precise factual details are captured in the retrieval process.

The final retrieval step combines outputs from both dense and sparse methods using z-score normalization. This fusion approach maximizes the strengths of both retrieval methods while minimizing their individual limitations.

Statement analysis: Large language models (Qwen2.5-72B-Instruct) are used to extract potentially fact-checkable statements. Filters exclude non-factual or speculative content. Statements are evaluated for "checkability" based on specificity, verifiability, and public interest.

Stance classification A multi-step prompting approach (Chain-of-Stance reasoning) is used to classify the relation between a statement and retrieved fact-checks. Three categories are applied: supports, refutes, not related.

Classification is accompanied by transparent reasoning chains, allowing for manual control over the AI reasoning.

Technical implementation: The system operates on HPC infrastructure using SLURM-based job scheduling for distributed processing. ChromaDB provides vector storage for semantic retrieval while Whoosh implements BM25-based keyword matching. Processing throughput achieves analysis of hundreds of articles per hour with horizontal scal- ing across multiple compute nodes. The pipeline demonstrates

[17] github.com/chroma-core/chroma

exceptional flexibility in model deployment, supporting both open-source and closed-source language models based on organizational requirements and constraints. Open-source models (such as Qwen, Llama, and Jina) can be deployed locally using vLLM12 inference engine for complete data sovereignty and cost control. The vLLM framework provides optimized serving capabilities with features like tensor parallelism, continuous batching, and efficient memory management for high-throughput model inference. Alternatively, closed-source models (such as OpenAI GPT, Anthropic Claude) can be integrated through API endpoints for enhanced capabilities and reduced infras- tructure requirements. This flexibility enables adaptation to different deployment environments, budget constraints, and data privacy requirements.

The analysis process used the HPC infrastructure at the Barcelona Supercomputing Center, and ran on 4 nodes equipped with 4x NVIDIA Hopper H100 64GB during 8 hours.

7. Conclusions

Studying social media spread patterns is a novel approach to disinformation fighting. Detecting and analysing *how*, *who* and *when* hoaxes spread across platforms is a research field still in its infancy, especially in the walled-garden ecosystem. In this regard, we have conducted a study on three common topics for disinformation (Covid, Climate change, Gender) and studied the patterns that communities adopt to spread false information.

Our first step has been to introduce a complete methodology for Telegram, flexible to other social media outlets, that relies only on content information. Traditionally, social media spread is studied using explicit links between users/messages, but this is not readily available information. Instead, this focus on content has allowed us to track down more than three hundred hoaxes and visualize their spread.

Keep in mind that our analysis is mostly done from a technical standpoint, offering insights based on data and some contextual clues. As such, one of the most glaring points of interest of this study is the heavy overlap in spread behaviour. All three topics heavily share: actors and hub/authority distribution. Many channels appear again and again across all graphs, many even unrelated to the topic; for instance, anti-vaxxers frequently post misogyny, flat-earthers are usually Covid-denialists, and so on. Overlap between topics is extremely high.

Despite each hoax having individual differences, the large-scale dynamics of propagation remains. A notable example was observed with misogyny posts, where the hoax spreads intermittently across several periods of time. Other hoaxes present very specific points in time where channels intensely cross-post in a short span. Some hoaxes have clearly defined leaders, who dominate the discourse. More disinformation dynamics can be found by analysing the fine-grain cascade graphs.

A cross-platform comparison further reinforces that while sub-topics like vaccine risks recur, the specific wording of false claims varies significantly, with no verbatim or closely-matched hoaxes found between Twitter/X and Telegram, underscoring the need for multi-platform analysis to understand the distinct yet impactful dynamics of misinformation propagation.

To conclude, this technical report aims to analyse disinformation dynamics in social media, further investigation could be performed in several ways; for instance, cleaning up visual representations of cascades, targeting known offender channels, diversifying our target social media, etc. This approach is still data-intensive, so access to social media is still a point of contention to assess the full potential of this type of technology. Our contribution to design a flexible analysis tools, despite being more versatile than before, can still be undermined by data scarcity.

8. Bibliography

Arcos, I., Rosso, P., & Salaverría, R. (2025, January 28). Divergent Emotional Patterns in Disinformation on Social Media? An Analysis of Tweets and TikToks about the DANA in Valencia. arXiv: 2501.18640[cs]. https://doi.org/10.48550/arXiv.2501.18640

Blanco, M. A., Santiago, M. P., & Cartea, P. Á. M. (2021). La sociedad española ante la emergencia climática: cognición, emoción y acción. In La comunicación del cambio climático, una herramienta ante el gran desafío (pp. 273-296). Dykinson.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901. https://doi.org/10.48550/arXiv.2005.14165

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Cover, T. M., & Thomas, J. A. (2005). Elements of Information Theory [Edition: 1]. https://doi.org/10.1002/047174882X

de Landa, J. F., & Agerri, R. (2025). Language Independent Stance Detection: Social Interaction-based Embeddings and Large Language Models. Procesamiento del Lenguaje Natural, 74, 139-157. https://doi.org/10.48550/arXiv.2210.05715

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformer For Language Understanding. Proceedings of the 2019 Conference of the North. https: //doi.org/10.18653/v1/n19-1423

Fernández Reyes, R. (2024). (n.d.). Aproximación a la contraargumentación ante el negacionismo y el retardismo climáticos. Abordaje de las trabas a la adaptación y mitigación en la comunicación climática. Zaragoza: ECODES. https://ecodes.org/images/que-hacemos/MITERD-2023/Aproximacin_a_los_argumentos_ante_la_inaccin_y_el_retardismo_climticos_DEF.pdf

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 46(5), 604-632. http://www.cs.cornell.edu/home/kleinber/auth.pdf

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Advances in neural information processing systems, 35, 22199-22213.

Lan, X., Gao, C., Jin, D., & Li, Y. (2024). Stance Detection with Collaborative Role-Infused LLM-Based Agents. Proceedings of the International AAAI Conference on Web and Social Media, 18, 891–903. https://doi.org/10.1609/icwsm.v18i1.31360

Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2024). Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with

deep transfer learning and bert-nli. Political Analysis, 32(1), 84-100. https://doi.org/10.1017/pan.2023.20

Ma, J., Wang, C., Xing, H., Zhao, D., & Zhang, Y. (2024, November). Chain of stance: Stance detection with large language models. In CCF International Conference on Natural Language Processing and Chinese Computing (pp. 82-94). Singapore: Springer Nature Singapore. https://doi.org/10.48550/arXiv.2408.04649

Meddeb, P., Ruseti, S., Dascalu, M., Terian, S. M., & Travadel, S. (2022). Counteracting french fake news on climate change using language models. Sustainability, 14(18), 11724. https://www.mdpi.com/2071-1050/14/18/11724

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016, June). Semeval-2016 task 6: Detecting stance in tweets. In Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016) (pp. 31-41). https://doi.org/10.18653/v1/s16-1003

Organización Mundial de la Salud (2022). (n.d.). Desinformación y salud pública. https://www.who.int/es/news-room/questions-and-answers/item/disinformation-and-public-health

Powell, J. (2017). Scientists reach 100% consensus on anthropogenic global warming. Bulletin of Science, Technology & Society, 37(4), 183-184. https://doi.org/10.1177/0270467619886266

Reimers, N., & Gurevych, I. (2020). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics.. https://arxiv.org/abs/2004.09813

Rodrigo-Cano, D. & Del Río Álvarez, M. J. (2021). (n.d.). Las redes de la comunicación. Estudios multidisciplinares actuales. Capítulo 39: La desinformación sobre salud y cambio climático en redes sociales. https://www.dykinson.com/libros/las-redes-de-la-comunicacion-estudios-multidisciplinares-actuales/9788413775609/

Salaverría, R. (2025, February 27). Qué hay detrás de los bulos: estructura y actores de la desinformación. https://doi.org/10.5281/zenodo.14939609

Sánchez, J. S., & Martín, I. R. (n.d.). Realidades conectadas: medios, cultura y sociedad en la era digital.Sparck Jones, K. (1972). A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL. Journal of Documentation, 28(1), 11–21. https://doi.org/10.1108/eb026526

Sánchez, J. S., Martín, I. R., & de Luna, Á. B. M. (2025). Realidades conectadas: medios, cultura y sociedad en la era digital. Realidades conectadas: medios, cultura y sociedad en la era digital. https://burjcdigital.urjc.es/server/api/core/bitstreams/4fe11281-4002-4d4e-9b82-ef293827e15c/content

Szczepański, M., Pawlicki, M., Kozik, R., & Choraś, M. (2021). New explainability method for BERT-based model in fake news detection. Scientific reports, 11(1), 23705. https://www.nature.com/articles/s41598-021-03100-6

Taranukhin, M., Shwartz, V., & Milios, E. (2024). Stance reasoner: Zero-shot stance detection on social media with explicit reasoning. arXiv preprint arXiv:2403.14895. https://doi.org/10.48550/arXiv.2403.14895

Vranić, A., Reis, J., Damião, Í., Almeida, P., & Gonçalves-Sá, J. (2025, October). Global Claims: A Multilingual Dataset of Fact-Checked Claims with Veracity, Topic, and Salience Annotations. In Proceedings of the 2nd International Workshop on Diffusion of Harmful Content on Online Web (pp. 85-94). https://doi.org/10.1145/3746275.3762201

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837. https://dl.acm.org/doi/10.5555/3600270.3602070

Zotova, E., Agerri, R., Nuñez, M., & Rigau, G. (2020, May). Multilingual stance detection in tweets: The Catalonia independence corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 1368-1375). https://aclanthology.org/2020.lrec-1.171.pdf

9. Annexes

Annex 1. The Digital Services Act (DSA) and Disinformation Research in the Iberian Context

As mentioned throughout the document, the current situation for online media accessibility is crucial. Here we review the recent EU regulation that addresses this issue directly.

A1.1. Background on DSA and disinformation

The Digital Services Act (Regulation (EU) 2022/2065) constitutes the European Union's most ambitious reform of digital governance since the 2000 E-Commerce Directive (European Union, 2022). It establishes a horizontal framework applicable to all online intermediaries while introducing differentiated obligations depending on the nature and scale of the service. Hosting providers and ordinary platforms must meet baseline duties, whereas Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs), defined as those reaching more than 45 million average monthly users in the Union, are subject to more demanding requirements. In April 2023 the European Commission designated nineteen services as VLOPs/VLOSEs (European Commission, 2023a), including Google Search, YouTube, Facebook, Instagram, TikTok, and X (Twitter), which fall under direct Commission supervision

The regulation entered into force on 16 November 2022 and its implementation has been slowly taking place. Obligations specific to VLOPs and VLOSEs have been binding since August 2023, while the general regime became applicable in February 2024. Since then, the regulatory framework has entered an operational phase. The Delegated Act on data access for vetted researchers, adopted in July 2025 (European Commission, 2025b), specifies the procedures, eligibility rules, and safeguards that give effect to Article 40 on researcher access, including the creation of a dedicated DSA Data Access Portal[18]. This platform is already online and expected to be fully functional during the last trimester of 2025. In parallel, the Implementing Regulation on transparency reporting, related to Article 42 and adopted in November 2024 (European Commission, 2024), harmonises the structure of reports and obliges VLOPs and VLOSEs to issue them every six months from July 2025 onwards, meaning that the first ones are due in early 2026. Besides these reports, the DSA Transparency Database[19] collects the reported content moderation decisions in almost real time. Naturally, while these mechanisms represent an important move toward transparency, both are at an early stage of implementation, and their practical value for the analysis of content dissemination over social networks in the Iberian context is still limited.

The DSA's enforcement is a two-fold system, shared between the European Commission, which holds exclusive supervisory powers over designated VLOPs and VLOSEs, and the Member States. At the national level, each country must appoint a Digital Services Coordinator (DSC), an independent authority responsible for supervising the intermediary services within its jurisdiction. These DSCs are endowed with powers to investigate, demand information, and impose penalties. To ensure

[18] https://data-access.dsa.ec.europa.eu/home
[19] https://transparency.dsa.ec.europa.eu

consistent application of the rules across the Union, all national DSCs form the European Board for Digital Services. This independent advisory group, established in early 2025 and chaired by the Commission, provides guidance and support for joint investigations. In Portugal, this role is held by the National Communications Authority (ANACOM); in Spain, the designated body is the National Commission on Markets and Competition (CNMC), although the Commission issued a formal referral in early 2025 concerning the full implementation of its supervisory powers under the Act (European Commission, 2025a).

Disinformation is addressed in the DSA within the framework of systemic risks, i.e., more general risks such as content that has negative effects for the exercise of fundamental rights, including freedom of expression and information. Articles 34 and 35 oblige VLOPs and VLOSEs to assess how their design and operation may contribute to these systemic risks, and to adopt proportionate mitigation measures. Failing to meet these obligations, platforms face sanctions of up to six per cent of their global turnover. The central innovation in relation to disinformation is the articulation between these duties and the Code of Practice on Disinformation (Ó Fathaigh et al., 2025). First launched in 2018 as a voluntary self-regulatory initiative and revised in 2022 to expand its commitments, the Code now operates under Article 45 of the DSA as a code of conduct. Signing the Code is voluntary, but once a provider adheres to it, the commitments are subject to monitoring and may be considered in assessing compliance with systemic-risk obligations. For platforms, the Code gives a clearer and easier way to demonstrate compliance with their obligations on systemic risks. At the same time, this dual approach—voluntary in the decision to join but linked to regulatory oversight once inside—illustrates both the possibilities and the difficulties of the European Union's strategy against disinformation. It encourages responsibility without leaving behind co-regulation, and although its results will need to be seen in practice, it can be read as a positive step to advance transparency, platform accountability and the protection of fundamental rights.

A1.2. Platforms and research access

A particularly relevant aspect of the DSA for misinformation research is Article 40, which enables independent researchers to request access to non-public, proprietary data from online platforms. This provision represents an important step toward greater transparency and accountability, allowing academics to investigate how misinformation spreads, how platform algorithms shape information exposure, and how content moderation policies are applied in practice. The implementation of Article 40, coincides with significant changes in how large platforms structure access to their data, and these developments shape the methodological options available. In this regard, the European Commission has issued requests for information to several VLOPs on the design and accessibility of their research tools, signalling that questions of scope, eligibility, and functionality are central to the DSA.

In particular, platforms are expected to provide catalogues and related information, such as codebooks, changelogs and architectural documentation, so researchers know what data is available to be requested. In parallel, initiatives such as the Social

Data Science Alliance[20] and the DSA 40 Collaboratory[21] aim to coordinate the scientific community in all stages of the DSA implementation, pooling resources, defining data and request taxonomies, and bringing stakeholders together to more effectively author data access applications and monitor compliance.

Different platforms have adapted their research access policies in distinct ways, reflecting varied interpretations of the new obligations. Table A1 summarises the research access available for a selection of platforms as of September 2025, with particular attention to academic research:

Meta has replaced CrowdTangle with the Meta Content Library and its API, available to vetted researchers through the ICPSR. It offers a structured environment but with scope and eligibility rules still under definition.

TikTok has opened a research API accessible to non-profit academic institutions in the European Union and the United States, while third-party services such as Tikapi provide additional, though unofficial, access pathways.

X (formerly Twitter) and Reddit reorganised their APIs in 2023, introducing paid tiers and discontinuing legacy academic-access routes; this has raised costs and reduced scale but still leaves possibilities for smaller projects or collaborations.

YouTube continues to offer its Data API, which grants access to metadata on videos and comments but not official bulk archives, and is subject to quotas and copyright restrictions.

Telegram, which is not reported as a VLOP, allows access to public channels through the Bot API and libraries such as MTProto/TDLib and Telethon, enabling large-scale monitoring though without standardised metadata on reposts.

In addition, while not currently identified as VLOPs or VLOSEs, it is easy to argue that large language models (LLMs), such as ChatGPT, clearly fall under the scope of the DSA: they act as large-scale intermediary services, used by millions of people in the EU, and can create systemic risks—for example, by sharing false information, helping shape opinions, or influencing elections. As LLMs become part of major search engines (such as Bing with ChatGPT) and social media platforms (such as Grok on X), they potentially further increase the social impact of existing platforms, and effectively act as main gateways to information. This regulatory blind spot should be closed in the near future.

In all, the current landscape of data access provided by these VLOPs illustrates the problem of misaligned incentives between platforms and researchers - researchers often prioritize transparency and data quality, while the platforms need to protect proprietary information and the interests of their stakeholders. This issue is compounded by the asymmetry in resources between platforms and the research and NGO communities, be they financial, legal, infrastructural, or in knowledge. Researchers are expected to provide privacy and security safeguards when receiving data they might not even know exists, especially before the release of transparent data

[20] https://social-data-science-alliance.org/
[21] https://dsa40collaboratory.eu/

catalogues. Moreover, there is a domain mismatch between researchers (more focused on research questions) and DSCs (more concerned with legal issues), which can further impair the process, leading to unnecessary limitations in the scope, quantity and duration of data requests.

We also note that the DSA's provisions on data access depend on who qualifies as a vetted researcher under Article 40, paragraph 4. A pilot program, led by the European Research Council and the European Commission's DG-CONNECT, has brought together regulators, researchers, DSCs and platforms, and this experience has been helping shape the future platform. It is expected that the first researchers will be vetted during 2025 and be able to complete the first request in the first semester of 2026. However, at present, this category is understood as academic or scientific institutions that meet statutory requirements, with requests evaluated by assigned (national or international) Digital Services Coordinators. Fact-checking organisations, despite being recognised in the EU Code of Practice on Disinformation as key actors in countering false content, are not automatically included. For formal NGOs, the DSA reinforces that they should have access to public data through Article 40(12) but, if they require access to non-public data, they must associate with researchers / research institutions or rely on public interfaces, transparency tools, or other partnerships. Not only does this asymmetry limit the diversity of actors who can directly benefit from Article 40 mechanisms, there is recent precedent of platforms rejecting such requests[22]. However, it is possible that alternative access pathways for these organisations may be explored in the future, for instance through their recognised role in monitoring the Code of Practice. The explicit legalization of alternative methods of data acquisition, including scraping, use of unofficial APIs and data donation may also help fulfill some of the needs of these organizations[23].

A1.3. Implications for disinformation research in Spain and Portugal

The gradual deployment of the DSA is beginning to shape the environment in which disinformation research takes place, although its practical contribution remains to be seen. As explained above, the most direct channel is Article 40 on data access, now specified through the delegated act adopted in July 2025. This provision establishes a formal pathway for vetted researchers to obtain non-public data from VLOPs and VLOSEs, yet, its effectiveness depends on the Digital Services Coordinators and on the implementation choices of the platforms themselves. This legal architecture offers, for the first time, a statutory basis for academic access to platform data, but also that its scope and enforcement mechanisms are still untested (Peterka-Benton, 2025).

The Transparency Database and the new six-monthly transparency reports represent a second major instrument. They provide systematic information on content moderation and recommender systems, thereby opening new opportunities for comparative

[22] https://www.politico.eu/article/x-challenges-german-court-decision-that-would-force-it-to-share-data-with-researchers/

[23] Mozilla has a comment on the DSA that includes these modalities: https://www.mozillafoundation.org/en/blog/the-digital-services-act-must-ensure-public-data-for-public-interest-research/

audits. Early studies, however, reveal inconsistencies in reporting formats and incomplete metadata, which limit their analytical utility for reconstructing disinformation dynamics in detail (Kaushal et al., 2024; Trujillo et al., 2025). Research on recommender systems also suggests that access to algorithmic explanations is too general to fully evaluate amplification patterns (Fabbri & Boratto, 2025). In short, while these transparency mechanisms advance regulatory accountability, their immediate value for empirical disinformation research is still modest.

For the Iberian context, the methodological consequences are clear. In the short term, platforms with observable public channels, particularly Telegram, continue to provide the most relevant material for analysing disinformation flows in Spain and Portugal, although the absence of explicit metadata on reposts and the opacity of cross-channel interactions requires the reconstruction of dissemination cascades, as done in the deliverable. By contrast, the restricted availability of systematic data from larger VLOPs, combined with the exclusion of fact-checking organisations from the category of vetted researchers, reinforces a structural dependence on partial datasets and heterogeneous tools. This asymmetry echoes broader concerns about the limited inclusiveness of the DSA's research provisions (Nannini, 2025).

Looking ahead, the first audit and reporting cycles scheduled for 2025–2026 are likely to refine both compliance practices and the comparability of platform data. As the first researchers get vetted and request access to platform data, it is expected that more collaborative platforms will appear, facilitating data sharing and transparency. Whether these instruments will translate into a tangible improvement for empirical disinformation research can only be evaluated after this point, but their introduction constitutes a significant step in aligning regulatory ambition with research needs.

Table A1: Research access conditions of major platforms (snapshot 2025, note that the conditions are rapidly evolving)

Platform	API availability	Data accessibility	Technical barriers	Non-technical barriers	Costs	Notes / Iberian relevance
X (Twitter)[24]	Paid/enterprise X API tiers; legacy academic access discontinued	Historical data available via paid tiers	Strict rate limits compared with former research tracks	Terms of service changes; evolving policies	Enterprise-grade	Previously central for propagation studies; now less accessible for academia
Meta (Facebook, Instagram, WhatsApp, Threads)[25]	Meta Content Library + API for vetted researchers	Structured access; transparency tools in parallel	Anti-scraping measures; metadata coverage still incomplete	Eligibility and vetting required	No fee for vetted researchers (policy-dependent)	High relevance; access formalised but selective
YouTube[26]	Official Data API	Metadata on videos, channels, comments; large-scale retrieval constrained	Quotas and processing requirements; copyright considerations	Terms of service restrictions on scraping	Free (quota extensions possible)	Relevant for climate and health narratives; longitudinal capture challenging

[24] https://developer.x.com/en/docs/x-api
[25] https://transparency.meta.com/es-es/researchtools/meta-content-library/
[26] https://developers.google.com/youtube/v3

Platform	API availability	Data accessibility	Technical barriers	Non-technical barriers	Costs	Notes / Iberian relevance
TikTok[27]	Research API for non-profit academics	Access via API; transparency resources; bulk retrieval subject to conditions; third-party tools such as Tikapi are not officially supported	Endpoint changes; technical obfuscation	Vetting and approvals required	Partnership/approval-based	Rapidly growing youth reach in ES/PT; highly focused on multimedia; recommendation system is very effective to increase engagement
Reddit[28]	Paid API since 2023 for everybody	Some third-party archives (via torrent) and tools	Rate limits and pricing barriers	ToS discourage large-scale scraping	Paid (tiered)	Smaller ES/PT footprint; relevant in niches
Telegram[29]	No dedicated research API; Bot API, MTProto/TDLib and Telethon libraries	Public-channel content retrievable	No standardised reshare graph; cross-channel mapping difficult	Legal diligence recommended for bulk collection, which may involve processing personal data	Free	Not a VLOP; central for Iberian monitoring; feasible at scale with caveats

A1.3. Annex references

European Union (2022). Regulation (EU) 2022/2065 on a Single Market for Digital Services (Digital Services Act). Official Journal of the European Union, L 277, 27 October. Web: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065 (Accessed September 2025).

European Commission (2023). Digital Services Act: Commission designates first set of Very Large Online Platforms and Search Engines. April. Web: https://ec.europa.eu/commission/presscorner/detail/en/ip_23_2413 (Accessed September 2025).

European Commission (2024). Commission harmonises transparency reporting rules under the Digital Services Act. November. Web: https://digital-strategy.ec.europa.eu/en/news/commission-harmonises-transparency-reporting-rules-under-digital-services-act (Accessed September 2025).

European Commission (2025a). Commission decides to refer CZECHIA, SPAIN, CYPRUS, POLAND and PORTUGAL to the Court of Justice of the European Union due to lack of effective implementation of the Digital Services Act. May. Web: https://ec.europa.eu/commission/presscorner/detail/en/ip_25_1081 (Accessed September 2025).

European Commission (2025b). Commission adopts delegated act on data access under the Digital Services Act. July. Web: https://digital-strategy.ec.europa.eu/en/news/commission-adopts-delegated-act-data-access-under-digital-services-act (Accessed September 2025).

Fabbri, M., Boratto, L. (2025). Auditing recommender Systems for User Empowerment in Very Large Online Platforms under the Digital Services Act. Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys'25). https://doi.org/10.1145/3705328.3748074

[27] https://developers.tiktok.com/products/research-api/
[28] https://www.reddit.com/dev/api/
[29] https://core.telegram.org

Kaushal, R., van de Kerkhof, J., Goanta, C., Spanakis, G., & Iamnitchi, A. (2024). Automated Transparency: Analysis of the DSA Transparency Database. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT). https://doi.org/10.1145/3630106.3658960.

Nannini, L., Bonel, E., Bassi, D., & Maggini, M. J. (2025). Beyond phase-in: assessing impacts on disinformation of the EU Digital Services Act. AI and Ethics, 5(2), 1241–1269. https://doi.org/10.1007/s43681-024-00467-w.

Ó Fathaigh, R., Buijs, D., & van Hoboken, J. (2025). The Regulation of Disinformation under the Digital Services Act. Media and Communication, 13(1), 142–153. https://doi.org/10.17645/mac.9615.

Peterka-Benton, D. (2025). Fighting Disinformation Online: The Digital Services Act in Context. Social Sciences, 14(1), 28. https://doi.org/10.3390/socsci14010028

Trujillo, A., Fagni, T., & Cresci, S. (2025). The DSA Transparency Database: Auditing Self-reported Moderation Actions. arXiv preprint. https://doi.org/10.48550/arXiv.2312.10269

Annex 2. Technical terms

Semantic similarity: *Semantic similarity scoring* is achieved by using *semantic embeddings* to extract similarity scores through *cosine distance*. Step by step: First, a transformer converts a phrase into a numerical representation (sequence of real numbers). Using the metric known as cosine distance we can measure how different two numerical representations of texts are. Distance and similarity are inverses of each other, therefore easy to infer using any of them.

Another approach to similarity is the cross-encoder strategy, used for fine-grained retrieval. This is another transformer model that receives two sentences simultaneously and outputs a single score value. As this transformer has been trained on a similarity identification task, the cross-encoder will tell how similar both sentences are.

IBERIFIER – Iberian Digital Media Observatory

IBERIFIER is a digital media observatory in Spain and Portugal funded by the European Commission, linked to the European Digital Media Observatory (EDMO). It is made up of thirteen universities, five fact-checking organizations and news agencies, and five multidisciplinary research centers.

Its main mission is to analyze the Iberian digital media ecosystem and tackle the problem of misinformation. To do this, it focuses its research on five lines of work:

1. Research on the characteristics and trends of the Iberian digital media ecosystem.

2. Development of computational technologies for the early detection of misinformation.

3. Fact-checking of misinformation in the Iberian territory.

4. Strategic reports on threats of disinformation, both for public knowledge and for the authorities of Spain and Portugal.

5. Promotion of media literacy initiatives, aimed at journalists and informants, young people and society as a whole.

For more information, please visit the project website at iberifier.eu

Website: iberifier.eu

X: @iberifier

Instagram: @iberifier

LinkedIn: IBERIFIER

Contacts

Report coordinators:
Javier Huertas (javier.huertas.tato@upm.es)

IBERIFIER coordinator:
Ramón Salaverría (rsalaver@unav.es)

IBERIFIER
Iberian Digital Media Observatory

www.iberifier.eu

Coordinator

Universidad de Navarra | SCHOOL OF COMMUNICATION

Partners

Associate

Iberifier – Iberian Digital Media Observatory has received funding from the European Commission under the Call DIGITAL-2023-DEPLOY-04, European Digital Media observatory (EDMO) — National and multinational hubs, with the reference IBERIFIER Plus - 101158511

Co-funded by the European Union