COMMUNICATION & SOCIETY

Miscellaneous

A fact-checking tool based on Artificial Intelligence to fight disinformation on Telegram

Fco. Javier Cantón-Correa
https://orcid.org/0000-0002-8466-1679
javicanton@ugr.es
Univ. Internacional de La Rioja / Univ. de Granada

Lucia Ballesteros-Aguayo
https://orcid.org/0000-0003-1191-4070
luciaballesteros@uma.es
Universidad de Málaga

Andrés Montoro-Montarroso
https://orcid.org/0000-0003-1893-3346
andres.montoro@uclm.es
Universidad de Castilla-La Mancha

Submitted
April 17th, 2024
Approved
December 10th, 2024

Communication & Society
ISSN 0214-0039
E ISSN 2386-7876
www.communication-society.com

2025 – 38(1)
pp. 247-262

How to cite this article:
Cantón-Correa, F. J., Ballesteros-Aguayo, L. & Montoro-Montarroso, A. (2025). A fact-checking tool based on Artificial Intelligence to fight disinformation on Telegram, *Communication & Society 38*(1), 247-262. https://doi.org/10.15581/003.38.1.019

Abstract

This article develops an automatic detection model based on natural learning by designing a useful tool for disinformation monitoring in Telegram that avoids the algorithmic bias of artificial intelligence. It is of relevance the monitoring of online platforms and messaging applications such as WhatsApp and Telegram because they have become tools for circumventing traditional verification controls, and, therefore, for conveying disinformation content. The goal of this work is to contribute with early warning mechanisms that allow early combatting of disinformation (prebunking), especially in media ecosystems prone to the dissemination of false content such as pre-election periods, wars or energy crises. The methodology used applies a systematic and structured approach that allows the design and development of the tool, as well as its subsequent evaluation and optimisation. The main result is the creation of an Artificial Intelligence model capable of detecting disinformation in Telegram, and which also allows the integration of these solutions in the information verification workflow through a friendly and easy-to-use interface, thus contributing to the field of fact-checking. The findings reveal lines of future research, including the adaptation of the tool for other platforms and the integration of new features.

Keywords

Semi-supervised learning, disinformation, fact-checking, Artificial Intelligence, Telegram, social networks, prebunking.

Funding

This work was supported by the IBERIFIER Plus project, co-funded by the European Commission under the Call DIGITAL-2023-DEPLOY-04, European Digital Media observatory (EDMO) – National and multinational hubs, grant number: IBERIFIER Plus – 101158511.

1. Introduction

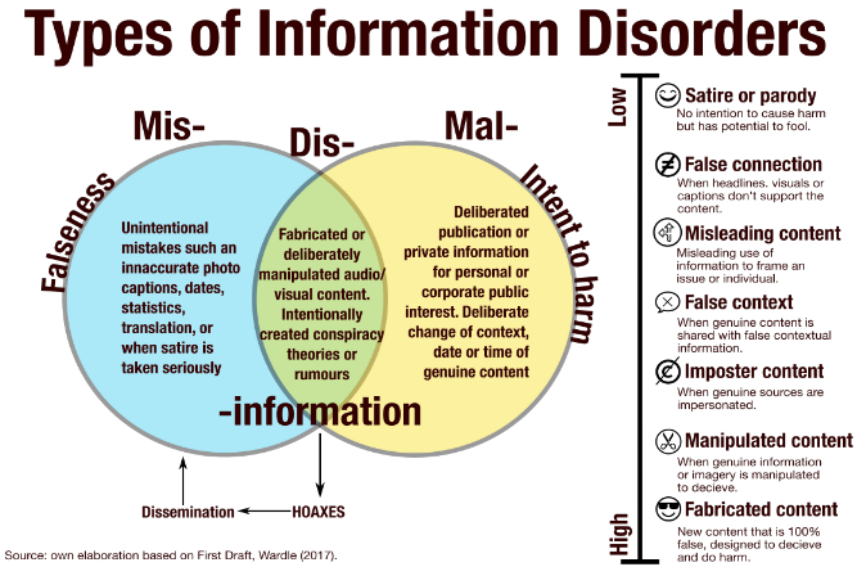
Reporteros Sin Fronteras (2023) warns in the latest World Press Freedom Index ranking of the "rise of the deception industry" in a highly volatile digital ecosystem that is driven by an industry "that shapes and distributes disinformation, while providing the tools to manufacture it."

According to EU DisinfoLab (Romero Vicente, 2023), Spain has been affected in recent years by various disinformation campaigns that promote democratic destabilization. The Report concludes that Spain is highly permeable to disinformation and almost any topic or event can be instrumentalized to generate misleading content: social and economic issues, race, gender, religion, technology, education, regional political tension, etc. Narratives often overlap –warns EU DisinfoLab– within the same hoax and are recycled from one crisis to another.

The so-called information disorders (Tandoc *et al*., 2018; Wardle & Deraskshan, 2017) are increasingly flooding the digital ecosystem around us. Beyond fake content, disinformation appears as a response to reality. Vázquez-Herrero *et al.* (2023) warn that even the most dramatic events are subject to an online jury based on interpretative polarization. The development of early warnings and pre-disinformation strategies that act preventively rather than intervening after the dissemination of false content –when it has already been virilized, shared and spread– together with the promotion of media literacy plans are key objectives to enshrining democratic health. In Spain, media such as Maldita.es or Newtral already incorporate early warning systems against disinformation, for example in Telegram[1].

Disinformation can take different forms (Figure 1). Wardle and Derakhshan (2017) set out seven general types of information disruption ranging from clickbait content, misleading content, genuine content reframed with a false context, imposter content –where an organization's logo or influential name is linked to false information– to manipulated and finally fabricated content. Manipulated and fabricated content is of relevance as it is altered or entirely false content –that is, it has no basis in fact– designed to mislead and damage. Its purpose is to deliberately distribute false information.

Figure 1. Typology of information disorders, according to Claire Wardle.



Source: Own elaboration based on First Draft, Wardle (2017).

Disinformation circulating through online platforms and messaging applications is particularly serious because of its impact on citizens' decision-making and in contexts of natural disasters

[1] This is the case of https://maldita.es/nosotros/20240930/telegram-canales-riesgos-documento-politicas/.

ISSN 2386-7876 – © 2025 Communication & Society, 38(1), 247-262

and armed conflicts, representing one of the main threats to democracy. The problem has become more pressing because of the normalization in political discourse of certain discursive practices based on disseminating disinformation to influence the actions of individuals, whether consciously or unconsciously. For Pérez-Curiel and Rivas-de-Roca (2022), a plan to rescue democracy is urgently needed, as leaders, instead of governments and parties, emerge as the dominant users of the networks and direct shareholders of computational propaganda. At the same time, a radicalized populism emerges that generates levels of fake news never seen before.

The result is, according to Zimmermann and Kohring (2020), a disinformation order that forms in opposition to the established information system to disrupt democracy. These authors researching the political debates leading up to the German parliamentary elections of 2017 point to the lack of institutional trust in the established media and politics as a crucial reason why people believe fabricated news to be true. Mistrust of traditional media and institutions causes a shift towards non-traditional or alternative media (Tsfati & Peri, 2006; Bennett & Livingston, 2018). Tsfati (2003) coins the concept of "media skepticism" as a subjective feeling of alienation and distrust toward mainstream media, that is, audience perceptions of how mainstream news institutions function in society. This breakdown of trust in the democratic institutions of the press and politics is not ephemeral but is based on the hollowing out of parties and the decline of electoral representation generating legitimacy problems in many democracies (Bennett & Livingston, 2018).

Online disinformation is also a priority for supranational bodies such as the EU or UNESCO, which implement legal policies by creating and promoting awareness-raising campaigns and improving digital literacy. This is the case, for example, of the *Code of Best Practices on Disinformation* (European Commission, 2022) and the *Digital Services Directive* (*DSA*) (Directive EU 2018) which aim to ensure a safe, predictable and trustworthy online environment.

The OMS also warns of the consequences of misinterpreting information on mental health such as polarization of opinions, increased fear and panic or decreased access to care (Borges do Nascimento *et al.*, 2022). Botha and Pieterse (2020) consider disinformation a dangerous threat to 21st century information security. Salaverría *et al.* (2004) warn about how disinformation is used to polarize and mobilize politically. Specifically, in sectors of the citizenry that align themselves with extreme and populist ideological positions, a state of disaffection spreads towards professional journalistic organizations, which are suspected of defending spurious interests, of submitting to the dictates of certain political or economic powers and of disseminating fake news (Salaverría & Cardoso, 2023).

2. Fact-checkers and early warning mechanisms

Identifying disinformation attacks in their early stages is of particular concern to media and journalism professionals. In Spain there are consortia of credible fact-checking agencies such as Maldita, Newtral, Verificat or Infoveritas. In addition, there are fact-checkers integrated in media such as EFE Verifica, Verifica RTVE, Verifica A3N or AFP Factual. These verification platforms were particularly important in the public sphere during the coronavirus health crisis (Pozo-Montes & León-Manovel, 2020; Aguado-Guadalupe & Bernaola-Serrano, 2020; García Vivero & López, 2021; Almansa-Martínez *et al.*, 2022). Data verification on platforms such as TikTok has become essential particularly due to its consideration as the preferred social network for younger people. Indeed, Digital News Report (2024) highlights the growing use of TikTok among a much younger age profile while warning that more than a quarter of TikTok users (27%) say they have difficulty detecting trustworthy news, the highest score of all the networks investigated.

The origin of fact-checking as an institutionalized practice dates to 1913, when the *New York World* newspaper founded the Bureau of Accuracy and Fair Play. The goal, in any case, is to contribute more effectively to the accountability of public representatives and better information for citizens (Ufarte-Ruiz *et al.,* 2020). In the same way, the authors add, these companies

ISSN 2386-7876 – © 2025 Communication & Society, 38(1), 247-262

favour the revitalization of journalism with the search for new business models, which shows a radically different panorama from the trends of traditional companies in recent years when launching news products on the market (Ufarte-Ruiz *et al.,* 2020). Lyons *et al.* (2020) demonstrate how politics influences opinion on data verification. Therefore, the results of this research show greater familiarity with and acceptance of data verification companies in Northern Europe (Sweden and Germany) than elsewhere (Italy, Spain, France and Poland). Furthermore, the researchers add that those less likely to trust fact-checkers may be more vulnerable to misinformation directed at these divisions, leading to a spiral of cynicism.

Data verification projects have grown exponentially in recent decades. Duke University's Reporters' Lab (Stencel *et al.,* 2023) estimates that by 2022 there were a total of 424 fact-checking organizations, a far cry from the 11 projects that existed in 2008. However, the same Duke observatory notes a slowdown in the emergence of such initiatives which in recent years have been growing at a slower pace, despite growing concerns around the world about the impact of manipulated media, political lies and other forms of deception and dangerous rumours. Among the causes of this slowdown, the authors point to the impact of the pandemic that contributed to slower growth, the fact that in some countries the fact-checking audience may be somewhat saturated –as of June, 71 countries had more than one fact-checker– and the challenge of implementing fact-checking initiatives in places with repressive governments, limited press freedom and concerns for the safety of journalists. In 2023 the Duke Reporters' Lab counted 417 fact-checkers who are active verifying and debunking misinformation in more than 100 countries and 69 languages. The average lifetime of an active fact-checking site is less than 6 years (Stencel *et al.,* 2023).

Solutions to combat false information range from user analysis, content analysis, propagation analysis and hybrid approaches. Ufarte-Ruiz *et al.* (2018) warn that fact checking ensures that journalistic texts are checked against reliable sources, official documents and solvent research results. Gupta *et al.* (2022) propose the creation of a stand-alone news verification service, i.e. an interface that provides an in-app (for mobile message exchange applications) or on-platform (for social networking platforms) news verification service for end-users. The objective, according to Gupta *et al.* (2022) for this service is to empower end-users to verify the authenticity of any news or content they encounter on social networks or in any message they receive. It, therefore, consists in building a contextual database of credible news information organized hierarchically according to location and *trending topics*.

It is essential to develop best-practice protocols to avoid the algorithmic bias of Artificial Intelligence (AI) through trustworthy models, i.e., to rely on explainable models for the automatic detection of disinformation to avoid uncertain conclusions (Schütz. *et al.*, 2023). The media are aware of the importance of data verification to achieve quality journalism and, consequently, journalistic projects devoted to this activity have skyrocketed in recent years (Ufarte-Ruiz *et al*., 2018).

In short, it is about training interpretable models with journalistic management based on verified data, true facts, reliable sources and cross-checking of information (Porlezza, 2023). One of the objectives of these models is to configure early warning tools in media ecosystems prone to disinformation, such as in the case of wars, electoral periods, or energy crises, to act not when disinformation has spread, but in media environments and ecosystems sensitive to the multiplication of false content: instead of debunking or unmasking disinformation, try to anticipate it through prebunking or anticipation. In the words of Cartwright *et al.* (2022), the decisive goal is to develop an integrated set of machine learning algorithms that can mobilize Artificial Intelligence to identify hostile disinformation activities in near real time.

The adoption of a multimodal approach capable of detecting and mitigating false information is still unknown due to the heterogeneous and diverse factors involved of a human,

ISSN 2386-7876 – © 2025 Communication & Society, 38(1), 247-262

technical, ethics and economic nature (Manfredi Sánchez & Ufarte Ruiz, 2020). This is compounded by the complexity introduced by the creation of human- and AI-generated multimedia content.

This paper proposes the need to contribute by providing tools to detect and consequently prevent disinformation using early warning mechanisms in a social network such as Telegram.

3. Fighting disinformation on Telegram

Although the social networks mainly monitored by verifiers are Facebook, Instagram (through the CrowdTangle tool), and X (formerly Twitter, although since June 2023, it is more challenging to monitor disinformation due to changes made to API access by the current owner), other platforms such as Kwai or TikTok are increasingly attracting the attention of fact-checkers due to their growing popularity and usability by extremist groups. The exponential increase in the use of mobile phones to access the internet and check online news has meant that messaging platforms such as WhatsApp and Telegram have become tools for circumventing traditional verification checks and, thus, for carrying disinformation content.

Recent research (Baumgartner *et al.*, 2020; Cazzamatta & Santos, 2023; Santini *et al.*, 2021) warns of the use of messaging platforms by extremist parties, especially in pre-election periods, as channels for disseminating disinformation. Due to their characteristics, these tools, which share their own narratives, act outside the mainstream media and allow them to speak directly to citizens. Specifically, WhatsApp and Telegram played a relevant role in Brazil's 2018 and 2022 elections. Salaverría *et al*. (2020) show how closed social networks such as WhatsApp concentrated most of the hoaxes in Spain.

Santini *et al.*, (2021) highlight the role of WhatsApp groups in the propagation of disinformation and the polarisation of voters in electoral campaigns in India and Brazil. These authors also consider that the possibilities of WhatsApp create the socio-technical conditions for audience fragmentation and *microtargeting* strategy. Cazzamatta and Santos (2023) highlight the important role of Telegram during the second round of the 2022 Brazilian elections among Bolsonaro's supporters to spread falsehoods that undermined the election results and incited outrage. Cartwright *et al.* (2022) also warn especially of Russia's use of such tools. Therefore, a complicating factor is the adoption of Russian strategies and techniques by far-right domestic groups.

Baumgartner *et al.* (2020) warn that mobile messaging platforms such as WhatsApp, Telegram and Signal can have deceptively large user bases and are often used as a meeting place for extremists. This is the case of far-right movements in Germany and the UK, which increasingly use public Telegram channels and group chats to spread hate speech, disinformation and conspiracy theories (Bovet & Grindrod, 2022; Gursky *et al.*, 2022; Santini *et al.*, 2022; Schulze *et al.*, 2022).

Other works (Muhammed T & Mathew, 2022; Sosa & Sharoff, 2022) show how Telegram has become a launching pad for disinformation, channels that serve as a "testing ground" for disinformers who, having verified the potential virality of a given message, turn to public social networks to try to spread the disinformation they have created to audiences who are less biased than the ones in their own channels.

This explains why many of the fake accounts or actors spreading hoaxes migrate to these types of platforms, also known as "fringe" or alternative platforms. Hence, there is a desirability to integrate these channels into the monitoring processes of information verifiers.

This study proposes the monitoring of booming online platforms such as Telegram using advanced natural language processing and machine learning techniques. The aim is to provide an effective solution that not only addresses the problem of disinformation on Telegram but also contributes significantly to the field of fact-checking.

The overall objective of the project is to develop software to monitor, analyse and detect disinformation on Telegram. This tool aims not only to identify false or misleading content, but

ISSN 2386-7876 – © 2025 Communication & Society, 38(1), 247-262

also to provide users and information verifiers with robust tools to combat the spread of disinformation on the messaging platform.

The specific objectives of the project are: first, to develop a useful tool for disinformation monitoring in Telegram; second, to create an Artificial Intelligence model, trained using this tool, capable of detecting disinformation early (prebunking); and third, to integrate these solutions into the information verification workflow through a user-friendly and easy-to-use interface. The tool is expected to demonstrate accuracy and efficiency beyond basic filtering techniques, showing potential for application across multiple platforms. Based on these objectives, this study addresses the following research questions:

- How effective is the proposed AI model in identifying and mitigating disinformation within Telegram channels?
- What are the key challenges and opportunities in integrating this AI tool into existing information verification workflows?
- How adaptable is the tool to other messaging platforms, and what adjustments might be necessary for broader applicability?

4. Methodology

The methodology adopted for this project is based on a systematic and structured approach. It starts with an introduction and justification phase, followed by the definition of the project phases. These phases include the identification of requirements, the design and development of the tool, and its subsequent evaluation and optimisation. Throughout the project, various tools and techniques have been used, including open-source solutions and specific tools for data analysis and processing in Telegram.

This is not the first time that fact-checking models based on automation tools (AI) have been used. Thus, several research efforts (Cartwright *et al.*, 2019, 2022) used the web crawling software tool TDC to capture web content from the open and dark web, as well as structured content from online discussion forums and various social media platforms. Therefore, these researchers developed an artificial intelligence tool to quickly and accurately help identify early-stage disinformation attacks, especially those that were orchestrated by the Russian Internet Research Agency during the 2016 US presidential election campaign.

Algorithmic solutions have proven effective in combating disinformation, as AI has proven useful in automating the fact-checking process, particularly through machine learning, natural language processing and other AI subfields (Jiang *et al.,* 2021; Kertysova, 2018; Santos, 2023).

Some (Demartini *et al.,* 2020) call for a combined system using similar algorithmic and data-driven methods to detect disinformation and control its spread. That is, combining automatic and manual checking approaches to combat the spread of disinformation on the Internet. In addition, the combination of AI and blockchain technologies leads to a more efficient computing system to combat disinformation (Santos, 2023).

To develop an effective tool, it is essential to understand not only the problem itself, but also the context in which the solution will be used. This involves identifying the target users, understanding their needs and expectations, and recognising the typical scenarios in which the tool will be used.

The disinformation monitoring tool on Telegram is designed to be used by a variety of actors interested in combating disinformation:

- **Researchers**: Academics and other professionals studying the spread of disinformation and looking for tools to collect and analyse data from platforms such as Telegram.
- **Journalists and *fact-checkers***: Media professionals who aim to verify the authenticity of information circulating on Telegram and similar channels.
- **Organisations**: Government entities, NGOs and other organisations working to combat disinformation and requiring tools to monitor and to respond to disinformation campaigns in real time.

ISSN 2386-7876 – © 2025 Communication & Society, 38(1), 247-262

It is also envisaged for use at different points in the information verification workflow, with daily monitoring, where the tool is used to actively monitor Telegram channels and groups, identifying and alerting about possible hoaxes or disinformation campaigns in real time, being common scenarios for its use; for retrospective analysis, where users can use it to analyse historical data by detecting patterns, trends and key players in the spread of disinformation in a given period; or for *fact-checking*, in situations where a specific news item or claim arises, the tool can be used to track its origin and spread on Telegram, assisting in the *fact-checking* process.

The methodology for requirements identification focused on a systematic and collaborative approach, ensuring that software features were aligned with the needs and expectations of users and other stakeholders. Collaboration with disinformation experts, including academics, journalists and practitioners, was fundamental to understanding the valuable functionalities required in such a tool. To this end, a thorough process of semi-structured interviews with fact-checkers and journalists specialised in fact-checking was carried out, as well as surveys and participant observation in various fact-checking organisations, thanks to the authors' participation in the European IBERIFIER[2] project, which allowed for a deep dive into the specific challenges of combating disinformation in Telegram.

This disinformation monitoring tool on Telegram has been designed with several essential requirements in mind, covering both specific capabilities and general features. In terms of functionalities, priority has been given to the ability to extract messages in real time, analyze their content to detect disinformation and automatically classify them using AI algorithms and NLP techniques. In addition, emphasis has been placed on creating an intuitive user interface that facilitates interaction and data analysis.

To ensure that the tool meets the quality expectations of its primary users, the development process includes a structured evaluation based on recognized software quality standards (Piattini Velthuis *et al.,* 2019). This evaluation covers key criteria such as functional adequacy, performance efficiency, usability, reliability, and maintainability. Feedback from professionals in fact-checking and data verification will be collected through a survey combining numerical scores and open-ended responses, providing both quantitative and qualitative insights into the tool's performance and areas for enhancement. This process is designed to guide iterative improvements and align the tool's functionality with real-world needs in disinformation monitoring.

In terms of the overall qualities of the tool, special attention has been paid to ensuring efficient performance, capable of handling large volumes of data without delays. Data security and privacy have been considered critical given the sensitive nature of the information handled. Usability has been treated as a key aspect, making the tool accessible even to non-technical users. In addition, scalability and adaptability have been identified as necessary to accommodate constant changes in the disinformation environment.

Finally, integration requirements have been defined to ensure that the tool can cooperate efficiently with other platforms and systems. This includes compatibility with other monitoring mechanisms, options to export data for further analysis and the provision of application programming interfaces (APIs) for broader integration. These aspects reflect the complexity of the disinformation problem in Telegram and the need for a flexible, adaptable and user-centric tool to effectively address it.

5. Results

The resulting tool (publicly available in a GitHub repository[3]), based on a modular architecture (a design that divides a system into independent components, modules, that can be changed or upgraded without affecting the rest of the system., as shown in Figure 2), offers a comprehensive solution for monitoring, analyzing and detecting disinformation in Telegram. The structure and

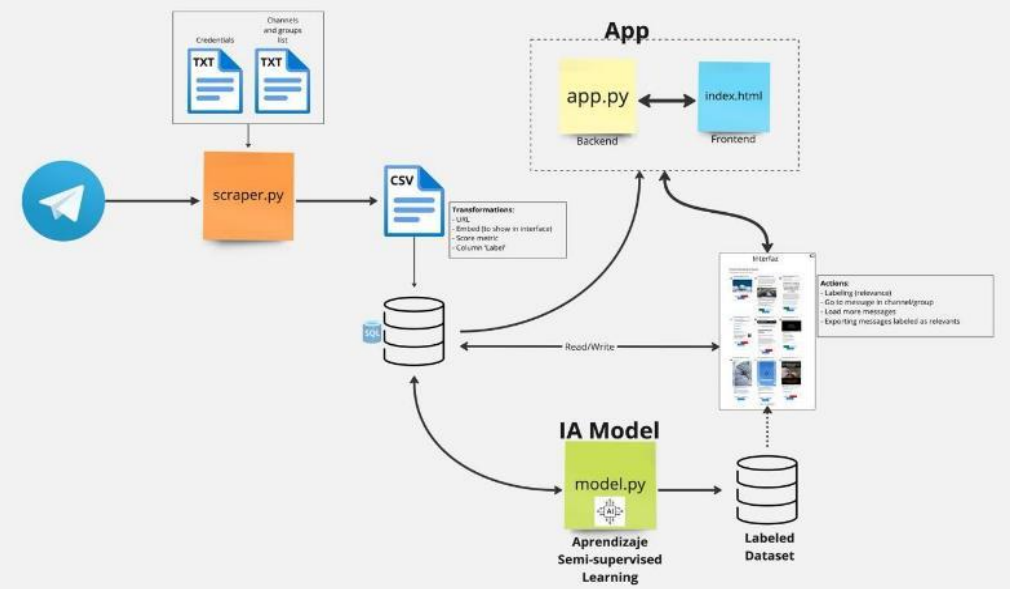
[2] https://iberifier.eu.
[3] https://github.com/javicanton/Telegram-app.

ISSN 2386-7876 – © 2025 Communication & Society, 38(1), 247-262

interaction between the different components, programmed in Python, HTML and CSS, is detailed below:

1. Telegram Scraper (scraper.py): extracts messages from different Telegram channels, using the Telegram API[4] to obtain messages and channel details from a closed list of channels and groups, in text format, provided by the user (who also must provide credentials for access to the Telegram public API). Messages are grouped by date and a daily average of views is calculated, saving the extracted data in CSV and Excel files.
2. Main application (app.py): uses the Flask framework to create an HTML web interface that displays Telegram messages. These messages are loaded from a CSV file and sorted according to a certain score called Overperforming Score, which calculates the performance of the messages of the monitored channels compared to the daily average. The interface allows the messages to be classified as relevant or not for the development of the AI model.
3. Classification model (model.py): uses a logistic regression model to classify Telegram messages. Message text is cleaned, tokenized, lemmatized and converted into TF-IDF vectors[5]. The model is trained with a labelled set through the interface and used to predict labels for unlabeled messages. Performance metrics such as precision, completeness and F1 are calculated and visualized using confusion matrices and ROC curves[6].

Figure 2. Summary diagram of the developed tool, with the interaction between its components.



Source: Own elaboration. IInteractive version retrieved from Miro: https://miro.com/app/board/uXjVMlzxfjk=/?share_link_id=273212052029.

The interaction between these components is essential for the functioning of the tool. The scraper (scraper.py) feeds the database with new Telegram messages. The main application (app.py) acts

[4] An API (Application Programming Interface) allows different software applications to communicate with each other. In this case, the Telegram API enables automated access to data from Telegram channels and groups, such as retrieving messages and channel details.

[5] Tokenization splits text into individual words or phrases, lemmatization reduces words to their base form, and TF-IDF assigns importance to words based on their frequency within a document relative to other documents.

[6] Performance metrics assess a model's effectiveness in making accurate predictions. The confusion matrix shows how often the model correctly or incorrectly classifies each category, helping to understand specific errors. The ROC curve visualizes the model's ability to distinguish between classes across different thresholds, indicating its overall classification performance.

ISSN 2386-7876 – © 2025 Communication & Society, 38(1), 247-262

as the user interface, displaying messages and allowing user interaction. Finally, the model (model.py) provides the classification capabilities, determining the relevance of each message.

The tool's interface allows users to review downloaded Telegram messages organized into individual cards, with each card displaying the message content, date, and source. A key feature, the overperforming score, is calculated based on the average daily views of each message, highlighting those that exceed typical viewership and signaling higher-than-usual engagement. Users can classify these messages as relevant or not, helping refine the tool's AI model. Additionally, each card includes a button that lets users access the message in its original context with a single click, aiding in a better understanding of the surrounding conversation. This setup enables efficient identification and categorization of potentially disinformation content based on its relative popularity and reach.

5.1. *Preliminary Validation Based on Software Quality Standards*

Following software quality standards, the tool was evaluated according to key parameters (Piattini Velthuis *et al.,* 2019): functional adequacy, performance efficiency, usability, reliability, and maintainability. This evaluation was conducted by five fact-checkers and data verification professionals (primary target users of the tool, to assess its readiness for real-world application) from the organizations Maldita.es, Newtral, and VerificaRTVE, which are widely recognized in the field of fact-checking. Feedback was collected through a survey that combined a numerical rating (on a 1-to-6 scale) with open-ended comments, allowing participants to provide quantitative and qualitative insights, providing specific feedback on each parameter:

- Functional Adequacy (average score 5.6/6): Functional adequacy assesses whether the tool fully and accurately delivers the functionalities it was designed for. The evaluation confirms that the tool successfully achieves its main purpose of monitoring and identifying potentially harmful disinformation on Telegram. Evaluators highlighted its user-friendly interface and practical functionalities. According to one evaluation, "The tool is easy to use and effective, though it could benefit from additional filters for identifying the most impactful messages in each channel." This feedback suggests that while the tool excels in its current state, enhancements like advanced filtering options could provide users with greater flexibility and improve its effectiveness in managing disinformation-related tasks.
- Performance Efficiency (average score 5.4/6): Performance efficiency evaluates the tool's speed and responsiveness in handling tasks. The evaluation rated this aspect positively, with participants praising the tool's capability to process messages and present results quickly. However, some evaluators noted slight delays when dealing with high volumes of data. As one professional pointed out, "The tool is generally efficient, but in channels with high traffic, additional optimization could improve its responsiveness." While the current performance is satisfactory for typical use, addressing these minor latency issues would ensure smoother operation under more demanding conditions and reinforce its reliability in intensive scenarios.
- Usability (average score 5.8/6): Usability measures how intuitive and accessible the tool's interface is, ensuring ease of interaction and learning. Evaluators consistently praised the interface for being clear, intuitive, and easy to navigate. One participant commented, "The layout makes it easy to find what's needed quickly and single-click access to message context is very helpful." Suggestions included adding a brief tutorial for new users to explain key features like the overperforming score. This addition would make the tool even more accessible for beginners, complementing its strong usability performance.
- Reliability (average score 5.6/6): Reliability evaluates whether the tool functions consistently and can recover effectively from potential failures. Participants reported stable performance throughout testing, with no interruptions or errors observed. A reviewer emphasized, "No errors were encountered during testing, though adding diagnostic tools

ISSN 2386-7876 – © 2025 Communication & Society, 38(1), 247-262

would be beneficial for future troubleshooting." While its stability is commendable, incorporating diagnostic features could enhance its ability to address potential issues, ensuring continued reliability under varied operating conditions.

- Maintainability (average score 4.2/6): Maintainability assesses how easily the tool can be updated, modified, and adapted for future needs. The tool's modular design was seen as a strength, facilitating updates and adaptability to changing disinformation trends. However, evaluators emphasized the need for more comprehensive documentation to support developers in making modifications. One participant noted, "The tool's design is simple and agile, but comprehensive guides for developers would enhance its usability for modifications." Improving documentation would ensure long-term maintainability and better support future enhancements.

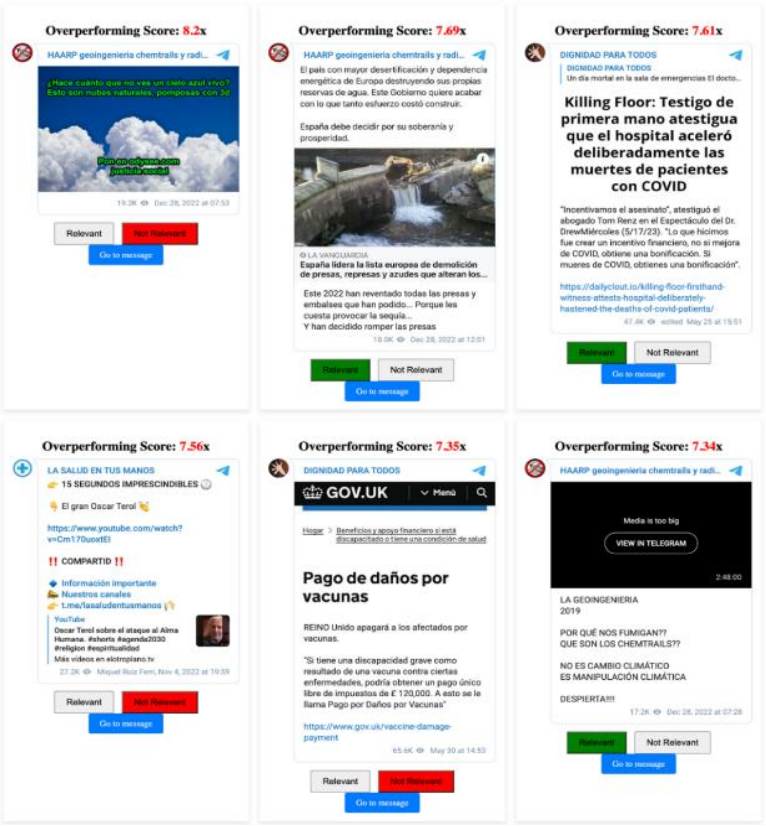
This evaluation highlights the tool's strengths in functional adequacy, usability, and reliability, demonstrating its suitability for deployment in Telegram monitoring. Recommendations such as additional message prioritization filters, diagnostic tools, and enhanced documentation align with the evaluation standards, ensuring the tool remains relevant and adaptable to evolving disinformation challenges. These preliminary findings provide a robust foundation for continuous improvement and further validation in professional verification contexts.

5.2. *Advanced AI Techniques and Future Extensions*

The tool integrates advanced Artificial Intelligence and Natural Language Processing techniques to analyze and classify the extracted content. This allows not only identifying potentially false or misleading messages but also understanding the underlying patterns and trends in the spread of disinformation. Its modular structure makes the tool scalable and easily extensible. For example, more classification models could be added, or other data sources could be integrated in addition to Telegram. The main contribution of this work is the combination of natural language processing and machine learning techniques to create a useful and practical tool that can be used in social media monitoring to fight disinformation.

ISSN 2386-7876 – © 2025 Communication & Society, 38(1), 247-262

Figure 3. A sample of the tool's interface, where the 'score' of each message, the message itself and the 'relevance' buttons for training the model are visible.



Source: Own elaboration.

Evaluation is a vital component in the development of any software project, and in the context of the fight against disinformation, its importance increases. The tool, still in its prototype phase, has been designed to be intuitive and efficient (Figure 3). The speed of data extraction and responsiveness are indicative of its performance. However, like any tool under development, it is open to improvements and adaptations. Security and privacy are of paramount importance, especially in a platform where users share personal information. In addition, ethical aspects such as ensuring that the tool does not introduce bias and respects users' rights are also considered. The contribution to the field of disinformation is evident with an innovative approach in Telegram, a less studied platform in this context.

A thorough evaluation is essential to ensure the effectiveness and relevance of the tool. A user evaluation is proposed that combines usability testing, applicability testing and feedback from experts in the field of disinformation. Through these evaluations, the aim is not only to validate the tool but also to identify areas for improvement and adaptation, thus ensuring that the tool remains relevant and effective in the evolving disinformation landscape.

This work has focused on combating disinformation on platforms such as Telegram using advanced AI techniques. The study approached disinformation from a multidimensional perspective, considering not only false content, but also the context in which it is propagated. The proposed tool, designed to monitor and verify information, has been validated through several phases and has proven to be effective in identifying and mitigating disinformation. The contributions of this study include a detailed overview of the state of the art in disinformation: the development of a practical tool and a proposal for a comprehensive evaluation.

ISSN 2386-7876 – © 2025 Communication & Society, 38(1), 247-262

6. Conclusions

We conclude that, faced with the challenge of disinformation, we are committed to trustworthy automatic detection models through the development of a useful tool for monitoring disinformation in Telegram that avoids the algorithmic bias of Artificial Intelligence, thus fulfilling the first objective set out at the beginning of this research.

This study successfully addressed its main research questions: (1) the AI model demonstrated promising effectiveness in identifying disinformation within Telegram, indicating its potential as an early warning tool; (2) integrating the tool within existing verification workflows was feasible, although further iterations may improve its ease of use; and (3) the modular design offers adaptability to other platforms, although additional adjustments may be required to maintain effectiveness in varied social media environments.

It is particularly interesting to focus the study on one of the most popular messaging platforms, Telegram, as experts (Baumgartner *et al.*, 2020; Cazzamatta & Santos, 2023; Santini *et al.*, 2021) warn of the use of WhatsApp, Telegram or Signal by extremist parties as channels for disseminating disinformation thanks to their own characteristics: they operate outside the mainstream media and allow them to talk directly to citizens.

This leads to far-right movements in Germany and the UK using such public Telegram channels and group chats to spread hate speech, disinformation and conspiracy theories (Bovet & Grindrod, 2022; Gursky *et al.*, 2022; Santini *et al.*, 2022; Schulze *et al.*, 2022). The result is that many of the fake accounts or hoax-spreading actors have migrated to these messaging platforms, hence the desirability of integrating these devices within online verifiers.

Disinformation content becomes especially virulent in certain contexts such as war conflicts, electoral periods or energy crises, hence the need to configure models based on early warnings that can act not when the information has been produced (debunking), but to try to anticipate it through early detection (prebunking). In this way, and in response to the second objective, this study creates an Artificial Intelligence model capable of detecting disinformation in Telegram in advance.

The third objective of this research was the suitability of integrating these solutions into the information verification workflow through a user-friendly and easy-to-use interface. This objective has been achieved by configuring a software tool to monitor, analyze and detect disinformation on Telegram. This device aims not only to identify false or misleading content, but also to provide users and information verifiers with robust tools to combat the spread of disinformation.

But despite its achievements, it is essential to reflect and discuss its impact and implications in the field of disinformation on Telegram. Therefore, while the tool has shown promise in its ability to identify and classify messages, there are several aspects to consider.

First, the dynamic nature of disinformation means that the tactics and strategies used by the propagators of fake content are constantly evolving. This raises the question of how the tool will adapt to these changing tactics and whether it will be able to maintain its effectiveness over time.

The contribution of this tool to the field of disinformation detection is significant. Not only does it provide a technical solution to a critical social problem, but it also offers a platform for future research, allowing other researchers and practitioners in the field to build and improve on this foundation. In short, the developed tool represents a step forward in the fight against disinformation on Telegram, offering a robust, scalable and research-based solution to address this ever-evolving challenge.

Secondly, although the tool focuses on Telegram, disinformation is a problem that affects multiple platforms. It would be interesting to discuss how the techniques and methods developed could be adapted or extended to address disinformation on other social media platforms. It is also crucial to consider the ethical implications of monitoring and analyzing

ISSN 2386-7876 – © 2025 Communication & Society, 38(1), 247-262

messages on platforms such as Telegram. While the goal is to combat disinformation, it is essential to ensure that users' privacy and rights are respected.

Future development of Telegram's disinformation monitoring tool focuses on key aspects to improve its effectiveness and usefulness. Explainability of Artificial Intelligence models is fundamental, seeking to ensure that users such as journalists and fact-checkers understand the decisions made by AI. This would increase trust and facilitate its adoption in professional environments. In addition, improvements to the user interface are planned, with the incorporation of interactive visualizations and customizable *dashboards* to enrich the user experience and facilitate the interpretation and analysis of results.

In addition, the integration of multimodality and the incorporation of the tool into the workflows of fact-checkers are areas of interest. Multimodality would allow analyzing and verifying information that combines different formats, such as text, images and videos, thus increasing the ability to detect more complex and varied disinformation. Integrating the tool into the workflows of fact-checkers would facilitate its use in real-world contexts, improving efficiency and effectiveness in detecting and refuting disinformation. These improvements and expansions open a range of possibilities in the field of disinformation and fact-checking, with potential applications in fields as diverse as journalism, communication, education and training.

While the results presented are promising, the discussion highlights the need for continued reflection and constant adaptation to ensure that the tool remains relevant and effective in combating disinformation in today's digital environment.

To support continuous improvement, this tool will undergo iterative refinements based on the feedback gathered from the preliminary validation by fact-checking professionals. Using recognized software quality standards has provided initial insights into both its strengths and areas for enhancement, ensuring the tool aligns effectively with user needs in disinformation monitoring.

The ethical risks associated with these models also need to be thoroughly assessed, while the financial barriers to the wider development of such tools need to be studied. In addition, sensitive issues related to privacy, transparency and accountability need to be analyzed. In the end, as Santos (2023) argues, it is crucial to emphasize the importance of responsible implementation and continued collaboration between technology and journalism.

Finally, this work addresses a highly relevant contemporary challenge: disinformation on the Telegram messaging platform. Through research and development, a tool has been created that combines advanced natural language processing and machine learning techniques to identify, classify and analyze potentially misleading or false messages. This tool represents a significant contribution to the field, offering a practical and research-based solution to combat the spread of disinformation.

However, as discussed above, combating disinformation is an ongoing effort that requires constant adaptation and evolution. While the tool has proven to be effective in its current state, future evaluations and iterations based on feedback will be crucial for maintaining its relevance. In addition, several lines of future research have been identified, including adapting the tool for other platforms and integrating additional features to improve the explainability of AI algorithms. Overall, this work lays the groundwork for future research and development in the field of disinformation detection, with the goal of creating a safer and more trustworthy digital environment for all users.

References

Aguado-Guadalupe, G. & Bernaola-Serrano, I. (2020). Verificación en la infodemia de la Covid-19. El caso Newtral. *Revista Latina De Comunicación Social*, *78*, 289-308. https://doi.org/10.4185/RLCS-2020-1478

Almansa-Martínez, A., Fernández-Torres, M. J., & Rodríguez-Fernández, L. (2022). Desinformación en España un año después de la COVID-19. Análisis de las verificaciones

de Newtral y Maldita. *Revista Latina De Comunicación Social*, *80*, 183-200. https://doi.org/10.4185/RLCS-2022-1538

Athira, A. B., Kumar, S. M., & Chacko, A. M. (2023). A systematic survey on explainable AI applied to fake news detection. *Engineering Applications of Artificial Intelligence, 122*, 106087. https://doi.org/10.1016/j.engappai.2023.106087

Baumgartner, J., Zannettou, S., Squire, M., & Blackburn, J. (2020). *The Pushshift Telegram Dataset* (arXiv:2001.08438). arXiv. https://doi.org/10.48550/arXiv.2001.08438

Bennett, W. L. & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, *33*(2), 122-139. https://doi.org/10.1177/0267323118760317

Borges do Nascimento, I. J., Pizarro, A. B., Almeida, J. M., Azzopardi-Muscat, N., Gonçalves, M. A., Björklund, M., & Novillo-Ortiz, D. (2022). Infodemics and health misinformation: A systematic review of reviews. *Bulletin of the World Health Organization*, *100*(9), 544-561. https://doi.org/10.2471/BLT.21.287654

Botha, J. G. & Pieterse, H. (2020). Fake news and deepfakes: A dangerous threat for 21st century information security. In B. K. Payne & H. Wu (Eds.), *Proceedings of the 15th International Conference on Cyber Warfare and Security,* Norfolk, Virginia, 12-13 March 2020 (pp. 57-66). Retrieved from http://hdl.handle.net/10204/11946

Bovet, A. & Grindrod, P. (2022). Organization and evolution of the UK far-right network on Telegram. *Applied Network Science*, *7*(1). https://doi.org/10.1007/s41109-022-00513-8

Cartwright, B., Weir, G., Frank, R., & Padda, K. (2019). Deploying Artificial Intelligence to Combat Disinformation Warfare Identifying and Interdicting Disinformation Attacks Against Cloud-based Social Media Platforms. *International Journal on Advances in Security, 12*(3 & 4), 203-222. Retrieved from https://www.iariajournals.org/security/tocv12n34.html

Cartwright, B., Frank, R., Weir, G., & Padda, K. (2022). Detecting and responding to hostile disinformation activities on social media using machine learning and deep neural networks. *Neural Computing and Applications*, *34*(18), 15141-15163. https://doi.org/10.1007/s00521-022-07296-0

Cazzamatta, R., & Santos, A. (2023). Checking verifications during the 2022 Brazilian run-off election: How fact-checking organizations exposed falsehoods and contributed to the accuracy of the public debate. *Journalism*, 14648849231196080. https://doi.org/10.1177/14648849231196080

European Commission (2022). *2022 strengthened Code of Practice on disinformation*. Shaping Europe's Digital Future. Retrieved from https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation

Demartini, G., Mizzaro, S., & Spina, D. (2020). Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities. *IEEE Data Engineering Bulletin, 43*(3), 65-74. Retrieved from http://sites.computer.org/debull/A20sept/p65.pdf

García Vivero, G. & López, X. (2021). La verificación de datos en Europa. Análisis de 5 iniciativas europeas: Maldita.es, Newtral, Pagella Politica, Les Décodeurs y BBC Reality Check. *AdComunica*, 235-264. https://doi.org/10.6035/2174-0992.2021.21.12

Gupta, A., Kumar, N., Prabhat, P., Gupta, R., Tanwar, S., Sharma, G., Bokoro, P. N., & Sharma, R. (2022). Combating Fake News: Stakeholder Interventions and Potential Solutions. *IEEE Access*, *10*, 78268-78289. https://doi.org/10.1109/ACCESS.2022.3193670

Gursky, J., Riedl, M. J., Joseff, K., & Woolley, S. (2022). Chat Apps and Cascade Logic: A Multi-Platform Perspective on India, Mexico, and the United States. *Social Media + Society*, *8*(2), 205630512210947. https://doi.org/10.1177/20563051221094773

Jiang, T., Li, J. P., Haq, A. U., Saboor, A., & Ali, A. (2021). A Novel Stacking Approach for Accurate Detection of Fake News. *IEEE Access, 9*, 22626-22639. https://doi.org/10.1109/ACCESS.2021.3056079

Kertysova, K. (2018). Artificial Intelligence and Disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. *Security and Human Rights, 29*, 55-81. https://doi.org/10.1163/18750230-02901005

Lyons, B., Mérola, V., Reifler, J., & Stoeckel, F. (2020). How Politics Shape Views Toward Fact-Checking: Evidence from Six European Countries. *The International Journal of Press/Politics, 25*(3), 469-492. https://doi.org/10.1177/1940161220921732

Maldita (2024, September 30). Los riesgos sin mitigar de los canales públicos de Telegram y pasos a seguir. Retrieved from https://maldita.es/nosotros/20240930/telegram-canales-riesgos-documento-politicas/

Manfredi Sánchez, J. L. & Ufarte Ruiz, M. J. (2020). Inteligencia artificial y periodismo - Artificial Intelligence and Journalism: una herramienta contra la desinformación. *Revista CIDOB d'Afers Internacionals, 124*, 49-72. Retrieved from https://www.jstor.org/stable/26975708

Muhammed T. S. & Mathew, S. (2022). The disaster of misinformation: A review of research in social media. *International Journal of Data Science and Analytics*, *13*(4), 271-285. https://doi.org/10.1007/s41060-022-00311-6

Newman, N. (2024, June 17). Resumen y principales conclusiones del Informe de noticias digitales de 2024. Reuters Institute. Retrieved from https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024/dnr-executive-summary

Pérez-Curiel, C. & Rivas-de Roca, R. (2022). Realities and Challenges of a Democracy in Crisis. Impact of Disinformation and Populism on the Media System. In Á. Rocha, D. Barredo, P. C. López-López & I. Puentes-Rivera (Eds.), *Communication and Smart Technologies*. ICOMTA 2021. Smart Innovation, Systems and Technologies, 259. Singapore: Springer. https://doi.org/10.1007/978-981-16-5792-4_10

Piattini Velthuis, M. G., García Rubio, F. O., García-Rodríguez de Guzmán, I., & Pino, F. (2019). *Calidad de Sistemas de Información* (5th Ed.). Ra-Ma.

Porlezza, C. (2023). Promoting responsible AI: A European perspective on the governance of Artificial Intelligence in Media and Journalism. *Communications*, *48*(3), 370-394 . https://doi.org/10.1515/commun-2022-0091

Pozo-Montes, Y. & León-Manovel, M. (2020). Plataformas *fact-checking*: las *fakes news* desmentidas por Newtral en la crisis del coronavirus en España. *Revista Española De Comunicación En Salud*, 103-116. https://doi.org/10.20318/recs.2020.5446

Reporteros Sin Fronteras. (2023, May 3). *Clasificación Mundial de la Libertad de Prensa de Reporteros Sin Fronteras (RSF)*. Retrieved from https://www.rsf-es.org/clasificacion-2023-analisis-general-los-peligros-de-la-industria-del-engano/

Romero Vicente, A. (2023). Disinformation landscape in Spain. *EU DisinfoLab*. Retrieved from https://www.disinfo.eu/wp-content/uploads/2023/03/20230224_SP_DisinfoFS.pdf

Salaverría, R., Bachmann, I., & Magallón Rosa, R. (2024). Desinformación y confianza en los medios: propuestas de actuación. *Index.Comunicación, 14*(2), 13-32. https://doi.org/10.62008/ixc/14/02Yconfi

Salaverría, R. & Cardoso, G. (2023). Future of disinformation studies: Emerging research fields. *Profesional De La información, 32*(5). https://doi.org/10.3145/epi.2023.sep.25

Salaverría, R., Buslón, N., López-Pan, F., León, B., López-Goñi, I., & Erviti, M.-C. (2020). Desinformación en tiempos de pandemia: tipología de los bulos sobre la Covid-19. *Profesional De La información, 29*(3). https://doi.org/10.3145/epi.2020.may.15

Santini, R., Salles, D., & Barros, C. E. (2022). We love to hate George Soros: A cross-platform analysis of the Globalism conspiracy theory campaign in Brazil. *Convergence: The International Journal of Research into New Media Technologies*, *28*(4), 983-1006. https://doi.org/10.1177/13548565221085833

Santini, R., Tucci, G., Salles, D., & de Almeida, A. R. D. (2021). Do You Believe in Fake After All? In Guillermo López-García, Dolors Palau-Sampio, Bella Palomo, Eva Campos-Domínguez

& Pere Masip (Eds.), *Politics of Disinformation: The Influence of Fake News on the Public Sphere* (pp. 49-66). Wiley-Blackwell. https://doi.org/10.1002/9781119743347.ch4

Santos, F. C. C. (2023). Artificial Intelligence in Automated Detection of Disinformation: A Thematic Analysis. *Journalism and Media, 4*(2). https://doi.org/10.3390/journalmedia4020043

Schulze, H., Hohner, J., Greipl, S., Girgnhuber, M., Desta, I., & Rieger, D. (2022). Far-right conspiracy groups on fringe platforms: A longitudinal analysis of radicalization dynamics on Telegram. *Convergence: The International Journal of Research into New Media Technologies*, *28*(4), 1103-1126. https://doi.org/10.1177/13548565221104977

Sosa, J., & Sharoff, S. (2022). *Multimodal Pipeline for Collection of Misinformation Data from Telegram* (arXiv:2204.12690). arXiv. http://arxiv.org/abs/2204.12690

Stencel, M., Ryan, E., & Luthor, J. (2023, June 21). *Misinformation spreads, but fact-checking has leveled off*. Duke Reporters' Lab. Retrieved from https://www.poynter.org/fact-checking/2023/duke-reporters-lab-fact-checking-census-2023/

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "Fake News": A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137-153. https://doi.org/10.1080/21670811.2017.1360143

Tsfati, Y. (2003). Media skepticism and climate of opinion perception. *International Journal of Public Opinion Research, 15*, 65-82. https://doi.org/10.1093/ijpor/15.1.65

Tsfati, Y. & Peri, Y. (2006). Mainstream Media Skepticism and Exposure to Sectorial and Extranational News Media: The Case of Israel. *Mass Communication and Society, 9*(2), 165-187. https://doi.org/10.1207/s15327825mcs0902_3

Ufarte-Ruiz, M.-J., Peralta-García, L., & Murcia-Verdú, F.-J. (2018). Fact checking: un nuevo desafío del periodismo. *Profesional De La información, 27*(4), 733-741. https://doi.org/10.3145/epi.2018.jul.02

Ufarte-Ruiz, M. J., Anzera, G., & Murcia-Verdú, F. J. (2020). Plataformas independientes de *fact-checking* en España e Italia. Características, organización y método. *Revista Mediterránea de Comunicación, 11*(2), 23-39. https://www.doi.org/10.14198/MEDCOM2020.11.2.3

Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) in view of changing market realities. *Official Journal of the European Union*, L303, 28 November 2018, pp. 69-92. Retrieved from https://www.boe.es/doue/2018/303/L00069-00092.pdf

Vázquez-Herrero, J., Negreira-Rey, M. C., & López-García, X. (2023). Misinformation on Trial: Media Coverage of a Murder, Public Conversation and Fact-Checking. *Journalism Practice*, *17*(10), 2218-2240. https://doi.org/10.1080/17512786.2022.2164328

Wardle, C. & Derakhshan, H. (2017). *Information Disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe. Retrieved from https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

Zimmermann, F. & Kohring, M. (2020). Mistrust, Disinforming News, and Vote Choice: A Panel Survey on the Origins and Consequences of Believing Disinformation in the 2017 German Parliamentary Election. *Political Communication*, *37*(2), 215-237. https://doi.org/10.1080/10584609.2019.1686095