MALDITA.ES

DISINFORMATION ON WHATSAPP:

MALDITA.ES' CHATBOT AND THE "FREQUENTLY FORWARDED" ATTRIBUTE

June 2021

A report by Maldita.es' academic research unit.

This report is developed by the academic research unit of Maldita.es. The purpose of this unit is to provide studies and research to the editorial staff of Maldita.es so that the work and decisions they make are supported by data and directed by the results of the research. The academic research unit works both independently and in collaboration with universities and research centers.

You can contact us at
academia@maldita.es

INDEX

01

Introduction

01

INTRODUCTION

The Maldita.es database contains the trace that disinformation has left in Spain in recent years. Every piece of data in these files provides specific information that allows us to know how disinformation works, when it appears, how it is shared or what techniques it uses to try to get as far as possible.

Back in July 2020 amid the COVID-19 pandemic, we launched our WhatsApp chatbot and began to collect data on how disinformation was being shared with us in an automated way, enriching our database and standardizing a methodology that until then had been manual. Since March 2021, a new element has been included in that database: messages marked with the Frequently Forwarded (FF) attribute, which indicates those that have already been forwarded on WhatsApp five or more times. This attribute is also known as 'Highly Forwarded Messages' (HFMs).

The academic research team at Maldita.es has made a first analysis of the meaning and relevance of this new attribute. After studying the patterns of appearance of the FF during the first month in which it was active, we have found indications that it is strongly linked to disinformation. 74% of the alerts with FF we received were linked to an investigation by the Maldita.es team and 78.72% of the investigated alerts were finally rated as a hoax or a disinformation.

In addition, in the patterns of appearance of the alerts with FF we observe that they tend to be included in hoaxes with a very high potential impact, either because they are likely to become very viral or because they are old hoaxes that already had a significant impact in the past and are back in circulation. There are also features that suggest that these types of alerts can signal the appearance of coordinated disinformation campaigns. This data open the door to the development of an early warning system that allows fact-checkers to react earlier to this type of disinformation. These are some of the promising lines of research that are opened after the analysis of only one month of operation of the FF attribute. A broader and more in-depth study can confirm these hypotheses and broaden the knowledge about the possibilities that FF alerts provide to facilitate the work of fact-checkers.

This report analyzes the importance and usefulness of the WhatsApp service for verification entities such as Maldita.es, both as a means of detecting viral disinformation in private conversations through warnings from the community and as a vehicle to convey verified information to citizens. It also evaluates the impact and progression of the WhatsApp chatbot developed by Maldita.es in July 2020 and the progress made by automation compared to the old manual WhatsApp Business system. In addition, the usefulness of the "Frequently Forwarded" attribute associated with the content for verifiers is explored through case studies. Maldita.es is the first fact-checker in the world to investigate its usefulness for the fact-checking community.

Our chatbot has been awarded the European Press Prize on the innovation category in 2021.

02

The use of WhatsApp in Spain: information and disinformation

02

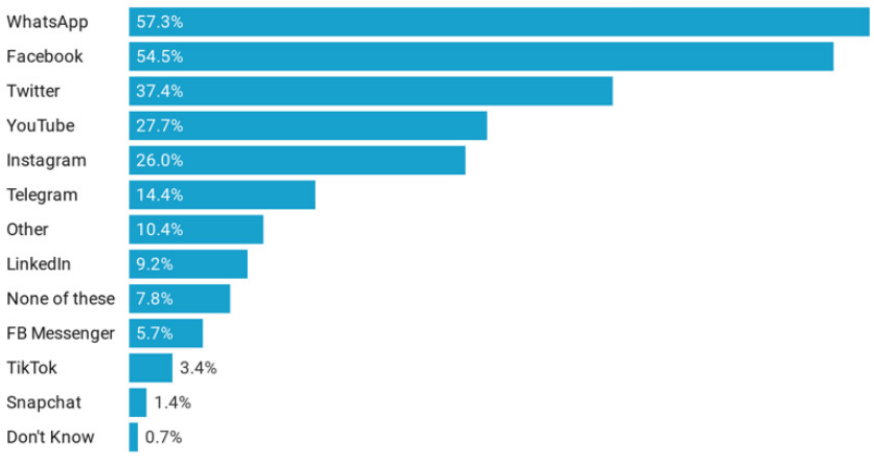
THE USE OF WHATSAPP IN SPAIN: INFORMATION AND DISINFORMATION

WhatsApp is the mirror of what is being talked about everyday in Spain. Statistics indicate that its use is massive: 85% of Spaniards use the messaging application on a daily basis according to the IAB Spain Annual Study of Social Networks. In addition, the use of WhatsApp is transversal. According to the 2020 Digital News Report of the Reuters Institute, the penetration of the application is enormous in people of all age groups, all economic situations and educational levels. In all these groups, a percentage of daily use is higher than 68%. But it is also very relevant what Spanish society uses it for. WhatsApp doesn't only let you bee in permanent contact with family, friends and acquaintances, it has also become an essential way to receive information. The same Digital News Report indicates that 34% of Spaniards use WhatsApp for information in 2020, which makes Spain the European country in which this application is most used to receive news from all those analyzed in the report.

It has also been a relevant information channel during the COVID-19 pandemic. The percentage of Spaniards who have searched for information in the application reached 57% in May 2020, according to this study by the International University of Catalonia (2020). In fact, another study by the Miguel Hernández University (2020) points out that, during the first months of the pandemic, WhatsApp became the most used platform for all age groups to share information about the coronavirus.

34% OF SPANIARDS USE WHATSAPP TO GET INFORMATION

Figure 2. Use of social media and mobile apps for news



Use of WhatsApp to receive information during the pandemic. Source: UIC

One year after the beginning of the pandemic, WhatsApp continues to be a source of information about COVID-19 for 26% of the Spanish population, according to a report by the Reuters Institute of Oxford (2021). It is the highest figure among the European countries that appear in this study. This same investigation detects that there is a problem with disinformation in messaging applications. 29% of Spaniards acknowledge that they have recently received a lot of disinformation about COVID-19 through messaging services. That percentage is, once again, the highest among European countries. According to this report, these applications are the type of digital platform that generates the least confidence when it comes to receiving information about the pandemic in Spain. Only 16% trust them, five points less than social networks. It is important to note that this research discovers a relationship between people who trust the content that they receive through messaging services and belief in hoaxes. The more a person trusts what they receive from these apps, the more likely it is that disinformation about the coronavirus vaccine will be believed. Spain is the third country after Germany and the United States in which this correlation is strongest.

We can conclude that WhatsApp is a fundamental pillar of communication and information in Spain, but it is very sensitive to the impact of disinformation and therefore a space that can become dangerous. Furthermore, as the messages are end-to-end encrypted, it is very difficult for verification institutions and organizations to locate and debunk hoaxes that are disseminated on this platform.

2.1 Verification in WhatsApp and the importance of the community

When we launched our WhatsApp Business service in July 2018, we did so being aware that a large part of the disinformation affecting Spanish citizens was shared on this conversational platform. Already then, the consumption of information through WhatsApp affected, according to the Digital News Report of that year, 36% of the Spanish population and our first number of WhatsApp helped us understand the phenomenon of disinformation and design what would be our strategy to fight it in the years to come.

The first step to fight disinformation is to detect it as soon as possible and to do so you need to have access to the place where it proliferates. On end-to-end encrypted platforms, such as WhatsApp, we can only know what is going viral in conversations if the users themselves let us know. That is why the strategy of Maldita.es in recent years has been to build a community that today has more than 48,000 registered users who collaborate with our newsroom when it comes to detect viral content on social networks, but also in their private

conversations. People committed to fight disinformation who notify us when they see suspicious content on WhatsApp, Telegram and other platforms.

48.000 REGISTERED USERS WHO COLLABORATE WITH OUR NEWSROOM WHEN IT COMES TO DETECT VIRAL CONTENTS

It's a two-way street because the same people who help us identify what's going viral on those platforms are also the ones who share verified content in their private conversations and groups. The places where Maldita.es cannot reach directly but where verified information is most necessary. Our job is simply not possible without our community.

For almost two years we used a WhatsApp Business account that, beyond identifying you as a legal entity, did not have great differences with the WhatsApp of any user: a number that could only be accessed by one Maldita.es journalist at a time who had to reply to inquiries one by one. In an account that received such inquiries from 200 to 300 users every single day, that meant dedicating a full-time journalist to just answer those messages. It was a repetitive and tough task because it required constantly exposing a staff member to hoaxes that came our way, very often linked to hate speech.

For almost two years we used this system and carried out a manual count of the number of times disinformation reached us in order to measure its virality. Many times our WhatsApp was saturated, unable to manage the flood of queries about the veracity of content related to an electoral campaign or an attack in which complaints could reach peaks of up to 800 users in one day. This meant not only that we could not access the viral content that users sent us, but also that those users were left without an answer to their queries.

03

Maldita.es' chatbot: a collective design

03

MALDITA.ES' CHATBOT: A COLLECTIVE DESIGN

Managing all the inquiries we received via WhatsApp became an impossible task during the COVID-19 crisis: 200 to 300 users reached out daily before the pandemic, but as coronavirus spread around the world we received inquiries from 2,000 people every day. Due to the high volume of messages, our WhatsApp service was failing to give us insights into what contents were going viral, nor was it providing contrasted information to those who were using it. That's when we found out about COVIDWarriors and the Wealize team, so we started working on the creation of an automated chatbot that would be able to receive, store and identify the contents that were being sent to us. The goal was to automatically respond when the inquiries had already been resolved by Maldita.es.

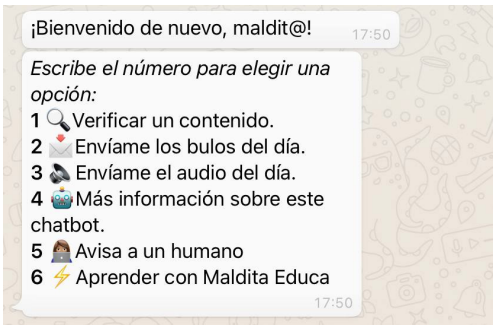
In July 2020 we launched our WhatsApp chatbot to automate part of that process and to immediately reply to those who were contacting us.

After nearly a year of functioning, over 26,000 people have used our automated service, of whom 61% have done so more than once. Our chatbot has sent over 400,000 messages, verified 108,000 contents and issued 143,800 daily debunks summaries in its text format and 18,400 in audio format.

26,000 PEOPLE HAVE USED OUR AUTOMATED SERVICE, OF WHOM OVER HALF OF THEM HAVE DONE SO MORE THAN ONCE

Just in 2021, 16,200 users have contacted us, with an average of over 950 daily conversations among users who sought to verify contents and those who requested other options in our chatbot.

Since we launched this automated service, when someone sends a message to the chatbot they receive a menu with several options. They include verifying contents, receiving a daily summary of Maldita.es' debunks or learning lessons with Maldita Educa, among others. Through a numeric menu users can choose from 6 different options:



Chatbot's main menu

QR to access the chatbot

1. Fact-check a piece of content

If a user has received a piece of content that he or she wants to fact-check, when choosing option '1', the chatbot will ask him to send the photo, video, audio or WhatsApp chain in question.

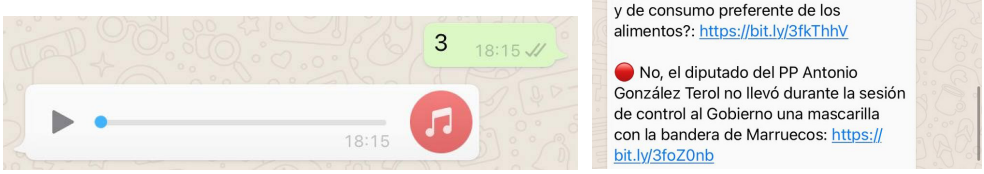
In the event that you send a piece of content that Maldita.es has already fact-checked, the user automatically receives the related article that debunks it. Otherwise, an editor will review the message sent by the user to verify it.

While the user receives an automated answer with Maldita.es' debunk; this allows the newsroom to know which hoaxes are being shared and how viral they have become.

Messages answering user's questions

2 & 3. Send me the daily fact-checking summary or audio

In order to make it easier for users to access Maldita.es 'articles, the chatbot includes the option to request the most important articles published that day. There is also an audio in which a journalist narrates the headlines of those articles.



The chatbot offers a daily fact-checking summary and audio

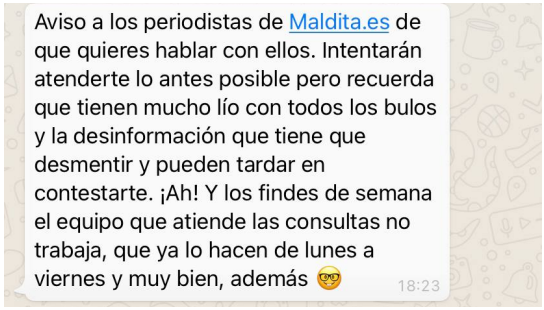
4. More information about this chatbot

Since not everyone is familiar with the concept of a 'chatbot', we believe it is important to offer an option to explain what it is and what it does. When users press '4', they will receive a message stating who developed the chatbot, what this chatbot is capable of doing and its privacy policy.

The chatbot includes an 'about me' option

5. Call a human

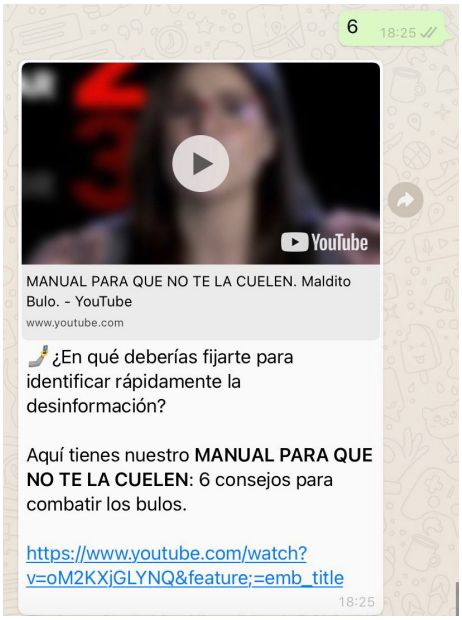
It is possible that the user is not satisfied with the answer to his or her query or simply that he or she prefers to speak with a human. By clicking on this option, the chatbot collects the query for a Maldita.es' journalist to review it.

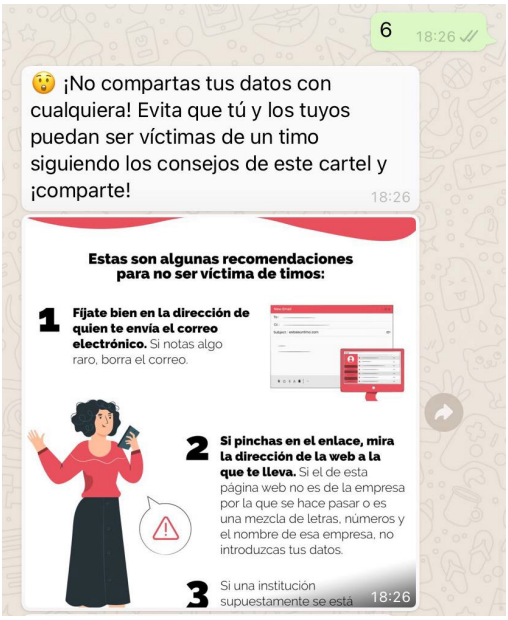


If the chatbot doesn't have an answer, it will alert a human

6. Learn about media literacy

Media literacy and giving society the necessary tools to differentiate true from false are part of Maldita.es' mission, therefore, we've built our chatbot so that there is an option that offers users resources and explanations around disinformation, fact-checking and media literacy courses.





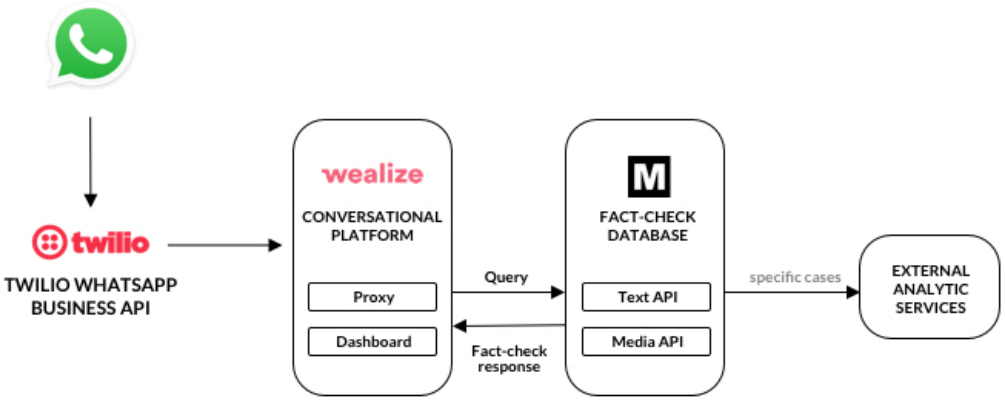
The chatbot also gives access to media literacy content

The chatbot's architecture

The chatbot is the main way in which our community alerts us of possible hoaxes on WhatsApp and it has become an essential element of our daily work. Not only does it process almost a thousand notifications on average every day, it also allows us to have a vision of what is going viral in closed WhatsApp conversations and groups, which we could not reach otherwise if it weren't for the reports from our users due to the end-to-end encryption WhatsApp alleges to have.

The architecture comprehends four different levels:

- The connection to the WhatsApp API via Twilio.
- The conversational platform Coloq.io.
- Maldita.es's database and its management frontend.
- External services used to process some of the disinformation formats like audios or images.



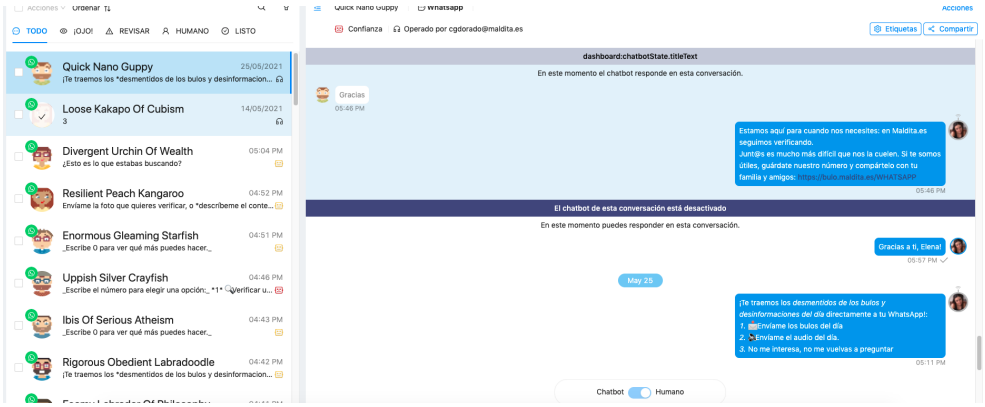
The chatbot's architecture

3.1 The conversational platform Coloqu.io

Maldita.es' WhatsApp chatbot is built on Coloq.io, a conversational platform that allows to connect with users in different conversation channels such as WhatsApp through Twilio, but also webchat, twitter, FB messenger and SMS, Google Business Messages; understand and respond to messages using conversational artificial intelligence tools, such as IBM Watson Assistant, Google Dialogflow or NLTK; consult information sources and integrate with work processes; it enables the process to be supervised by people with rules and personalized alerts; and get enough data to be able to generate reports like this one. Colloq.io is a product of Wealize, a digital products company specializing in blockchain and conversational artificial intelligence.

When the user sends a message through WhatsApp, it is received by the WhatsApp Business API in Twilio and it is processed with conversational artificial intelligence, holding a conversation to check news, answer questions or offer more information.

For each conversation, a series of rules and alerts are established that allow the verified information to be offered to users. In addition, Coloq.io allows the conversation to be intervened by a human improving the user experience, as well as sending individual messages or to several users, as necessary.



Coloq.io's frontend management

3.2 Maldita.es' database

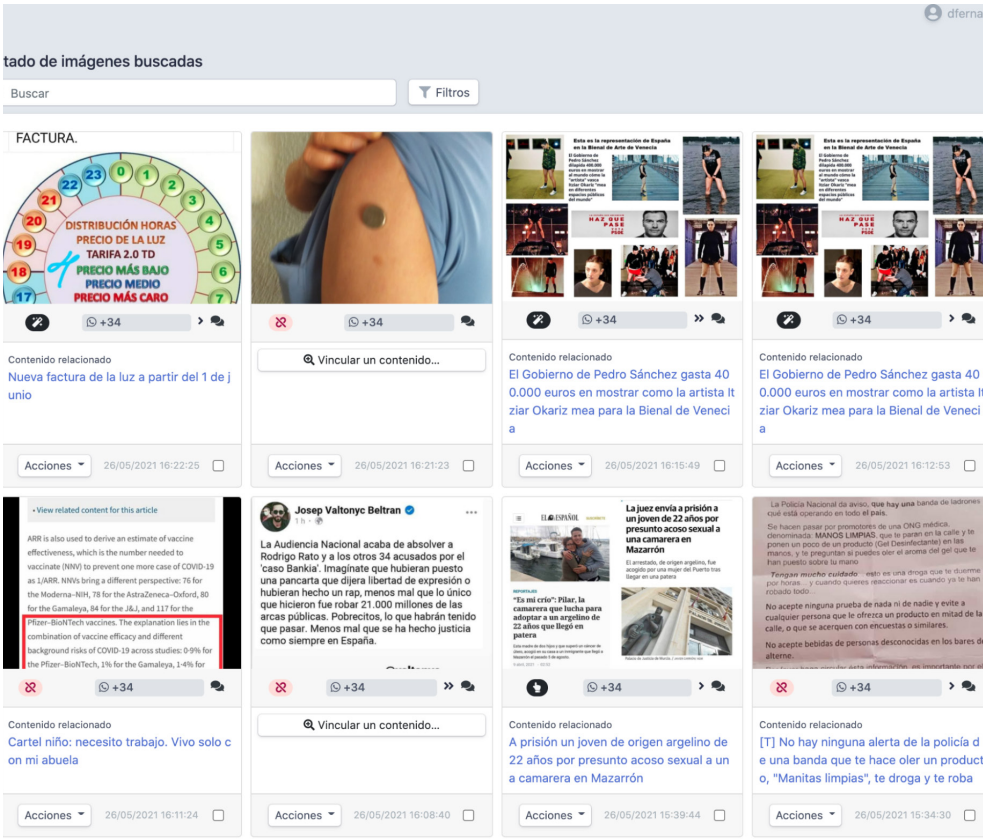
Maldita.es' database has three differentiated pillars: the API, the search system and the management frontend.

The inquiries that are received through the API are handled differently depending on their format:

- Images, videos and audios are analyzed internally and with external services to extract possible coincidences from the database or any other identifying element of text that will enable completing searches in the database.

- Text contents are directly searched for within the search system.

The results of that operation, along with the search itself, are registered in the database and can be consulted through the management frontend.



Maldita.es' management frontend

With this frontend, we can complete searches in different formats (text, video, audio or image), which enables us to access data such as the transcription of an audio or an image, the interpretation of the language processing system of the bot, the conversation on the Coloq.io platform, the number of times that we have received a piece of content or the different formats in which it has been sent to us. Also, it's used to register new contents that are likely disinformation so we can monitor them and, from there on, identical contents will be added to measure the virality of them.

Once the content is linked with a debunk the chatbot automatically sends the article to all users who have previously inquired about such content. Those who ask about the content after we have linked it to a debunk also receive the article automatically.

In addition, the recollection of the frequency with which we receive the notices gives us important clues on how a hoax evolves (when it's created, the diffusion intensity, the moment it starts to decrease, etc.). Every new data that we receive about a hoax gives us the chance to know it better and give it a more effective response.

To properly understand the importance of the alert system through our WhatsApp chatbot it's important to be familiar with Maldita.es' work process. The tool's goal is to resolve the user's doubts as soon as possible in order to let them know if the content they are suspicious of is a hoax or not, but the goal is to also let us know what is becoming viral at that particular time. We don't have access to user's private conversations as, according to WhatsApp, they are encrypted from end-to-end, so we need the users to be the ones who send these contents to us.

1. We first analyze every incoming recurrence. We receive hundreds of them every day and we have automated the process, although there's still part of the process that is done manually.

2. If we find that the notice is related to a hoax that we have already identified and debunked, the chatbot automatically responds with our verification.

3. If we haven't debunked it or identified it, a content access is created to measure its virality which, along with how dangerous it is, is the main variable we consider when deciding if we debunk it or not. We have limited resources, so we focus our efforts on the hoaxes that are either viral or that can present a real danger to citizens if they continue to be shared.

04

Analysis: How can the "Frequently Forwarded" attribute help fact-checkers

04

ANALYSIS: HOW CAN THE "FREQUENTLY FORWARDED" ATTRIBUTE HELP FACT-CHECKERS

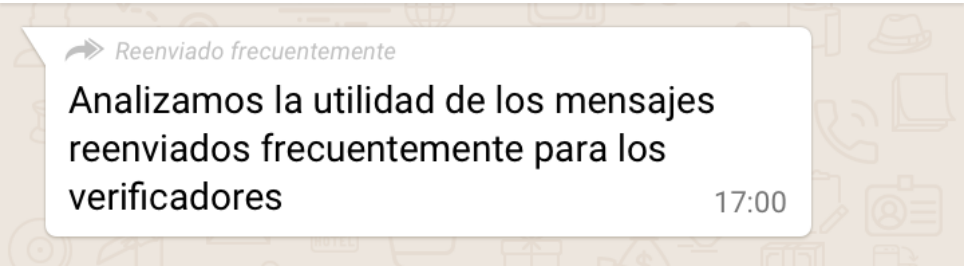
From March 2021, the messages that were sent to our WhatsApp chatbot contained additional information.

Depending on the number of times they had been sent they included a different attribute facilitated by the WhatsApp API. From then, our database registers three types of messages.

- SENT: Messages that have only been sent once and have no special attribute.

- FORWARDED: Messages that have been sent more than once but less than five times. They are marked with an arrow.

- FREQUENTLY FORWARDED (FF): Also know as 'Highly Forwarded Messages' (HFMs), these are messages that have been sent five times or more. They are marked with a double arrow. According to WhatsApp, these messages can no longer be sent to various groups at the same time, but they can still be sent individually to other recipients.



Although all organizations that have access to WhatsApp's API have access to the FF attribute, Maldita.es is the first fact-checking organization to investigate its utility for the fact-checking community.

This attribute grants very relevant information to fact-checkers regarding the virality of the content, which helps us improve our work when choosing what contents we verify following objective criteria.

MALDITA.ES IS THE FIRST FACT-CHECKING ORGANIZATION TO INVESTIGATE THE UTILITY OF THE FF ATTRIBUTE FOR THE FACT-CHECKING COMMUNITY

We have analyzed a full month of our chatbot's activity, from the 5th of March at 00:00 to the 4th of April at 23:59, and the data highlights how important it is for Maldita.es' work. During that month we wrote and published 207 articles related to hoaxes and contents that needed context: in 163 of them, the notices had been sent to our chatbot. This means that a notice had been sent to our WhatsApp chatbot in 78.74% of our debunks.

THE WHATSAPP CHATBOT ALERTED OF THE VIRALITY OF 78,74% OF HOAXES WE DEBUNKED

In that period of time, our chatbot has received 8,014 notices, of which 1,162 were tagged with the FF attribute. That's 14.5% of the total. It's important to point out that our chatbot is unable to receive text messages with more than 1,600 characters due to Twilio's limitations.

We have not investigated all of these 8,104 notices as the resources at Maldita.es are limited. Also, among text and audio notices we receive lots of irrelevant user input that has not to be counted. The ones that look initially false are selected for further investigation, like the notices that are highly dangerous or have been sent to us significantly and are about a disputed topic. As we have previously explained in the process description followed by Maldita.es, we apply an automatic and manual filter in order to find meaning in the messages that are sent to us and link them to the investigations we do to confirm if we are before a hoax.

We monitored 2,089 of potentially false notices to check their virality, equalling 26% of the total. If we exclusively focus on the contents associated with the FF attribute, of the 1,162 contents that were sent to the chatbot during that month we tracked 860 (74%). Therefore, it's three times more likely we monitor content that is associated with this attribute, following our own journalistic and methodological criteria.

Do all contents that are monitored by Maldita.es end up being catalogued as a hoax or disinformation in an explicative article? No. Firstly, we have limited resources. We've also got to bear in mind that some of the things we monitor can't be independently verified or can be real.

In our analysis, we have found that 78.72% of contents that are associated with the FF attribute and are monitored by our team (alerts) are finally tagged as hoaxes or disinformations.

78% OF ALERTS WITH FF END UP BEING RATED AS HOAXES OR DISINFORMATIONS

Our analysis has let us certify that receiving a normal notice is not the same as receiving one that is associated to the FF attribute. The latter gives us a more precise idea of the level of danger and virality each of these hoaxes present. Particularly, we have seen that FF notices help us spot two type of phenomenons:

- They are useful to quickly spot hoaxes that will likely be very viral.
- They help us detect what we call 'zombie hoaxes', those that are periodically shared and repeatedly appear in our alert system.

Additionally, we have seen that FF tags don't just offer clues on how particular hoaxes evolve. The analysis also presents interesting conclusions regarding the disinformation topics that need to be investigated in the future. We have found that there are sensitive matters in which a large amount of FF alerts concentrate, like large political controversies or fraud-related topics. We have also found a link between FF messages and organized campaigns that criminalize vulnerable groups such as immigrants.

We will now share a detailed analysis of the FF identified by our chatbot during the first month and explore, with case studies, the possible utilities of FF for fact-checkers.

1. FF and formats: data and analysis
2. FF and disinformation categorization
3. FF: indicative of high virality rate
4. FF and 'zombie hoaxes': the announced resurrection
5. FF in case studies

4.1 FF AND FORMATS: DATA AND ANALYSIS

Our chatbot is able to receive, classify and identify alerts in four different formats: text, video, image and audio. During the analysis of the notices received during the studied period, we have found substantive differences among the four formats when applied to frequently forwarded messages. We have also seen that on WhatsApp, normally, each hoax tends to be shared in one format only. That is why separately analysing the data from the four different formats gives us precious insights into how disinformation is shared on WhatsApp.

Format	Total content	FF%
Text	5,532	9.9%
Image	1,651	16.5%
Video	736	42.4%
Audio	95	29.4%

Distribution of alerts with FF attribute by formats

The first big difference we find is the percentage of notices with a 'frequently forwarded' label that we get in each of the formats:

- In **text**, we receive lots of inputs but less than 10% is marked as FF. That is due to irrelevant records that are filed in the database during the user's interaction with the chatbot.

- In **image** the percentage is low as well, but in this case we do not attribute it to the irrelevant user inputs. Our hypothesis to explain why the percentage of FF in images is so low when compared to video or audio would be that the option to automatically download images on to the phone is often activated by default, meaning it's easier to share them unlike videos and audios, which may be configured to not be downloaded automatically in order to save space.

When we only look at the data from contents that we have been watching over time, we can also extract some interesting conclusions. We can see that in text, the percentage of notices that end up linked (matched) is very low, 12%, and 36% for audio, substantially lower than the figures for video and image. This happens because in text and audio the chatbot receives a great deal of irrelevant user inputs that are not acted upon: answers to the chatbot's messages, greetings, badly-formulated or even incomprehensible questions, personal messages...

Format	Total notices	Alerts matched	% alerts
Text	5,532	666	12%
Image	1.651	971	58.81%
Video	736	417	56.65%
Audio	95	35	36.84%
TOTAL	8,104	2,089	26%

Percentage of alerts linked by format

4.2 FF AND DISINFORMATION CATEGORIZATION

When we focus solely on the messages we've tracked that also have the FF attribute, the data shows that notices are much more significant: 74% of FF associated notices, regardless of the format, are matched with content that is being monitored by our team as there is a strong belief that it's false. In audio and text messages the percentage of inputs tagged with FF is very high: text ones are close to 85% and audio reaches 75%. Additionally, if we compare these percentages with that of the notices we have followed up that don't have the FF attribute we can see that the difference is very large in all formats. It's especially important in text and audio formats, but even in videos it doesn't reach 50%.

Format	Matched alerts with FF atribute	Matched alerts without FF attribute
Text	84.54%	4.03%
Video	67.62%	48.58%
Image	63.60%	17.69%
Audio	75%	2.,89%
TOTAL	74.01%	9.7%

Matched alerts with FF vs Matched alerts without FF

The FF attribute is a solid clue to flag potential disinformation content. According to the data analyzed this month, there is a big difference between the percentage of alerts that we analyze as possible hoaxes when they carry FF and when they do not. When they have FF, 74.01% appear linked to some content that we are analyzing as possible disinformation. When they do not carry it, that percentage falls to 9.7%. Furthermore, this difference occurs in all formats. Even in video, which is the format with the highest percentage of alerts without linked FF, there is a difference of almost 20 points between carrying FF and not carrying it. All of these data support our theory that the FF attribute is a robust indicator that the content they flag likely contains disinformation.

There is another piece of data that highlights a strong link between the FF and disinformation. Not all the content that we investigate ends up being rated as a hoax. After verifying them we find that some are true, others are irrefutable and others, following the Maldita.es methodology, we consider that we do not have enough evidence to flag them as disinformation. According to the data we collected during the month we analyzed, there are 860 alerts linked to some of these investigations and a further 677 (78.72%), were related to content that Maldita.es rated as a hoax or a disinformation.

Format	Alerts with FF which was disinformation
Text	62.3%
Video	22.2%
Image	13%
Audio	2.5%
TOTAL	**100%**

Percentage of FF alerts that pointed to a hoax

The latter also holds important differences when we compare different formats. Of the 677 FF notices that warned of a content that was later rated as a hoax by Maldita.es, 62.3% were text messages. The second most common format is videos with 22.2%, followed by images with 13%. Audios represented 2.5% of the total.

4. 3. FF: INDICATIVE OF HIGH VIRALITY RATE

In this month's analysis we have seen how the FF alerts clearly point at great sources of disinformation. The contents with more alerts in our database during this time frame match the contents associated with the FF attribute; that is, the FF clearly marks the most viral content.

THE FF ATTRIBUTE MARKS THE MOST VIRAL CONTENT IN 80% OF THE CASES

To verify this statement based on the data, we have analyzed two types of content:

- The Top 10 hoaxes with the most alerts throughout the analyzed period: the 10 hoaxes or pieces of disinformation that have been sent to our chatbot the most during that month.
- The Top 10 disinformation contents with the most alerts associated with the FF attribute in the analyzed period.

8 contents appear in both rankings. This means that the hoaxes that were sent the most to our chatbot match the ones that were tagged with the FF more often. Hence, the FF attribute marks 80% of the contents that are most viral.

Alerts with FF focus very intensely on the 10 hoaxes with the most recurrences this month. A third of all the matches marked with FF during the month appear in these ten hoaxes and disinformation contents. Specifically, 387 of the 1162 that were sent to us this month are concentrated in the top 10.

The volume of alerts with FF in these hoaxes is also very high. In total, these 10 hoaxes add up to 803 notices, and half of them are disinformation. That percentage more than triples the 14.5% on average that we observed in the total database during this month.

In fact, the Top 10 list of hoaxes that were sent to us the most during that month is very similar to the Top 10 list of hoaxes with the most FF alerts. Eight of the hoaxes appear in both. The two that receive a lot of FF (but not so many alerts) are in text format, which as previously stated favours the appearance of a high volume of notices with the FF attribute.

The data shows that alerts with FF appear in greater numbers and in a much higher percentage in the main hoaxes of the month, those that have been more viral and, therefore, carry a greater danger.

Top 10 of the hoaxes with the most alerts

Hoax	Alerts	Alerts FF	% Alerts FF	Format
Chillán volcano in Chile	188	119	62.29%	Text
Pablo Iglesias' dismissal	119	29	24.36%	Image
Danish Parliament about Cataluña	111	54	48.64%	Video
Begoña Gómez excluded from the University	101	73	72.27%	Image and text
Trojan hoax Can't believe it's you	54	4	7.4%	Image
Danish Parliamente about monarchy	51	16	31.37%	Video
WhatApp's new rule	51	49	96.07%	Text
Shell company	45	22	48.88%	Image
Call from 626634795	43	32	74.41%	Image and text
30% of vaccinated will die	40	4	10%	Image
TOTAL	**803**	**402**	**50%**	

Top 10 hoaxes with more alerts associated with the FF attribute

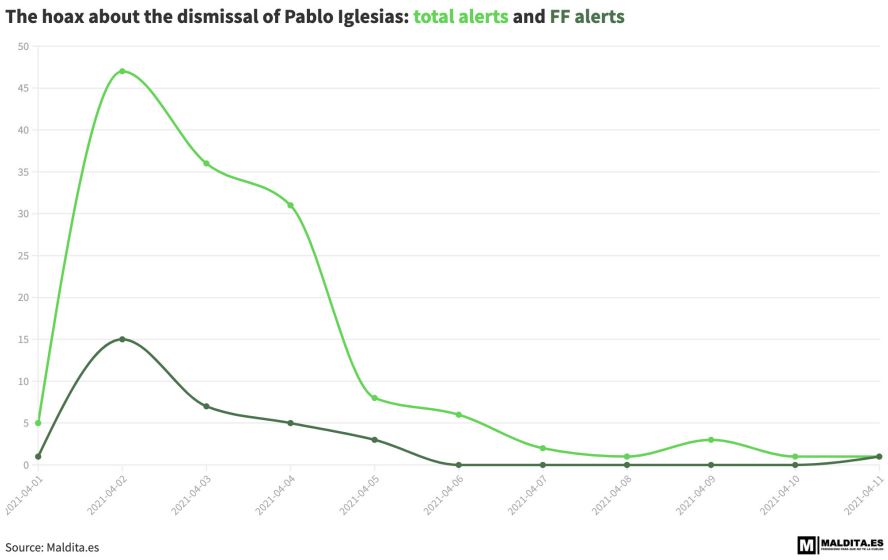
Hoax	Alerts	Alerts FF	% Alerts FF	Format
Chillán volcano in Chile	188	119	62.29%	Text
Begoña Gómez excluded from the University	101	73	72.27%	Image and text
Danish Parliament about Cataluña	111	54	48.64%	Video
WhatApp's new rule	51	49	96.07%	Text
Call from 626634795	43	32	74.41%	Image and text
Pablo Iglesias' dismissal	119	29	24.36%	Image
Cuts in pensions	26	26	100%	Text
Shell Company	45	22	48.88%	Image
Fired from Príncipe Felipe	34	19	55.88%	Text
Danish Parliamente about monarchy	51	16	31.37%	Video
TOTAL	**769**	**439**	**57.08%**	

The hoax about the dismissal of Vice President Pablo Iglesias

This hoax is one of the case studies that exemplify how the FF attribute precedes a great viralization. This hoax was shared as an image and, among all recent disinformations, is the one that carried the most alerts during the analyzed period. In just 10 days we received 146 notices in total. In the message they falsely accuse the former Vice President of the Spanish Government of doing a legal trick to receive a very high compensation when leaving office, according to the disinformation. To "prove it" they attach a publication of the Official State Gazette of March 31, in which the vice president officially resigns, just 15 hours before the first alert reaches our chatbot.

Image of the hoax on the dismissal of Vice President Pablo Iglesias

The first alert arrived on April 1 at 3:36 p.m. and that same night we already received the first warning with FF, which is repeated first thing in the morning of the 2nd. Shortly after, two FF were registered in a row, which gave us an idea that it is a potentially viral hoax. On April 2, we were alerted 47 times that this hoax was being shared on WhatsApp. In the afternoon of that day the volume of inputs with FF grew. Six out of the seven alerts we received between 17:01 and 21:13 included FF. The diffusion during the next two days is still very intense and begins to decline as of April 5.



The hoax about the dismissal of Pablo Iglesias. Total alerts vs FF alerts

It must be taken into account that it is a hoax in image format, in which we detect a lower percentage of notices with FF, probably because in the process of resending the images it causes some of those marks to be lost. Even so, we see a very considerable amount of FF that, in addition, marks in a very clear way that it was a hoax that was achieving a high degree of virality. The concentration of notices with FF in the afternoon of April 2 is a clear sign that it is a very widespread hoax and it is important to react quickly against it.

4. 4. FF AND ZOMBIE HOAXES: THE ANNOUNCED RESURRECTION

Just like a zombie is a resurrected being, there are hoaxes that reappear from time to time. They are disinformations that are shared every few months despite having been debunked in the past

Based on our analysis we have detected that zombie hoaxes often return associated with the Frequently Forwarded attribute, and this is useful for verifiers since those contents that recur in time are highly likely to continue to do so. That's the reason why it's convenient to debunk them.

These types of hoaxes have been in our database for years but they never fully disappear as they are shared from time to time. In this study we have looked at those we detected months ago for the first time. They accumulated more than 100 recurrences in our database and reappeared through our alert system this month.

In total, there are 14 of those big zombie hoaxes that meet these three conditions. Between all there we added up 503 recurrences. 298 of them, almost 60%, are marked with FF. This percentage indicates that old contents that have been shared on WhatsApp for a long time, tend to be forwarded and tagged with the FF attribute very often. In the chart we see that when there is a high number of alerts, that is, when the hoax is being heavily shared again, the percentage of alerts with FF tends to be very high. At the bottom of the list there are other hoaxes that have been widely shared in the past but have not reappeared in the analyzed month. They remain in a latent state. Further studies about the behavior of zombie hoaxes over time could confirm whether there is a direct relationship between FF alerts and the reappearance of zombie hoaxes.

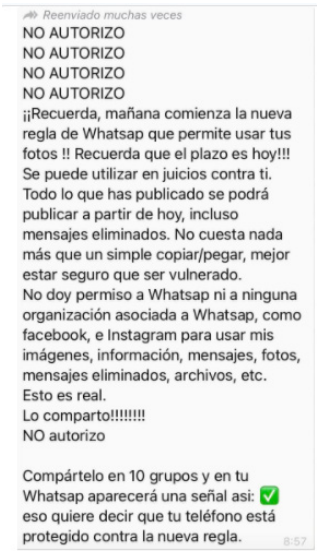
Zombie hoaxes with the most recurrences this month

Hoax	Total alerts	Alerts this month	Alerts this month with FF	% Alerts this month with FF
Chillán volcano in Chile	307	188	119	62.29%
Trojan: Can't believe it's you	337	54	4	7,4%
WhatsApp's new rule	413	51	49	96.07%
Shell Company	352	45	22	48.88%
Call from 626634795	151	43	32	74.41%
Mail from the Dian	232	28	16	57.14%
Laboratory in Wuhan	234	27	26	96.29%
Cuts in pensions	155	26	26	100%
European Parliament and communism	178	18	1	5.55%
"Clean Little Hands"	969	14	0	0
Official cars in the scrap yard	128	4	2	50%
Alkaline foods and coronavirus	145	3	1	33.33%
Money transfer by Ria	187	1	0	0
Sponges for a stealr	166	1	0	0
TOTAL		**503**	**298**	59.24%

The patterns we have detected in alerts tagged with the FF attribute give us very valuable clues on how to spot the most dangerous disinformation that is being shared on WhatsApp at that particular time. This lets us make data-driven decisions on what contents we should invest our limited resources in. We have already seen how this type of alert concentrates on the most viral hoaxes or the ones that tend to reappear from time to time. Now we will analyze concrete examples that show how FF indicates that a hoax is becoming viral and how it helps to quickly detect when a piece of disinformation that was widely shared in the past is spreading once again.

The WhatsApp chain about privacy that always comes back

One of the most prominent zombie hoaxes that we have seen in the analyzed period was shared with us in text format and it references a new alleged WhatsApp rule that would allow the platform to use the photos we send and make all the shared content public.


Reenviado muchas veces
NO AUTORIZO
NO AUTORIZO
NO AUTORIZO
NO AUTORIZO
¡¡Recuerda, mañana comienza la nueva regla de Whatsap que permite usar tus fotos !! Recuerda que el plazo es hoy!!! Se puede utilizar en juicios contra ti. Todo lo que has publicado se podrá publicar a partir de hoy, incluso mensajes eliminados. No cuesta nada más que un simple copiar/pegar, mejor estar seguro que ser vulnerado.
No doy permiso a Whatsap ni a ninguna organización asociada a Whatsap, como facebook, e Instagram para usar mis imágenes, información, mensajes, fotos, mensajes eliminados, archivos, etc.
Esto es real.
Lo comparto!!!!!!!!
NO autorizo

Compártelo en 10 grupos y en tu Whatsap aparecerá una señal asi: ✅ eso quiere decir que tu teléfono está protegido contra la nueva regla.
8:57

The false new WhatsApp alert, one of the zombie hoaxes

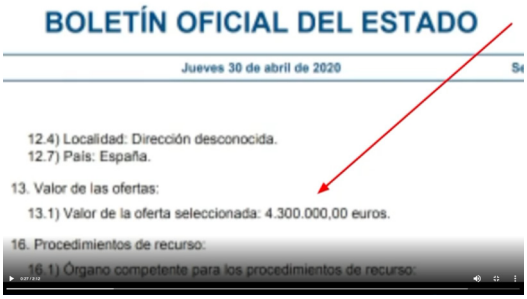
There are 413 notices regarding this hoax registered in our database. It appeared for the first time on January 14, 2021 at 8:16 p.m. and its development is explosive. Only on January 15th we received 296 new notices in our chatbot, more than 1 every 5 minutes, on average. At that time we did not have the alert system with FF and therefore we do not know if they were Frequently Forwarded. As of January 16, the rhythm began to drop and it is noticeable that its diffusion decreased until the end of January. In February it hardly appeared but it returned strongly in March, especially from the 13th on. By then we had already activated the alerts with FF and it is seen in a very clear way how this type of text message chains are marked with this tool.

Of the 51 alerts we received about this hoax in the studied period, 49 were marked by the FF, 96%.

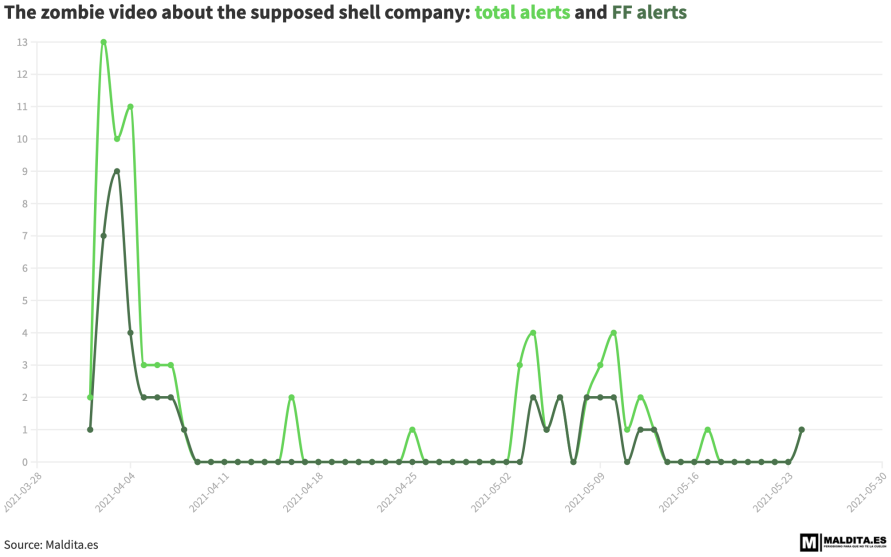
The zombie chain of WhatsApp about privacy. Total alerts vs FF alerts

The zombie video about the supposed shell company

This is a case study of a zombie content in video format that needs context. In it, the Spanish government is accused of doing business with a shell company to buy material to detect COVID-19. 365 alerts appear in the Maldita.es database over time.



The first one arrived on May 3, 2020 when we had not yet activated our chatbot and we were doing a manual data collection: only that day we received 168 alerts on our WhatsApp service. On the 4th, although the intensity dropped to 70 warnings, it is still very active and thereafter it declines. Disinformers usually try to make the most out of the hoaxes that are successful and get such an intense diffusion as this one. Our hypothesis is that this is the reason why on April 1, after a month and a half without receiving any alert about this hoax, we see it again with a significant degree of virality. In three and a half days we received 36 alerts and almost half with FF. We also see that the alerts with FF begin to reach us very soon. 3 of the first 6 are marked. This is a symptom that it is not new content, but rather a video that had previously been circulating on WhatsApp and, therefore, is more likely to accumulate forwards. But it can also alert us that it is recurring content that should be verified.



The zombie video about the supposed shell company. Total alerts vs FF alerts

4. 5. FF ACCORDING TO TOPIC: CASE STUDIES

In the detailed analysis of this month's hoaxes, there are a series of case studies that put forward hypotheses for future research in relation to the high presence of the FF attribute and specific patterns of appearance of these alerts.

Political hoax: Duel in Parliament

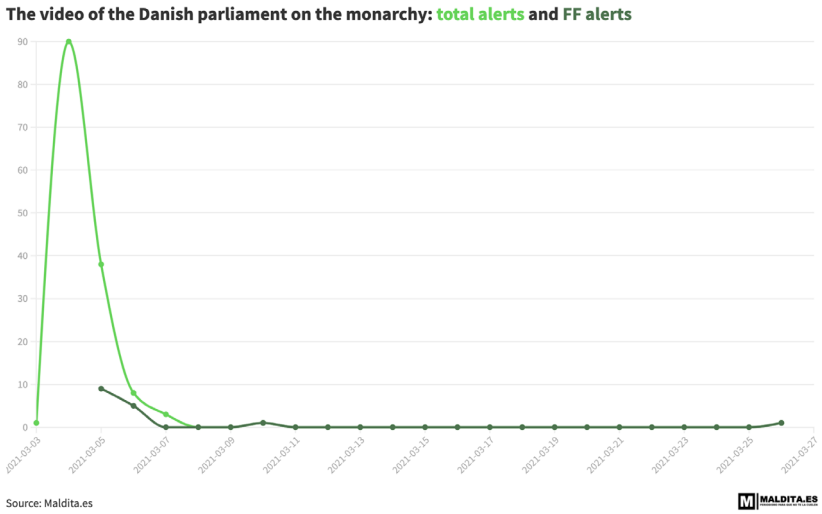
One of the most relevant motivations that we identify based on our experience is when generating disinformation has to do with political reasons. They are hoaxes that seek to modify the perception of a reality to favor a certain ideology: a large part of the hoaxes that we detect have an ideological background.

In the studied period we have been able to see a disinformation duel between two political tendencies that manipulated the same video but with different content. The video they used was of a session of the Danish Parliament in which the deputies began to laugh during the Prime Minister's speech about the government's purchase of elephants from a circus in order to give them a decent life. As they speak Danish, a little-known language, the disinformers manipulated the subtitles and changed them so they had nothing to do with what was actually said, a common practice in disinformation. First, they spreaded a video in which it was simulated that the deputies were laughing at the Spanish monarchy. This was answered from the other side of the political spectrum with another version in which what was simulated is that the laughter was due to the Catalan independence process.

This duel of satirical videos with political content are the most prominent in this format in the period we have investigated. In fact, between the two they add up to 162 notices, 37% of the total alerts in video format linked to some content. Of these notices, 72 have the FF attribute: 23% of the 312 FF in video format that we have received this month.

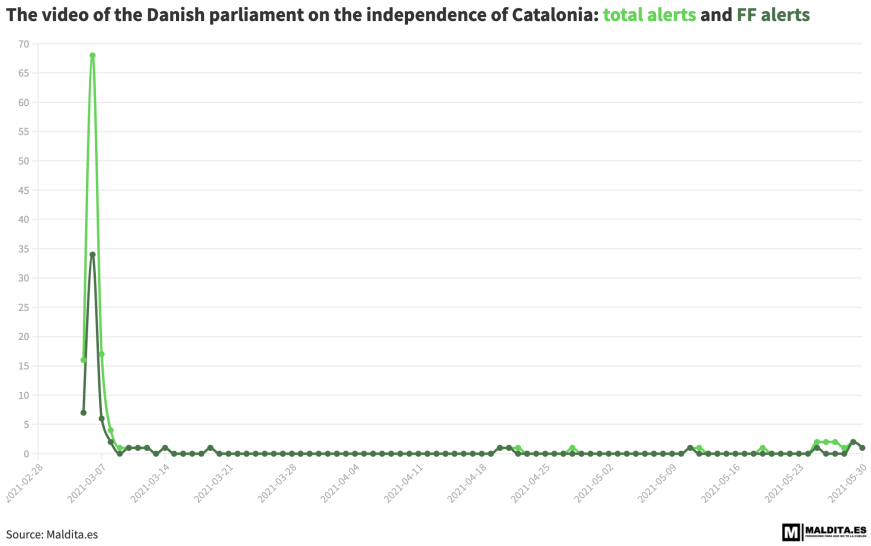
The same video is used against the monarchy (left) and against the Catalan independence movement (right)

The hoax against the monarchy begins on March 3, it's very active on days 4 and 5. On the morning of March 5, the Frequently Forwarded alert system in our chatbot was activated. Very shortly afterwards the first alert with FF about this hoax arrives. It 's at 13:24. When we start to receive the FF alerts, this hoax is already in its decline phase. It continues to appear, although with less intensity, until March 7. Still, in that time we received 16 alerts with FF. The number is not very high but if we look at the structure of the complete warning line of this hoax, the logical thing is to think that in the two days prior to the activation of the system there was also a significant amount of FF that were not yet registered like this in our database.



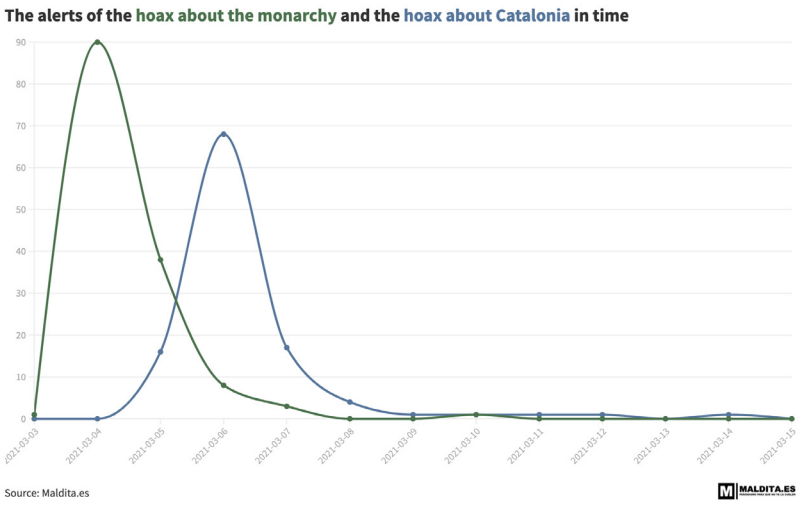
The video of the Danish parliament on the monarchy. Total alerts vs FF alerts

When the hoax against the monarchy began to decline, the response was launched: the doctored subtitles that satirized the Catalan independence movement. It entered the scene on March 5 in the afternoon, we received the first warning at 7:12 p.m., and it was widely shared with us. In this case, as the FF system was already active, we can see the complete evolution of the hoax. The fifth notice we receive about this hoax is already an FF and from then on practically half of the alerts that reach us are marked. In 72 hours we received 101 notices, 47 with FF. But most of these notices are concentrated on March 6. That day alone we received 68 notices, an average of one notice every 21 minutes. There are moments of special intensity. On the 6th of March,, between 8:16 p.m. and 9:26 p.m., we received 9 notices, all with FF, a clear indication that at that time the video was circulating on WhatsApp at full speed.



The video of the Danish parliament on the independence of Catalonia. Total alerts vs FF alerts

The two cases have a similar pattern, that of intensely viral hoaxes. They get a very strong diffusion shortly after being created and that causes queries about this hoax to skyrocket in the following hours. After a peak in which they almost monopolize the chatbot they begin to decline until they disappear. This structure is common in big political hoaxes. They follow the news and have many actors willing to spread them through social networks. When some news gives them the opportunity to share, all the machinery that supports the ideological group benefited by the hoax tries to spread it as much as possible. That is why it is common for our chatbot to receive an avalanche of queries in a short period of time about these hoaxes.



The alerts of the hoax about the monarchy and the hoax about Catalonia in time

False hacking alerts

There is a type of disinformation that is not linked to the present but they return periodically and they do so with great strength. They are the text message chains that warn about alleged hacking attempts. They get a lot of virality, sometimes taking advantage of the good intentions of users who forward messages to alert their family and friends of the alleged danger reported in the message.

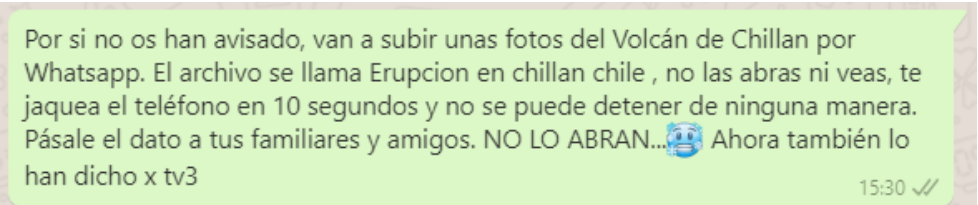
In these cases we see that the same text has been viral for years. The records that reach the chatbot show avalanches of warnings about these hoaxes that last only a few days but acquire a high intensity. These messages concentrate many FFs for several reasons:

- It is easy for text chains to accumulate forwardings because it is simple to make and to copy and paste them.
- Because of the high intensity with which they circulate.
- Because they are old messages that come back to life and that characteristic is also related to a high number of FFs.

In the month that we have analyzed we have several examples of this:

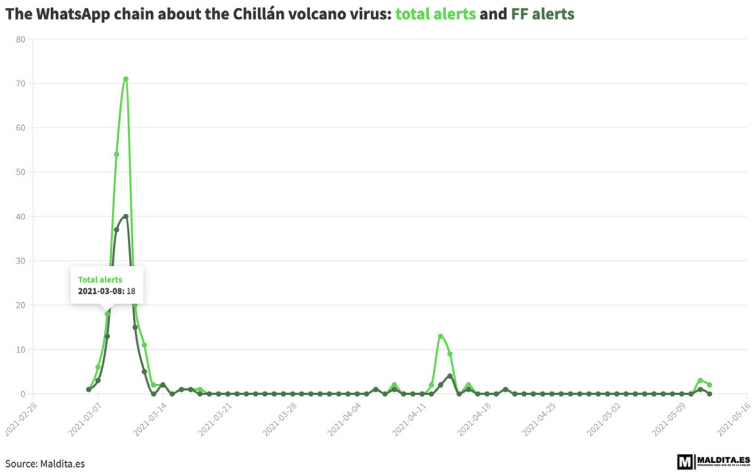
Chile's Chillán volcano

In our database there are 321 alerts in total about an alleged mobile hacking that is activated when opening a photo of a Chilean volcano called Chillán that circulates on WhatsApp. It is the content with the highest number of alerts in the period analyzed although it is also a zombie hoax.



WhatsApp chain about the Chillán volcano y Chile

The first one that we received was on October 9, 2019. It was shared a lot between October and November of that year, it appeared with less intensity until February 2020, then it disappeared. But the disinformers put it back into circulation on March 6, 2021. It reappeared with great force, especially between March 8 and 11. In those 4 days this hoax practically monopolized the alerts in text format that we received in our chatbot. Of the 186 that we received,163 were on the Chilean volcano. Predictably, the FF percentage was also very high. 65% of the alerts in those four days had FF.

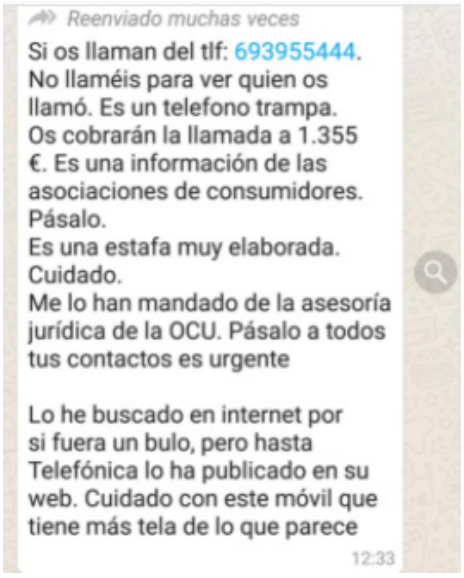


The WhatsApp chain about the Chillán volcano virus. Total alerts vs FF alerts

This hoax is the number one of the month as it's the one that was brought to our attention the most by our community during that period. In total there are 188 notices that were concentrated in just 12 days, 119 of them with FF.

Very expensive calls

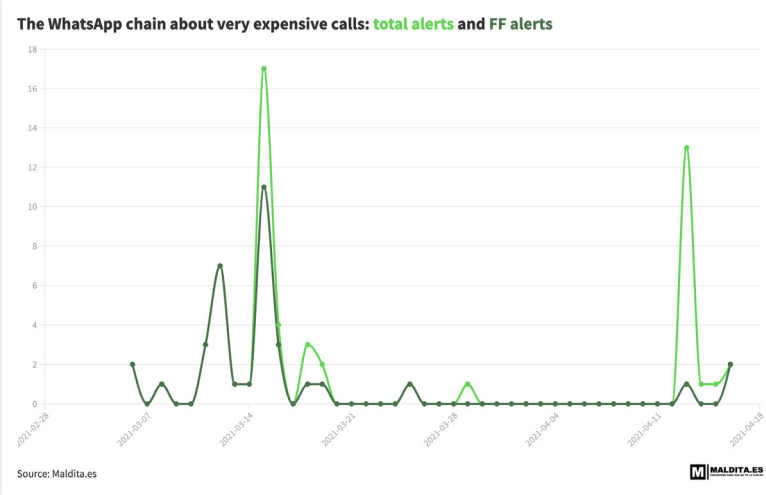
We find almost the same pattern in this other hoax with a very similar theme. Here the disinformers show two phone numbers and warn that they are going to charge you a high amount of money if you answer their calls. It is false, and it tries to go viral by inciting users to share it with their contacts to prevent them from falling into the allegeds call.



The hoax about the call is in the Maldita.es database since 2019

In our database there are 161 notices about this hoax. We saw it for the first time in September 2019, it disappears in January 2020. After a year without hearing from it, we saw it again in January 2021 when we received 5 notices. But it is in March, within the month that we have analyzed, when it goes viral again. The first 19 alerts we received in March are all with FF. In other words, as soon as the alert system with FF was activated, it was clear that this hoax was highly viral. In addition, this hoax has another peculiar characteristic. As of March 14, we detected that the disinformers were trying to expand the public to which this hoax could reach and we started to receive

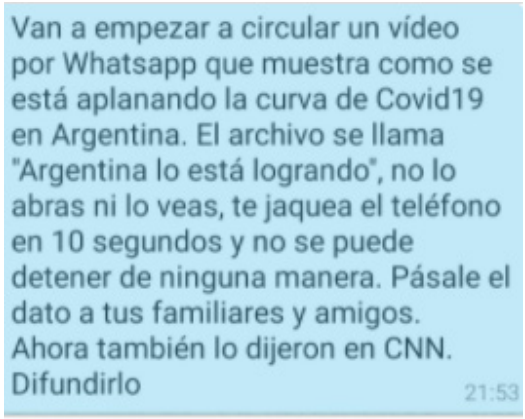
warnings about a version translated into Catalan that was also being shared. In total, during that month we received 43 notifications about this hoax in the chatbot, 32 with FF, that is, 74%.



The hoax about the call is in the Maldita.es database since 2019

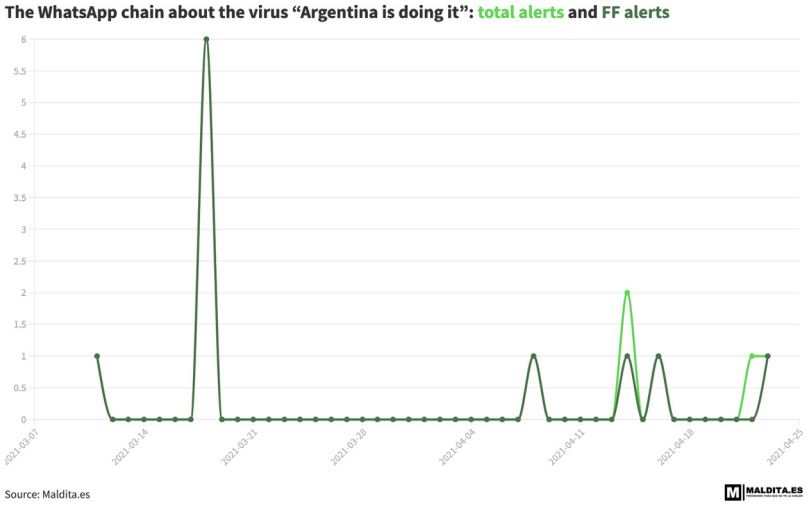
The traveling hoax

It is another example of a text chain that circulates on WhatsApp and that attracts a very high percentage of notices with FF. In this case, it also warns of an alleged mobile hacking through a video titled "Argentina is doing it." It is an old hoax that we first detected in April 2020 and we saw it again in March 2021.



Example of hoax detected in Latin America and Spain

But it's worth mentioning it in this report because it has a feature that increases its chances of attracting FF. It is a hoax that has also been shared in Latin America. For example, it has also been debunked by the Argentine verification organization Chequeado or the Bolivian Bolivia Verifica. As it is a message that spreads from one country to another, the chances of accumulating forwardings increase. In our database, during the study period we received 7 notices, 6 of them with FF.



The WhatsApp chain about the virus "Argentina is doing it". Total alerts vs FF alerts

Possible disinformation campaign against migrants

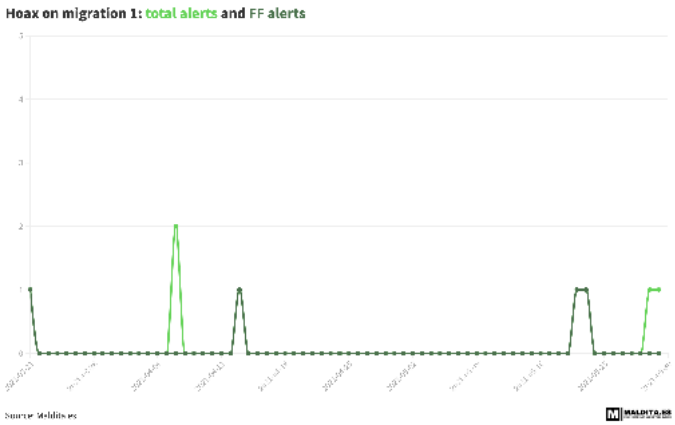
In the analysis of the Frequently Forwarded, we have also detected indications that could point to coordinated disinformation actions. In the month we studied, we discovered that between March 21 and 22 we had been warned of four different disinformations that tried to criminalize immigrants. They were four old hoaxes that users had once again warned us about in a short space of time and, although they weren't followed-up too much, they enabled us to hypothesize that a disinformation campaign had been launched against the immigrant population.

We detected the hoaxes we are referring to within 24 hours, between March 21 at 10:08 p.m. and March 22 at 10:23 p.m., and it consists of four videos. All attack immigrants and three of them accuse them of receiving large benefits from the state in an irregular way. The four have been in our database for months and the notices with FF suggest that they they had been previously shared because in the period we studied they did not manage to go viral, according to our records. We see them one by one.

Hoax on migration 1: The woman who rejects a job to continue receiving benefits

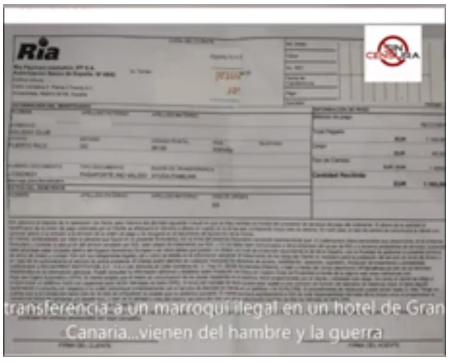
The first of the hoaxes claims that a Muslim woman rejects a job in order to continue receiving benefits. This lie was reported to us for the first time in March 2019. It continues to appear sporadically during 2019 and 2020. On March 21 at 10:08 p.m. we detected it again. It was the first time that it was shared

with us with the FF system in-place, and the particular alert was marked with Frequently Forwarded. In April and May we have seen it another 4 times, 2 of them with FF.

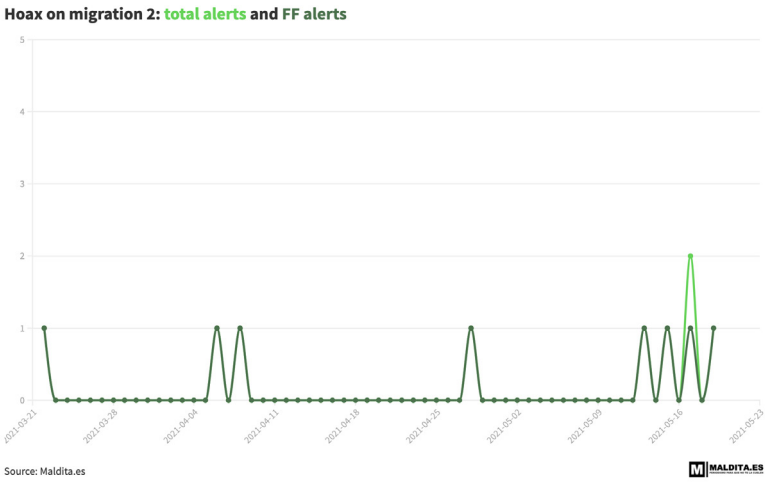


Hoax on migration 1. Total alerts vs FF alerts

Hoax on migration 2: the irregular immigrant who receives benefits

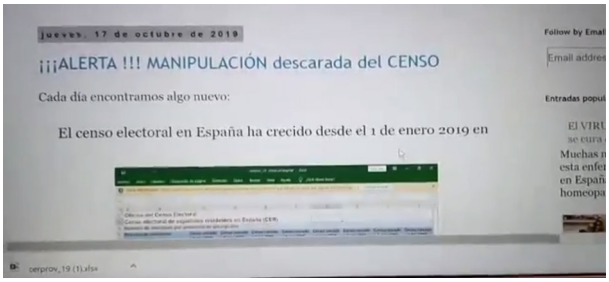


Another hoax that reappears in those days is refered to an undocumented immigrant who is accused of receiving irregular benefits. We saw this disinformation a lot in December 2020 and it continued to appear regularly until February 22. Then it disappeared for a month and returned on March 22 at 1:23 a.m. Outside the study period, beyond April 4, we found evidence that it is a video that is routinely used by disinformers who attack immigrants. Between April and May we have been alerted about this hoax 8 more times, 5 of them marked with FF.



Hoax on migration 2. Total alerts vs FF alerts

Hoax on migration 3: the inflated electoral roll

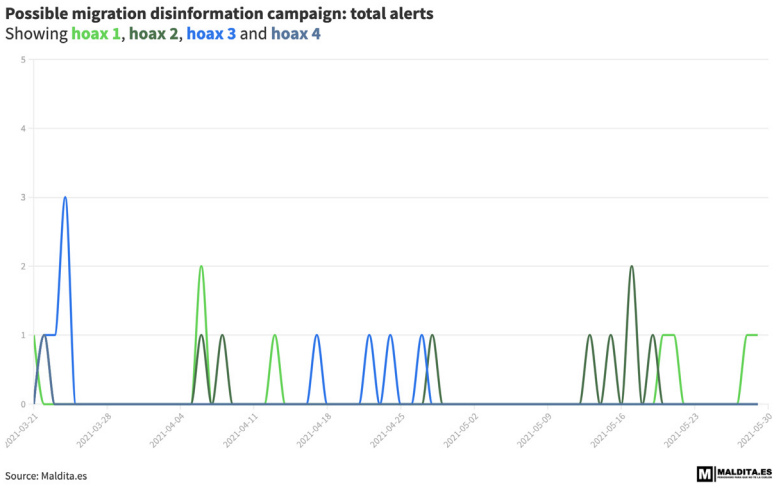


The third video hoax against immigrants was first sent to us on March 22 at 8:47 p.m. and refers to a false manipulation of the electoral roll to let 200,000 irregular immigrants vote. This one also appeared in 2019. We see it again intermittently in the following months. In 2021 we were alerted of this hoax on February 14 and it stopped being shared for a month until March 22, when we received an alert marked with FF. Since then we have received eight more alerts about this hoax, 3 of them with FF.

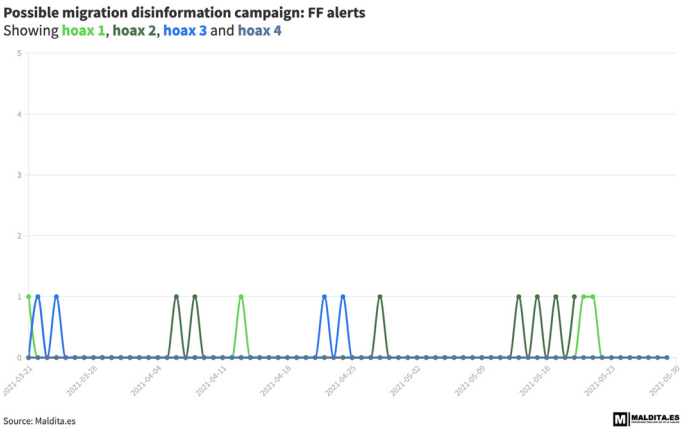
Hoax on migration 3. Total alerts vs FF alerts

Hoax on migration 4: the irregular trafficker who receives public benefits

The last video also tries to link an immigrant who is supposedly involved in drug trafficking with irregular State benefits. They alerted us about it at 22:23 on March 22. This other hoax that had not appeared for a month, but between January 31 and February 9, 2021 we had received 20 further notices about this disinformation.



These four hoaxes manifest an attempt, marked with alerts associated with the FF, to resurface them all at once in 24 hours, which makes us think that it could be an orchestrated campaign and that therefore in the future we must investigate whether this attribute can help us identify organized disinformation campaigns.



There is another piece of data that supports the suspicion that these four videos were part of an attempt to launch a campaign. From May 18, 2021, we detected a wave of disinformation against immigrants as a result of the irregular entry of several thousand people from Morocco into the Spanish city of Ceuta. At Maldita.es we detected two dozen hoaxes created from that day on. In this context, on May 19 and 20 we were again alerted of two of those videos. Specifically, we received notices about the hoax of the Muslim woman who rejects a job and that of the black immigrant who is accused of collecting illegal benefits.

Despite all these elements, the period we have analyzed is short and the number of examples we have detected is small. A more in-depth study of the Maldita.es database would be necessary to confirm whether these patterns that we have detected on a small scale are repeated regularly and the hypothesis that FFs may be a solid indication of a coordinated action in a specific issue such as immigration.

06

Conclusions

06

CONCLUSIONS

In countries where the consumption of information through WhatsApp is important, a WhatsApp service such as the Maldita.es chatbot is one of the main tools for a verifier when it comes to detecting possible hoaxes: it is a spyhole to detect the disinformation that is affecting the conversations that citizens have in closed environments. Thanks to the collaboration of our community, we can find out what is happening within WhatsApp despite being an end-to-end encrypted network, according to its creators: their messages help us to know what is going viral at any given moment. 78% of the content debunked by Maldita.es in the analyzed period had been sent to us through WhatsApp.

For a WhatsApp service to be useful, it is necessary for the verifier tto create a community that is willing to collaborate by denouncing possible disinformation and viralizing the debunks in their own conversations. An automated chatbot allows users in that community to scale without collapsing either the system or the team of journalists in charge of verification.

In addition, the FF attribute associated with the content can help verifiers identify and prioritize the most viral content and therefore having the biggest impact. Our analysis has identified that in 80% of cases the most viral content is also the one with the most alerts with associated FFs. In addition, when the verifiers apply the methodology choosing what contents should be investigated according to virality and dangerousness, the alerts associated with the FF are three times more likely to be identified as possible disinformation by journalists than those that do not have it. Finally, it is important to note that in the month analyzed 78% of the alerts with FF investigated ended up being flagged as hoaxes or disinformation. Therefore, the FF attribute indicates that the most viral contents are the ones that verifiers are most likely to follow up,according to their methodology, and the ones that tend to be classified as disinformation. This attribute can therefore be useful for verifiers in scaling up their work and deciding what to investigate first in a situation of limited resources.

In short, the WhatsApp chatbot and the alerts with FF help us to articulate an early warning system that can allow us to react quickly to stop the virality of a hoax that begins to spread, neutralize recurrent hoaxes as soon as they reappear and to quickly detect disinformation campaigns organized with specific motivations, although this needs further study.