Disinformation and vaccines on social networks: Behavior of hoaxes on Twitter

Desinformación y vacunas en redes: Comportamiento de los bulos en Twitter

José Manuel Noguera-Vivo.
Universidad Católica de Murcia. Spain.
jmnoguera@ucam.edu

María del Mar Grandío-Pérez.
Universidad de Murcia. Spain.
mgrandio@um.es

Guillermo Villar-Rodríguez.
Universidad Politécnica de Madrid. Spain.
guillermo.villar@upm.es

Alejandro Martín.
Universidad Politécnica de Madrid. Spain.
alejandro.martin@upm.es

David Camacho.
Universidad Politécnica de Madrid. Spain.
david.camacho@upm.es

This article is part of the R+D+I project Ecosystems of innovation in the communication industries: Actors, technologies and configurations for the generation of innovation in content and communication (INNOVACOM), with reference PID2020-114007RB-I00 funded by the Ministry of Science and Innovation AEI/10.13039/501100011033. Similarly, it has been carried out thanks to the support of the Ministry of Science and Education, within the framework of the FightDIS (PID2020-117263GB-100, 2021-2024) and XAI-Disinfodemics (PLEC2021-007681, 2021-2023) projects; as well as thanks to the Autonomous Community of Madrid, within the framework of CYNAMON (S2018/TCS-4566, 2018-2022); the Fundación BBVA, within the call for research teams on SARS-CoV-2 and COVID-19, with the CIVIC project: Intelligent Characterization of the Accuracy of Information related to COVID-19 (2021-2022); the European Commission, with the IBERIFIER project (2021-2023): Iberian Digital Media Research and Fact-Checking Hub (CEF-TC-2020-2: 2020-EU-IA-0252), and the Barcelona Mobile World Capital Foundation, with the DisTrack project: Tracking disinformation in Online Social Networks through Deep Natural Language Processing.

Received: 26/05/2022. **Accepted:** 29/08/2022. **Published:** 02/01/2023.

How to cite this article / Normalized reference.

Noguera-Vivo, J. M., Grandío-Pérez, M. del M., Villar-Rodríguez, G., Martín, A., and Camacho, D. (2023). Disinformation and vaccines on social networks: Behavior of hoaxes on Twitter. *Revista Latina de Comunicación Social*, 81, 44-62. https://www.doi.org/10.4185/RLCS-2023-1820

ABSTRACT

Introduction: Anti-vaccine disinformation is highly dangerous due to its direct effects on society. Although there is relevant research on typologies of hoaxes, denialist discourses on networks, or the popularity of vaccines, this study provides a complementary and pioneering vision about the anti-vaccine discourse of COVID-19 on Twitter, focused on its spreaders' behavior. **Methodology:** Given an initial sample of a hundred hoaxes (from December 2020 to September 2021) for the download of 200,246 tweets, around 36,000 tweets (N=36.292) that support or deny disinformation have been filtered through an algorithm for Natural Language Inference (NLI) to analyze their spreaders' through their metrics in the platform. **Results:** In relative numbers, the results show, among others, more hoaxes with original content (not retweets) among accounts with more followers and those verified; more irruption of disinformation as opposed to its objection by accounts created between 2013 and 2020, and the association of the acknowledgment (more presence in lists or many more followers than followed users) to the preference for denying false information instead of approving it. **Discussion:** The article shows how the typology of the accounts can be a predictive factor about the behavior of users who spread disinformation. **Conclusions:** Similar behavioral patterns of anti-vaccine discourse are revealed according to the accounts' Twitter-related indicators. The size of the sample and the techniques used give a solid foundation for other comparative studies on disinformation about health and other phenomena on social networks.

KEYWORDS:
Disinformation; Hoaxes; Vaccines; Twitter; Artificial Intelligence; Health Information; Spain.

RESUMEN

Introducción: La desinformación antivacunas tiene un gran peligro por sus efectos tangibles en la sociedad. Existen investigaciones relevantes sobre tipologías de bulos, discursos negacionistas en redes o la popularidad de las vacunas, pero este estudio aporta una visión complementaria y pionera sobre el discurso antivacunas de COVID-19 en Twitter, centrada en el comportamiento de sus propagadores. **Metodología:** Dada una muestra inicial de un centenar de bulos (de diciembre de 2020 a septiembre de 2021) para la descarga de 200.246 tuits, se han filtrado mediante un algoritmo para la inferencia del lenguaje natural (NLI) alrededor de 36.000 tuits (N=36.292) que apoyan o desmienten la desinformación para analizar a sus difusores a través de sus métricas en la plataforma. **Resultados:** En números relativos, los resultados muestran, entre otros, más bulos con contenido original (no retuits) entre las cuentas con más seguidores y aquellas verificadas; más irrupción de desinformación frente a su objeción por cuentas creadas de 2013 a 2020, y la asociación del reconocimiento (mayor presencia en listas o muchos más seguidores que seguidos) a la preferencia por negar información falsa en lugar de aprobarla. **Discusión:** El artículo muestra cómo la tipología de las cuentas es un factor predictivo del comportamiento de usuarios que expanden desinformación. **Conclusiones:** Se revelan patrones similares de comportamiento del discurso antivacunas según indicadores de las cuentas de Twitter. El tamaño de la muestra y las técnicas empleadas dan una base sólida para otros estudios comparativos en desinformación sobre salud y en otros fenómenos en redes sociales.

PALABRAS CLAVE:
Desinformación; Bulos; Vacunas; Twitter; Inteligencia Artificial; Información de Salud; España.

Received: 26/05/2022. **Accepted:** 29/08/2022. **Published:** 02/01/2023.

Translation by **Paula González** (Universidad Católica Andrés Bello, Venezuela)

1. Introduction

In September 2018, New York suffered a measles outbreak that lasted eleven months and cost the health system millions of dollars. The fight against the outbreak had an added difficulty: the anti-vaccine discourse emerged strongly in a context of distrust of the authorities and an argument based on the freedom of parents and the rejection of the pharmaceutical industry (Zucker et al., 2020). It was not the first time that anti-vax discourses had been recorded on social networks around measles, especially on Twitter and Facebook (Deiner et al., 2017), but in that year, Twitter registered a record of anti-vaccine narratives, even higher than that experienced with the coronavirus pandemic – a fact found in preliminary research for this study. We are not, therefore, facing a new phenomenon when we talk about the anti-vax discourse of COVID-19 on social networks, but we are dealing with a new opportunity to monitor behavior patterns of disinformation flows with serious direct consequences on society.

Since the start of the pandemic in 2019, disinformation has been spread both through traditional media and social networks, contributing to generating certain opinions. International organizations such as the World Health Organization or the European Union have established the fight against disinformation as one of their priority lines. We understand disinformation as "deliberately fake information, disseminated for economic, ideological, or any other reason" (Ireton and Posetti, 2018, p. 44). Within this term, we find a specific type of information, such as hoaxes, which consist of "fake messages made on social networks by users and/or groups to create a certain opinion" (Aparici et al., 2019, p.3). At present, this process of information intoxication is so visible, and such is the perception among Internet users, that disinformation is the main concern in the world regarding the use of social networks and media, and more than half of these users, 53%, explicitly indicate being concerned about this phenomenon (Knuutila et al., 2020).

Disinformation circulates very easily in online environments, especially in social networks (Kouzy et al., 2020). Previous research works show us some characteristics of the typology of hoaxes that circulate on social networks about COVID-19 (Salaverría et al., 2020) or how the creators of disinformation – with the help of bots– are more prolific than those who publish truthful information (Saby et al., 2021). Specifically, public personalities play a very important role in spreading hoaxes about COVID-19 as they are found to have a high level of engagement in social networks and become super-spreaders, especially on Twitter (Shahi et al ., 2021; Bodaghi and Oliveira, 2022). A study based on 38 million articles on the most prominent topics of COVID-19 disinformation shows that only 16.4% spread online had been verified (Evanega et al., 2021).

Misinformation about vaccines is having a high impact on public health worldwide. Recent studies have observed how misinformation has not only flowed but has also influenced people's perception of the pandemic (Islam et at., 2020; Kim et al., 2020). Specifically, and through experimental studies, there is evidence of how exposure to misinformation about COVID-19 vaccines directly affects vaccination intention (El-Mohandes et al., 2021; Loomba et al., 2021).

Twitter is presented as a fundamental tool for public conversation around COVID-19 vaccines compared to other platforms, as has previously occurred with other diseases (Zhou et al., 2015; Surian et al., 2016; Larrondo-Ureta et al., 2021; López-Martín et al., 2021). However, the dialogue is sometimes not completely horizontal, and, paraphrasing a classic network theory (Granovetter, 1973), it could be said that the logic of network communication acts by turning some weak ties into occasional strong ties, giving them greater influence. Previous studies have shown that users with verified accounts

Received: 26/05/2022. **Accepted:** 29/08/2022. **Published:** 02/01/2023.

had almost fifty times more spreading power on vaccines than unverified users (Carrasco-Polaino et al., 2021). Fake news continues to be spread, especially on Twitter and WhatsApp, one year after the start of the state of alarm in Spain (Almansa-Martínez et al., 2022) and there is an extensive body of research on vaccines and emotional content (Blankenship et al., 2018; Himelboim et al., 2020; Kummervold et al., 2021).

We also found analyses closer to the object of study of this work, such as some research on denialist discourses around COVID-19 (Morel, 2021), doubts about vaccines (Nowak et al., 2020; Thelwall et al., 2021; Subbaraman, 2021), or the popularity of vaccines and their level of controversy on Twitter (Carrasco-Polaino et al., 2021). On this last topic, the present work offers a new complementary perspective and focuses directly on the hoaxes generated by anti-vaccine users. The analysis of the physiognomy or morphology of the accounts responsible for the production and dissemination of disinformation, specifically around the anti-vax discourse of COVID-19, offers a greater perspective to understand the nature of these flows, as well as a tool to predict future behaviors in the face of other similar disinformation phenomena.

2. Objectives

The general objective of this research is to find out if the type of Twitter account influences the behavior of the disinformation flows of the anti-vaccine discourse. From this objective, other more specific ones arise, derived from the different variables under study:

1. To identify the majority form of dissemination of anti-vax discourse disinformation on Twitter (creation of original content, citation of another tweet, response to another tweet, or retweet) based on the number of followers of the account.

2. Detect differences in the forms of dissemination of disinformation between verified and unverified accounts.

3. Point out patterns of disinformation dissemination behavior based on the volume of content published in the account.

4. Compare trends between pro-hoax accounts and those that refute those messages, based on the year of creation of the Twitter accounts.

5. Discover if the presence in public Twitter lists can be a significant trait of the accounts that spread disinformation or of those that deny hoaxes.

6. Locate trends in the ratio of followed and followers in the behavior of the disinformation flow main accounts on Twitter.

2.1. Hypothesis

The reviewed literature offers us scientific evidence on the accounts disseminating hoaxes on Twitter, not only of their greater productive activity compared to other accounts, but also of significant differences in the spread of the hoax concerning their level of engagement on the Web. Hence, we assume the following research hypotheses:

H1. The morphology of Twitter accounts makes it possible to distinguish different behavior patterns in the production and dissemination of the discourse of anti-vaccine COVID-19 hoaxes.

H2. Variables such as the number of followers of the account, its status as a verified account or not, the volume of content it regularly publishes, the year the account was created, its presence on public lists, or its ratio of followed/followers, are factors that allow identifying particular behavior patterns in the communication flows of these hoaxes on Twitter.

3. Methodology

This experimental study requires a mixed methodology designed by a multidisciplinary team that applies research techniques from neuroscience (physiological emotional impact), psychology (impact on anxiety and stress), and communication (cognitive impact) to the subjects of this study.

3.1. Sample, procedure, and instruments

The collection, processing, and analysis of data consist of five phases, inspired by the FacTeR-Check methodology (Martín et al., 2021): 1) the hoaxes (in this case of anti-vaccines) are collected from the denials of fact-checkers; 2) queries are automatically created that allow searching for tweets about the considered hoaxes; 3) the Twitter API (Application Programming Interface) is used, which facilitates the download of tweets through previously defined queries; 4) an Artificial Intelligence technique known as Natural Language Inference (NLI) (MacCartney, 2009) is used to filter between those publications that affirm or contradict the hoax; 5) and, finally, the data of the users of tweets related to the hoax is downloaded.

Automatic queries in combination with the NLI use transformers as a base technology (Vaswani et al., 2017), a type of neural network that works with vector representations of the text that collect the semantic properties of the words and that take into account the context in which they occur. Based on the similarity between vectors (Huertas-García et al., 2021a; Huertas-García et al., 2021b) and the inference between them (Huertas-Tato et al., 2021), transformers have been successfully applied in different applications for the detection of disinformation.

In the first place, and for the selection of hoaxes, the denials of the Maldita fact-checking foundation have been chosen. The choice of Maldita alone as a verification medium with which to select anti-vaccine hoaxes is given by the significance of this digital medium. Since, in 2014, the journalists Clara Jiménez and Julio Montes opened a small minimum viable product in the form of a WordPress blog and a Twitter account (Maldita Hemeroteca), this verification medium was not only the turning point to motivate the creation of other verification media in Spain, but also a journalistic project that has not stopped growing. Since 2017, it has been part of the International Fact-Checking Network (IFCN) and in 2018 it was the only Spanish media selected to be in the High-Level Group on Fake News and Disinformation created by the European Union to advise on this matter. Currently, they are still endorsed by the IFCN seal (Mantzarlis, 2018).

Of the hoaxes on vaccination, this work has explored the examples that go directly against the use of vaccines for COVID-19. Furthermore, within the screening, false statements that caused problems in identifying disinformation on Twitter through a query have been eliminated. More specifically, the hoaxes eliminated have been those that contained:

1. Content that is expressed on multimedia elements (image, video, audio) but not textually since the queries allow tweets to be located only based on the words. That is, the text of the hoax must contain the false fact.

2. False accusations against a media outlet of having published a type of information on vaccines, because there are two focuses in the hoax, the media being questioned and the alleged information issued, which make it difficult to create queries and tag tweets.

3. Ambiguous statements that are taken out of context, in which the verification is a clarification of everything that happened, but not a denial of it.

4. Disinformation that is already repeated in the selection, so the same hoax does not appear several times.

Once the analysis and screening of the selected hoaxes have been carried out, for each disinformation piece verified and selected for this study, an automatic query has been generated with its most relevant keywords, together with the logical operators that allow defining the syntax of the search for tweets on the social network.

In this way, the hoax "Vaccines contain graphene oxide" would give rise to the query "(graphene vaccines oxide) OR (graphene vaccines contain) OR (graphene oxide contain) OR (vaccines contain oxide)" and would be used for the automatic search of related information in the selected social network. From these queries and the Twitter API, the Twitter posts between January 2020 and November 2021, consisting of a total of 220,246 tweets from 54 hoaxes, have been extracted and stored in a database (MongoDB).

Next, and since the inclusion of the query keywords does not guarantee that the tweets really contain disinformation, the probabilities that their content corresponds to an entailment (the tweet says the same thing as the hoax), a contradiction (the tweet denies the hoax), or it can only be set to be "neutral" (the algorithm cannot determine if the tweet affirms or contradicts it, so either it does not refer to the considered hoax, or cannot establish a correlation between its content and its semantic meaning) are inferred using the aforementioned NLI method. For this study, examples with more than 99% entailment and contradiction have been selected to compare those who spread disinformation and those who contradict it, respectively.

Finally, the data of the users who have published or disseminated the selected tweets have been collected on the platform. Specifically, the variables used for each Twitter account for this article are: the date of its creation; the number of tweets, followers and followed, and if they have the verified symbol on the social network. Those examples whose accounts have been suspended or deleted have been discarded.

Thus, the final sample is made up of user data from 36,292 tweets (originals, retweets, replies, and quotes from other tweets) from 49 hoaxes denied by Maldita (from an original selection of 100 anti-vaccine hoaxes, detected between December 30th, 2020, and September 9th, 2021, an original list available through the email of the authors). Of this amount, 15,513 posts have more than 99% entailment and 20,779 more than 99% contradiction. According to Twitter's data use policy, the texts, metrics, and users of the tweets cannot be published, but their IDs (Twitter identification numbers) can. These data can be obtained from Figshare [https://figshare.com/s/399422354519a2a51fc5] and the email addresses of the authors of this article.

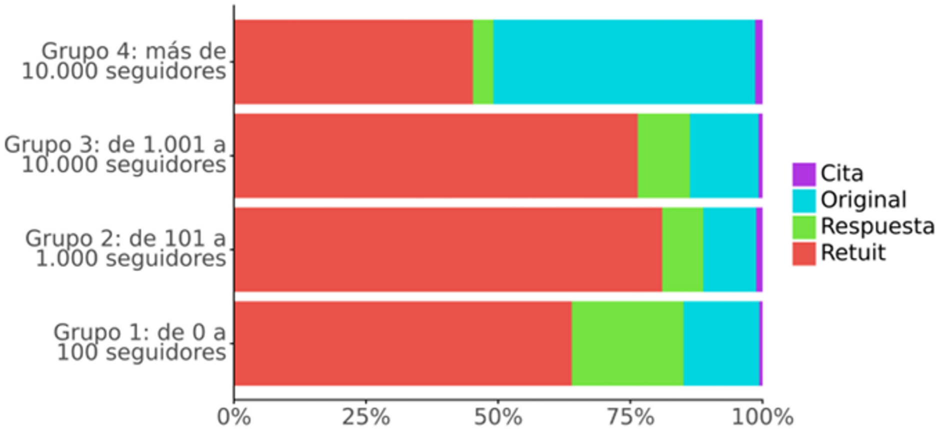
4. Results

As reflected in the hypotheses, the general objective and the consequent specific objectives indicated at the beginning, this section presents the most significant results regarding the different variables, grouped into three blocks.

4.1. According to the number of followers and verified accounts

Of the accounts that support disinformation, that is, that say the same thing as the hoax (entailment), those that develop their own content the most (in blue) are the ones that accumulate the largest number of followers since in this group we find up to 49.51% original content. On the other extreme, those with a smaller number of followers, there is greater use of retweets to spread disinformation (Chart 1), this resource being used up to 63.9% of the time within this type of account. As has been pointed out, the group of users that exceeds ten thousand followers is the one that generates the most original content regarding disinformation. In other words, the disinformation has been created originally by them, without coming from any other previous information, as occurs with other dynamics on Twitter such as direct retweets or quoted tweets, which start from a previous source. Therefore, these users generate misinformation in a very high percentage compared to other groups of users classified by their followers. The tendency of the rest of the groups with fewer followers is to retweet false information about vaccines, that is, to be loudspeakers of information that circulates through the social network and that they amplify even more, without any reflection or personal comments.

Figure 1. *Type of tweets (%) by groups with the highest "entailment", according to the number of followers.*



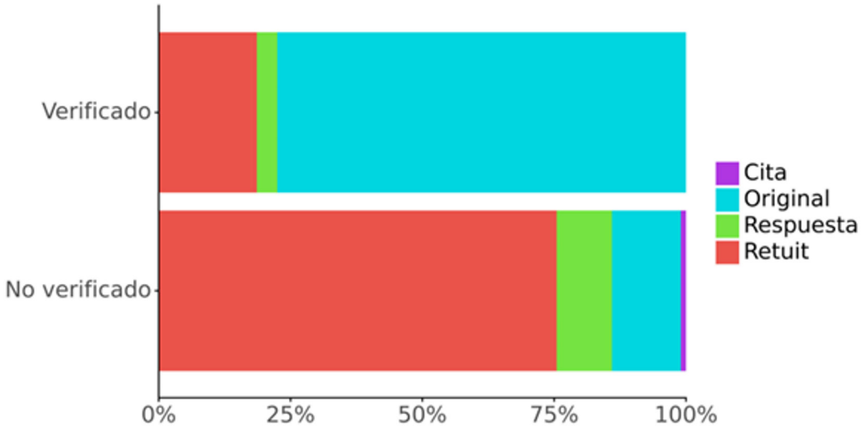
Source: Own elaboration.

Another interesting fact to highlight in chart 1 is the little commented diffusion that is made of disinformation (cited tweet, in purple in the chart), that is, the scarce comment or explanation about that information that users personally offer when they spread that content, something that appears residually in all groups, always with percentages of less than 2% (from highest to lowest number of followers, the cited tweet only appears in percentages of 1.39%, 0.77%, 1.21%, and 0.62%, respectively). This may be because the quote can act in support of the original disinformation tweet, without the need to formulate the disinformation again, since it is already expressed in the tweet that motivates its citation. In other words, the user would not need to repeat the misinformation but only show their support for a tweet that they already amplified by the mere fact of quoting it. On the other hand, this also reminds us of how little reflection or how impulsive the dissemination of content is on this social network. In other words, we are talking about a social network with difficulties in providing information with context (something that Twitter tried to alleviate in part in December 2017, with the creation of threads).

Regarding the ability to generate conversation or start a dialogue (responses to other users, in green in chart 1), it is striking how the disinformation of group 1 (from 0 to 100 followers) occurs more through replies than in other groups, since this action is found up to 21.15% of the time. It is, therefore, the group with the fewest followers that is most closely related to misinformation through responses to other users. An attempt at dialogue that should be qualified at least, since this low number of followers may indicate that we are dealing with a group of accounts made up mostly of bots.

For its part, chart 2 confirms with data one of the logical assumptions that we could have regarding the spread of anti-vax hoaxes, such as the low presence of verified accounts in the discourses since it should be noted (in the case of a percentage blocks chart) that the block of verified accounts operates on a total of just 129 units of analysis, while the block of unverified accounts operates on 15,384 units. It is striking how the verified accounts, perhaps due to their own a priori credible source condition, do not use at all (0%) the resource of the cited tweet (in chart 2, in purple), a resource that is used by unverified accounts, although it is true that also residually, barely reaching 1% (0.98%).

Figure 2. *Type of tweets (%) with the highest "entailment", according to the type of account (verified or not).*



Source: Own elaboration.

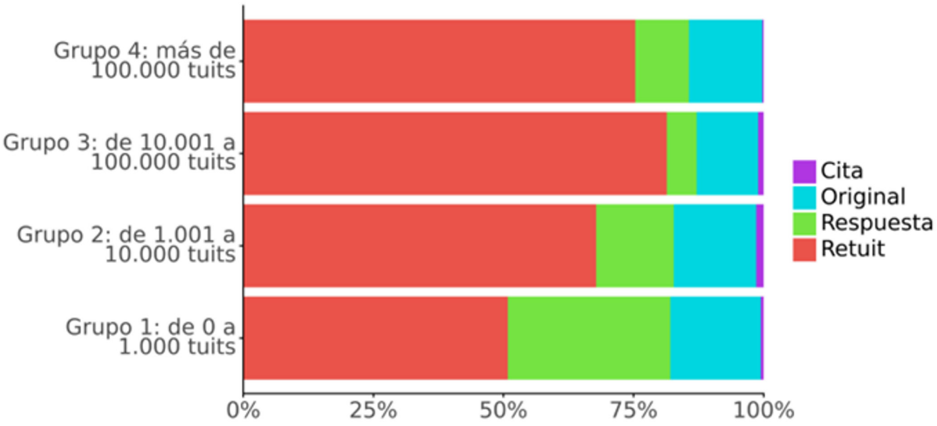
On the other hand, a lot of asymmetries can be seen, since while the verified accounts have a lot of original content (77.52%) and few retweets (18.6%) -and even fewer responses to users-, in the unverified ones there is a majority of retweets (75.47%) and little original content (13.06%). In any case, it should be clarified, given that we are talking about tweets with a high level of entailment (messages that do not contradict the hoax), that the difference regarding units of analysis is logical since it is not usual for a verified account to be the victim of a hoax, although these cases may exist since the verified condition is granted by Twitter not only to accounts of organizations and institutions but also to public figures from any field.

Paradoxically and dangerously, these personalities sometimes do not deny disinformation but are active actors in spreading hoaxes. For example, in Spain, the case of famous singer Miguel Bosé can be cited, whose account was closed by Twitter in August 2020, for spreading false information precisely about the coronavirus.

4.2. According to the number of published tweets and the age of the account

Another aspect to take into account, as shown in chart 3, is the variable related to the number of tweets that the user has published, whose division by groups also presents trends similar to the analyzes of previous charts. In the case of the first group, accounts with less than a thousand published tweets, we could be talking about both very recently created accounts and older ones, although with little activity. Both types of profiles are grouped in that first group and the first type (recently created accounts) can logically include false accounts or accounts carried out by bots.

Figure 3. *Type of tweets (%) by groups with the highest entailment, according to the number of tweets published.*

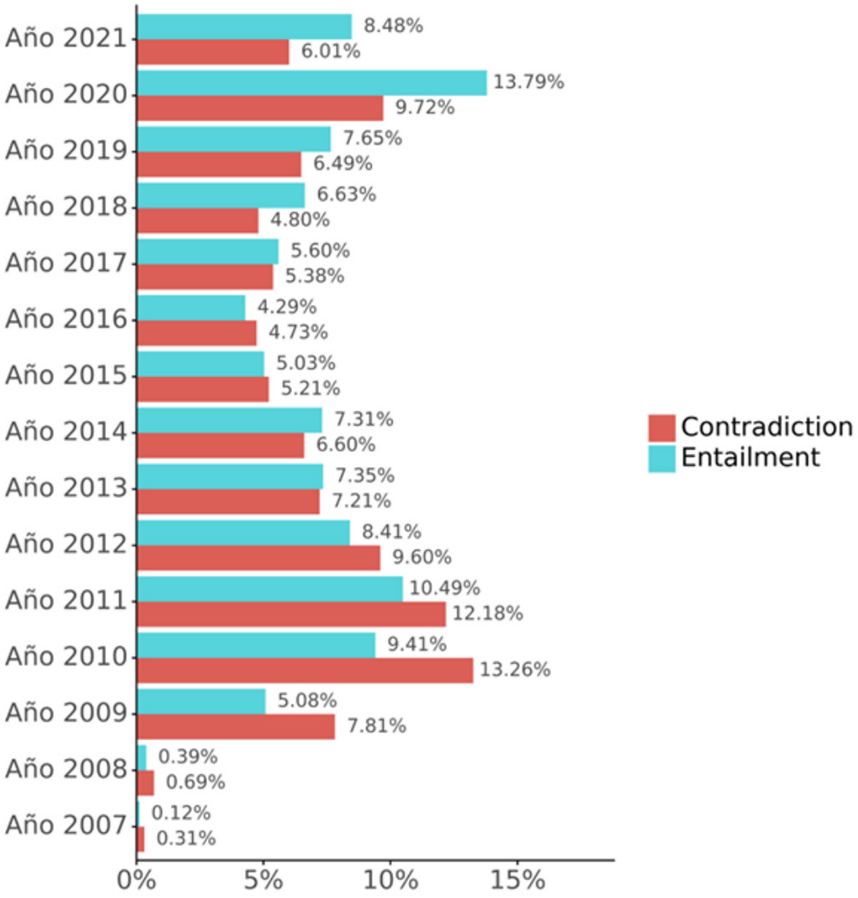


Source: Own elaboration.

Regarding this first group with little published content, it is the type of user that most resorts to responding to other users as a resource for disseminating disinformation, since we found this resource used up to 31.24% of the time. In any case, if the four types of accounts in this chart 3 share something, it is that the retweet is always the majority form of hoax dissemination (from the least to the greatest amount of published content, the retweet appears in 50.84%, 67.82%, 81.38%, and 75.33%, respectively). Regarding the distribution or size of these categories, the smallest is precisely the aforementioned group 1, with a total of 893 examples, while the most abundant group is number 3, with a total of 7,668 units (groups 2 and 4 are more even, with 3944 and 3008 units of analysis, respectively).

For its part, chart 4 allows us to observe an overview of users in favor of or against this type of anti-vaccine hoaxes over the last fifteen years and visually presents significant trends, both in terms of the presence of entailment (pro hoax speeches) as later, in its comparison with the tweets that deny the hoaxes (contradiction).

Figure 4. *Entailment vs Contradiction of the number of tweets, by year of creation of the user.*



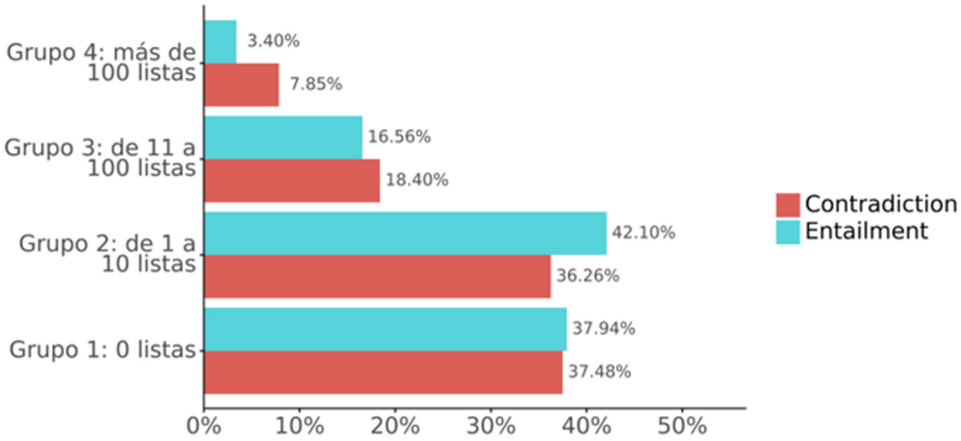
Source: Own elaboration.

If we look only at the blue chart (entailment), the 2020 period coincides with the highest peak in the creation of accounts producing disinformation about anti-vaccine discourse related to COVID-19. Another trend in this regard, started even before the pandemic, is that from 2013 to date (with only two exceptions, 2015 and 2016) we are witnessing a period in which each year the number of accounts created that support hoaxes (entailment) is greater than the number of accounts created that deny them (contradiction). Twitter, a social network that was created in 2006, had six years after its birth (2007-2012) where the number of accounts created against hoaxes was always greater than the number of pro-hoax accounts. However, in 2013 the trend changed to have (except for 2015 and 2016) a predominance of accounts that supported anti-vaccine hoaxes.

4.3. According to the presence in public lists and ratio of followers/followed

As chart 5 analyzes, presence on public Twitter lists is another aspect that can give us clues about the nature of the accounts that participate in the production and dissemination of disinformation about COVID-19 vaccines. The lists are a native tool of Twitter with which the community gives a certain identity to a specific user of this social network. An identity that, on the other hand, does not have to coincide with the self-name of the user in question (a user can be named in a specific way in the Twitter biography, but then be qualified by the community differently by including it in Lists). And here lies the value of lists: as an exogenous element that configures our digital identity.

Figure 5. *«Entailment» vs «Contradiction», according to the presence in public Twitter lists.*



Source: Own elaboration.

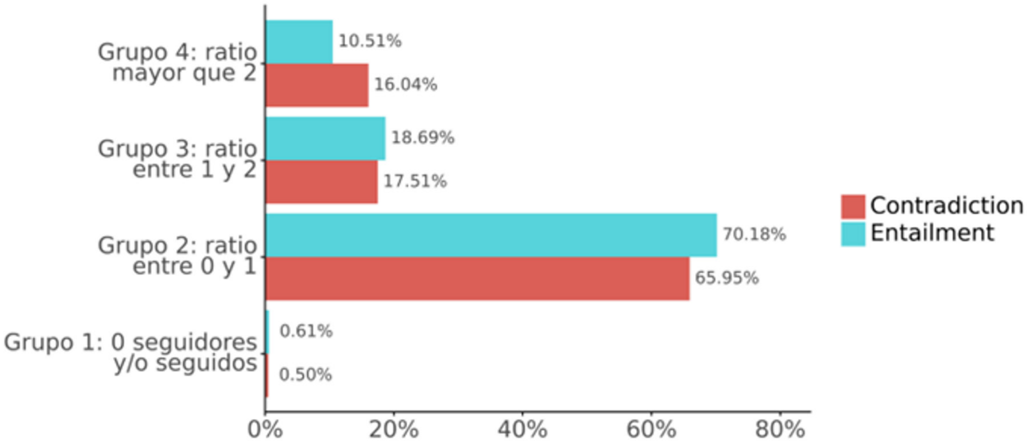
It should be added that being on public lists on Twitter is not something usual for accounts, since very few users receive this type of public recognition from the rest of the community. This would explain the majority of the data in the chart: the majority of accounts, whether entailment (37.94%) or contradiction (37.48%) are not on lists or are on very few (from one to ten lists). In this second assumption, we are talking about 36.26% for contradiction and up to 42.10% for entailment.

The comparison progresses evenly and with a slight advantage towards entailment until the groups with the most presence on the lists are reached. When we talk about users recognized by the Twitter community more than a hundred times through the Lists tool (group 4), the accounts that deny disinformation have an advantage of almost four and a half points (7.85% vs. 3.40%) over those that spread disinformation. They also win, although by less (18.40% compared to 16.56%) in group 3, of the presence in 11-100 lists. A quick reading of these data is that the greater the presence in public Twitter lists, the less likely that we are talking about an account that supports disinformation. However, it would be a first reading of the data because in this chart, as occurs in chart 2 (behavior of verified accounts), what a priori is an element that gives credibility to an account (here, the Lists) is not always that way. This would be due, in this case, to the particular behavior of certain public personalities who, due to their fame, will logically have a presence in a high number of public Twitter lists.

Finally, in chart 6 we have the opportunity to know if there is any special tendency of the accounts that support disinformation in terms of their ratio of followers and followed. The opportunity to know how many followed users they need to have followers. In other words, if it is about accounts that practice or seek follow back (that is, that there is mutual follow) or that, on the contrary, have a genuine and organic recognition by the Twitter community. The higher the ratio, the greater the organic (true) recognition that account has from the rest of the users since it does not need to practice the aforementioned follow-back to accumulate followers.

Received: 26/05/2022. **Accepted:** 29/08/2022. **Published:** 02/01/2023.

Figure 6. *«Entailment» vs «Contradiction», according to the ratio of followers and followed.*



Source: Own elaboration.

5. Discussion and conclusions

The discourse on social networks around COVID-19, with its corresponding hoaxes, is one of the clearest current examples we have of how health misinformation can lead to tangible public health problems. Sometimes it is about crowds in places falsely designated as vaccination centers or, on the contrary, misinformation that weakens the desire to be vaccinated and affects the attendance of certain groups to the real vaccination centers. Between these two extremes of falsehoods, we also find a multitude of variants that give fake side effects to vaccines and that, likewise, influence trust in them and the effectiveness of vaccination campaigns.

This preliminary work has shown that knowing the morphology of the accounts that are the protagonists of disinformation flows can help to better understand the behavior patterns of anti-vaccine discourses and hoaxes. Although there is a hidden level of dissemination of hoaxes about the coronavirus, which occurs on private messaging networks such as WhatsApp (Salaverría et al., 2020), the online public and media discourse of these hoaxes takes place above all on social networks and, in particular, on Twitter. For this reason, the present work has wanted to provide a generic vision that offers clues to identify some behavior patterns of the anti-vaccine discourse in this social network. Further studies, with similar denial speeches, samples, and techniques (the high reproducibility of this article allows it), will show to what extent the trends that have been outlined in this analysis can be extrapolated to other types of hoaxes, health disinformation, or denial discourses.

Regarding the hypotheses raised in this research, it has been confirmed that there are particular behaviors related to the characteristics of Twitter accounts and that they affect both the production and dissemination of disinformation about vaccines for COVID-19 (H1). On the other hand, and regarding the second hypothesis (H2), in the case of the entailment attribute (support for the hoax) there are specific behavior patterns linked to the number of followers of the account, the number of tweets published, and its status as a verified account or not. For its part, in the case of the contradiction attribute, the comparison with its opposite allows us to see other patterns in specific groups of accounts, especially when we group them according to the year of creation of the accounts, their presence on public Twitter lists, or their ratio of followers and followed.

Concerning the first objective (majority form of dissemination of anti-vaccine hoaxes), the creation of original content emerges as the preferred form of disinformation for those accounts with the largest number of followers, while, at the opposite extreme, the predominance of the retweet emerges in those accounts with very few followers and that remind us of the propagandistic role of bots, very possibly framed in this group of accounts.

The second objective highlights the importance of interacting with verified accounts in periods of disinformation, although the results show that this is not the preferred mode of dialogue for this type of account. It is conceivable that verified accounts, aware of having their own voice recognized by the community, opt above all for original content. The verified account condition also applies, as mentioned in this article, to public figures with an obvious role of influence over society. And although the social network has acted in obvious cases of disinformation despite being verified accounts (Twitter's cancellation of the singer Miguel Bosé's account), the truth is that the presence of this type of profile in discourses about anti-vaccine hoaxes, either supporting them (entailment) or denying them (contradiction), is anecdotal.

The search for the third objective (behavior of the account based on the number of published tweets) also refers us, like the first objective, to the debates on state propaganda and the irruption of bots that, framed in the group of accounts with less published content (along with newly created accounts), would opt for replying to other users as the preferred way of disseminating vaccine misinformation. The data has shown that in recent years on Twitter there is a predominance of accounts that support disinformation compared to those that combat it. The current context of polarization, hate speech, and government propaganda, with its consequent creation of bots, could be one of the explanations for this trend detected in recent years.

Regarding the fourth objective, we can trace a clear evolution of the Twitter social network which, although in this case refers to anti-vaccine hoaxes, would indicate that it may contain many similarities if we try to extrapolate it in future studies to other areas. The general trend detected consists of a first stage, just after the birth of this network (2007-2012), where users who currently deny anti-vaccine hoaxes had more weight than those who now spread disinformation on this subject, while, starting in 2013 and especially in 2016, a trend began where the visibility of the hoax (entailment) prevailed over the accounts that denied it (contradiction). It does not seem like a strange discovery if it is put in the context of hate speech and the current polarization in the communicative scenario (Pérez-Escolar and Noguera-Vivo, 2022) and especially on Twitter (Garimella and Weber, 2017; Yardi and Boyd, 2010). A polarization where emotions play a key role (Döveling et al., 2018) and determine the scope of disinformation in the digital field (Serrano-Puche, 2021).

The last two objectives were raised around the analysis of elements of exogenous identity, granted by the recognition of the community and not by the user himself, such as public Twitter lists and the number of followers as an element of credibility, based on the ratio of followers and followed. Social networks, and in particular Twitter, are an example of a mosaic identity (Caro-Castaño, 2015) where the digital identity is configured both by the user and by their community of followers (Kietzmann et al., 2011). Curiously, both variables have offered similar results in the sense that, the greater the recognition of the community, the less likely there are accounts supporting disinformation. Having organic and natural recognition by the community (in the case of the last chart, a high ratio of followers/followed accounts) increases the chances that we are facing an account that refutes misinformation about COVID-19 vaccines.

This work broadens the horizons of research on disinformation in social networks, based on an analysis that relates the morphology of Twitter accounts with certain online behavior patterns of anti-vaccine discourse and the hoaxes it generates.

6. References

Almansa-Martínez, A., Fernández-Torres, M. J. y 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

Aparici, R., García-Marín, D. y Rincón-Manzano, L. (2019). Noticias falsas, bulos y trending topics. Anatomía y estrategias de la desinformación en el conflicto catalán. *El Profesional de la Información*, 28. https://doi.org/10.3145/epi.2019.may.13

Blankenship, E. B., Goff, M. E., Yin, J., Tse, Z. T. H., Fu, K. W., Liang, H. y Fung, I. C. H. (2018). Sentiment, contents, and retweets: a study of two vaccine-related Twitter datasets. *The Permanente Journal*, 22. https://doi.org/10.7812/tpp/17-138

Bodaghi, A. y Oliveira, J. (2022). The theater of fake news spreading, who plays which role? A study on real graphs of spreading on Twitter. *Expert Systems with Applications*, 189, https://doi.org/10.1016/j.eswa.2021.116110

Caro-Castaño, L. (2015). *La identidad mosaico como modo de subjetividad propio de las redes sociales digitales y sus formas de comunicación paramediáticas: La microcelebridad y la marca personal* [tesis doctoral, Universidad de Cádiz]. https://bit.ly/3mJwShQ

Carrasco-Polaino, R., Martín-Cárdaba, M. y Villar-Cirujano, E. (2021). Citizen participation in Twitter: Anti-vaccine controversies in times of COVID-19. *Comunicar*, 69, 21-31. https://doi.org/10.3916/C69-2021-02

Deiner, M. S., Fathy, C., Kim, J., Niemeyer, K., Ramirez, D., Ackley, S. F. y Porco, T. C. (2017). Facebook and Twitter vaccine sentiment in response to measles outbreaks. *Health Informatics Journal*, 25, 1116-1132. https://doi.org/10.1177/1460458217740723

Döveling, K., Harju, A. y Sommer, D. (2018). From mediatized emotion to digital affect cultures: New technologies and global flows of emotion. *Social Media + Society*, 4, 1-11. https://doi.org/10.1177/2056305117743141

El-Mohandes, A., White, T. M., Wyka, K., Rauh, L., Rabin, K., Kimball, S. H., Ratzan, S. C., & Lazarus, J. V. (2021). COVID-19 vaccine acceptance among adults in four major US metropolitan areas and nationwide. *Scientific Reports*, 11, https://doi.org/10.1038/s41598-021-00794-6

Evanega, S. Lynas, M., Adams, J. y Smolenyak, K. (2021). Coronavirus misinformation: quantifying sources and themes in the COVID-19 infodemic. *JMIR*, 10. https://bit.ly/3HoN1RM

Garimella, V. R. K. y Weber, I. (2017). A long-term analysis of polarization on Twitter. En: *Proceedings of the International AAAI Conference on Web and Social Media*, 1. https://bit.ly/3mCmCb5

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology, 78*(6), 1360-1380. https://bit.ly/3sD5AgO

Himelboim, I., Xiao, X., Lee, D. K. L., Wang, M. Y. y Borah, P. (2020). A social networks approach

Received: 26/05/2022. **Accepted:** 29/08/2022. **Published:** 02/01/2023.

to understanding vaccine conversations on Twitter: Network clusters, sentiment, and certainty in HPV social networks. *Health Communication*, 35, 607-615. https://doi.org/10.1080/10410236.2019.1573446

Huertas-García, Á., Huertas-Tato, J., Martín, A. y Camacho, D. (2021a). CIVIC-UPM at CheckThat! 2021: integration of transformers in misinformation detection and topic classification. *Faggioli*, 33. https://bit.ly/3JrhwIi

Huertas-García, Á., Huertas-Tato, J., Martín, A. y Camacho, D. (2021b). *Countering Misinformation Through Semantic-Aware Multilingual Models*. International Conference on Intelligent Data Engineering and Automated Learning. Springer. https://bit.ly/319DiiF

Huertas-Tato, J., Martín, A. y Camacho, D. (2021). SML: a new Semantic Embedding Alignment Transformer for efficient cross-lingual Natural Language Inference. *arXiv preprint*, arXiv:2103.09635. https://bit.ly/3qtsbts

Ireton, Ch. y Posetti, J. (eds). (2018). *Journalism, 'fake news' and disinformation: Handbook for journalism education and training*. Unesco. https://bit.ly/3mGXoZd

Islam, M.S., Sarkar, T., Khan, S., Mostofa Kamal, A.-H., Hasan, S. M. M., Kabir, A., Yeasmin, D., Islam, M. A., Amin Chowdhury, K. I., Anwar, K. S., Chughtai, A. A., & Seale, H. (2020). COVID-19-related infodemic and its impact on public health: a global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, 103. 1621-1629. https://dx.doi.org/10.4269%2Fajtmh.20-0812

Kietzmann, J. H., Hermkens, K., McCarthy, I. P. y Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, 54, 241-251. https://bit.ly/3mGwujY

Kim, H. K., Ahn, J., Atkinson, L. y Kahlor, L.A. (2020). Effects of COVID-19 misinformation on information seeking, avoidance, and processing: a multi-country comparative study. *Science Commun*, 42. https://doi.org/10.1177/1075547020959670

Knuutila, A., Neudert, L.-M. y Howard, P. (2020). Global fears of disinformation: Perceived Internet and Social Media Harms in 142 countries. *COMPROP Data Memo*, 8. https://bit.ly/3FEoCXD

Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M. B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E. W. y Baddour, K. (2020). Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. *Cureus*, 12. https://doi.org/10.7759/cureus.7255

Kummervold, P. E., Martin, S., Dada, S., Kilich, E., Denny, C., Paterson, P. y Larson, H. J. (2021). Categorizing Vaccine Confidence with a Transformer-Based Machine Learning Model: Analysis of Nuances of Vaccine Sentiment in Twitter Discourse. *JMIR Medical Informatics*, 9, https://doi.org/10.2196/29584

Larrondo-Ureta, A., Fernández, S.-P., & Morales-i-Gras, J. (2021). Desinformación, vacunas y Covid-19. Análisis de la infodemia y la conversación digital en Twitter. *Revista Latina de Comunicación Social*, 79, 1-18. https://doi.org/10.4185/RLCS-2021-1504

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K. y Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, 5, 337-348. https://doi.org/10.1038/s41562-021-01056-1

López-Martín, Á., Gómez-Calderón, B. y Córdoba-Cabús, A. (2021). Desinformación y verificación de datos. El caso de los bulos sobre la vacunación contra la Covid-19 en España. *Revista Ibérica de Sistemas e Tecnologias de Informação*, 431-443.

MacCartney, B. (2009). *Natural language inference*. Stanford University. https://bit.ly/3qsAtla

Mantzarlis, A. (2018). Fact-checking 101. 85-100. En: Ireton, Ch. y Posetti, J. (eds). *Journalism, 'fake news' and disinformation: Handbook for journalism education and training*. UNESCO. https://bit.ly/3pAWizZ

Martín, A., Huertas-Tato, J., Huertas-García, Á., Villar-Rodríguez, G. y Camacho, D. (2021). FacTeR-Check: Semi-automated fact-checking through Semantic Similarity and Natural Language Inference. *arXiv preprint, arXiv:2110.14532*. https://bit.ly/32xKfux

Morel, A. (2021). Negationism of the COVID-19 and popular health education: to beyond the necropolitics. *Trabalho, Educação e Saúde*, 19. https://doi.org/10.1590/1981-7746-sol00315

Nowak, S. A., Chen, C., Parker, A. M., Gidengil, C. A. y Matthews, L. J. (2020). Comparing covariation among vaccine hesitancy and broader beliefs within Twitter and survey data. *PloS One*, 15. https://doi.org/10.1371/journal.pone.0239826

Pérez-Escolar, M. y Noguera-Vivo, J. M. (eds.) (2022). *Hate speech and polarization in participatory society*. Routledge. https://bit.ly/3FE8EwB

Saby, D., Philippe, O., Buslón, N., del Valle, J., Puig, O., Salaverría, R. y Rementeria, M. J. (2021) Twitter Analysis of COVID-19 Misinformation in Spain. En: Mohaisen, D. y Jin, R. (eds.) Computational Data and Social Networks. CSoNet 2021. *Lecture Notes in Computer Science*, 13116. https://doi.org/10.1007/978-3-030-91434-9_24

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

Serrano-Puche, J. (2021). Digital desinformation and emotions: exploring the social risks of affective polarization. *International Review of Sociology*, 31, 231-245. https://10.1080/03906701.2021.1947953

Shahi, G., Dirkson, A. y Majchrzak, T. (2021). An exploratory study of COVID-19 misinformation on Twitter. *Online Social Networks and Media*, 22, https://doi.org/10.1016/j.osnem.2020.100104

Subbaraman, N. (2021). This COVID-vaccine designer is tackling vaccine hesitancy-in churches and on Twitter. *Nature*, 377-377. https://doi.org/10.1038/d41586-021-00338-y

Surian, D., Nguyen, D. Q., Kennedy, G., Johnson, M., Coiera, E. y Dunn, A. G. (2016). Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection. *Journal of Medical Internet Research*, 18. https://doi.org/10.2196/jmir.6045

Received: 26/05/2022. **Accepted:** 29/08/2022. **Published:** 02/01/2023.

Thelwall, M., Kousha, K. y Thelwall, S. (2021). COVID-19 vaccine hesitancy on English-language Twitter. *El Profesional de la información*, 30. https://doi.org/10.3145/epi.2021.mar.12

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. y Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998-6008. https://bit.ly/3pFbpYY

Yardi, S. y Boyd, D. (2010). Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of Science, Technology & Society*, 30, 316-327. https://doi.org/10.1177%2F0270467610380011

Zhou, X., Coiera, E., Tsafnat, G., Arachi, D., Ong, M. S. y Dunn, A. G. (2015). Using social connection information to improve opinion mining: Identifying negative sentiment about HPV vaccines on Twitter. MEDINFO. https://bit.ly/3mGQ2ow

Zucker, J., Rosen, J., Iwamoto, M., Arciuolo, R., Langdon-Embry, M., Vora, N., Rakeman, J., Isaac, B., Jean, A., Asfaw, M., Hawkins, S. Merrill, T., Kennelly, M., Maldin-Morgenthau, B., Daskalakis, D. y Barbot, O. (2020). Consequences of Undervaccination-Measles Outbreak, New York City, 2018–2019. *The New England Journal of Medicine*, 382, 1009-1017. https://doi.org/10.1056/NEJMoa1912514

7. Authors

José Manuel Noguera-Vivo
Universidad Católica de Murcia. Spain.

Associate Professor at the Universidad Católica de Murcia (UCAM), where he is a researcher in the area of Journalism and directs the Department of Communication Sciences. At this same institution, he is the Chief Researcher of the "Comunicación, Política e Imagen" research group. With more than half a hundred publications on digital journalism, participation, and social networks, his research focuses on the intersection between media, technology, and society. He has been a postdoctoral fellow at The University of British Columbia (Vancouver, Canada), as well as a guest researcher at several universities and conferences. He is currently responsible in Spain for the Online News Association (ONA). Among his latest publications, it is worth highlighting the book *Hate Speech and Polarization in Participatory Society* (Routledge, 2022).
jmnoguera@ucam.edu

Índice H: 16
Orcid ID: https://orcid.org/0000-0002-7189-7017
Google Scholar: https://scholar.google.com/citations?user=yQGsjxkAAAAJ&hl=en
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=36617884600

María del Mar Grandío-Pérez
Universidad de Murcia. Spain.

Associate Professor at the Universidad de Murcia, she is part of the Spanish Management Committee of the European COST Action INDCOR (Interactive Narrative Design for Complexity Representations, 2019-2023), as she was previously of another COST, Transforming Audiences, Transforming Societies (2010 -2014). She received the Knowledge Transfer Award from the Universidad de Murcia in 2017, she has carried out research stays at the University of Missouri, University of California, University

of Maryland, and Georgetown, among other institutions. Her lines of research focus on entertainment content and audience studies, having scientific publications in journals such as *The International Journal of Audience Research* or *Media Studies*, among others.
mgrandio@um.es

Índice H: 18
Orcid ID: https://orcid.org/0000-0002-2577-4059
Google Scholar: https://scholar.google.com/citations?hl=es&user=eV4aG_wAAAAJ
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=57131060900

Guillermo Villar-Rodríguez
Universidad Politécnica de Madrid. Spain.

Postgraduate researcher associated with the project on disinformation CIVIC (Intelligent Characterization of the Accuracy of Information related to COVID-19 by its acronym in Spanish) and doctoral student in the area of computing at the Universidad Politécnica de Madrid. After obtaining his master's degree in data journalism, he worked at the RTVE LAB and the newspaper El País, where he consolidated his specialization in the area. He received the Extraordinary Award for Career in Journalism and studied an official master's degree in data science and society in the Netherlands to bring computational techniques to journalistic practice. His latest publications analyze automated verification practices through natural language processing and the forms of production and dissemination of hoaxes around COVID-19 on Twitter.
guillermo.villar@upm.es

Índice H: 1
Orcid ID: https://orcid.org/0000-0001-7942-2879
Google Scholar: https://scholar.google.com/citations?hl=es&user=WGpTH9cAAAAJ

Alejandro Martín
Universidad Politécnica de Madrid. Spain.

Assistant Professor at the Universidad Politécnica de Madrid, his main areas of interest are deep learning, modeled language, cybersecurity, and natural language processing. He has been a guest researcher at the University of Kent (UK) and the Universidad de Córdoba. Besides being a lecturer, reviewer, and organizer of numerous international congresses, he is the Chief Researcher of the Intelligent Characterization of the Accuracy of Information related to COVID-19 (CIVIC) project, financed by the Fundación BBVA within the call for research teams on SARS-CoV-2 and COVID-19 (2021-2022). He has published in journals such as *Communication & Society, Information Fusion,* and *Applied Soft Computing,* among others.
alejandro.martin@upm.es

Índice H: 12
Orcid ID: https://orcid.org/0000-0002-0800-7632
Google Scholar: https://scholar.google.es/citations?user=b3J9VRsAAAAJ&hl=es
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=57143182900

David Camacho
Universidad Politécnica de Madrid. Spain.

Full Professor of the Department of Information Systems at the Universidad Politécnica de Madrid and Chief Researcher of the Applied Intelligence and Data Analysis Group (AIDA). His main areas of interest are disinformation, social network analysis, data mining, Machine Learning, or artificial intelligence, in particular the specific area of swarm intelligence. Between articles, books, and conferences, he credits more than 300 publications. He has published in scientific journals such as *The Journal of Supercomputing, Information Fusion,* or *The Journal of Ambient Intelligence and Humanized Computing*. He is part of the Spanish management committee of the European COST Action INDCOR (Interactive Narratives for Complexity Representations).
david.camacho@upm.es

Índice H: 36
Orcid ID: https://orcid.org/0000-0002-5051-3475
Google Scholar: https://scholar.google.com/citations?user=fpf6EDAAAAAJ&hl=en
Scopus ID: https://www.scopus.com/authid/detail.uri?authorId=57191344221

Received: 26/05/2022. **Accepted:** 29/08/2022. **Published:** 02/01/2023.