MALDITA.ES

DESINFORMACIÓN EN WHATSAPP:
EL CHATBOT DE MALDITA.ES Y EL ATRIBUTO "REENVIADO FRECUENTEMENTE"

Junio 2021

Un informe de la unidad de
Investigación Académica de Maldita.es

Este es un informe desarrollado por la unidad de investigación académica de Maldita.es. El fin de esta unidad es proveer de estudios e investigación a la redacción de Maldita.es con el fin de que su trabajo y las decisiones que ésta toma estén apoyadas en datos y dirigidas por los resultados de la investigación. La unidad de investigación académica trabaja tanto en colaboración con universidades y centros de investigación como de manera independiente.

Puedes ponerte en contacto con ella en
academia@maldita.es

ÍNDICE

01

Introducción

01

INTRODUCCIÓN

En la base de datos de Maldita.es se recoge el rastro que está dejando la desinformación en España en estos últimos años. Cada uno de los elementos de esos archivos aporta una información concreta que permite conocer cómo funciona la desinformación, cuándo aparece, cómo se mueve o qué técnicas se utilizan para tratar de llegar lo más lejos posible.

Cuando en julio de 2020, en mitad de la pandemia, lanzamos nuestro chatbot de WhatsApp, comenzamos a recopilar esos datos de llegada de desinformación de manera automatizada, enriqueciendo nuestra base de datos y normalizando una metodología que hasta el momento había sido manual. Desde marzo de 2021 se ha incluido un nuevo elemento en esa base de datos, los mensajes marcados con el atributo Frequently Forwarded (FF), que señala aquellos que ya han sido reenviados en WhatsApp cinco o más veces. Este atributo también se conoce como "Highly Forwarded Messages' (HFMs), mensajes altamente reenviados en castellano.

El equipo de investigación académica de Maldita.es ha hecho un primer análisis del significado y relevancia de ese nuevo atributo. Estudiando los patrones de aparición del FF durante el primer mes en el que estaba activo se han encontrado indicios de que está fuertemente relacionado con la desinformación. El 74% de las alertas con FF pudieron ser vinculadas por los periodistas de Maldita a una de nuestras investigaciones ya abiertas por parte del equipo de Maldita.es y el 78,72% de las alertas investigadas se acabó confirmado que señalaban a un bulo o desinformación.

Además, en los patrones de aparición de las alertas con FF observamos que aparecen asociadas a bulos con un impacto potencial muy alto, ya sea por que es probable que lleguen a ser muy virales o porque se trata de bulos antiguos que ya tuvieron un impacto importante en el pasado y vuelven a estar en circulación. También hay rasgos que apuntan a que este tipo de alertas pueden señalar la aparición de campañas coordinadas de desinformación.

Estos datos abren la puerta a la elaboración de un sistema de alerta temprana que permita reaccionar antes a este tipo de desinformaciones. Son algunas de las líneas de investigación prometedoras que se abren después del análisis

de solo un mes de funcionamiento del atributo FF. Un estudio más amplio y profundo permitirá confirmar estas hipótesis y ampliar el conocimiento sobre las posibilidades que aportan las alertas con FF para facilitar el trabajo de los verificadores.

En este informe se analiza la importancia y utilidad del servicio de WhatsApp para entidades de verificación como Maldita.es tanto como vía de detección de la desinformación viral en conversaciones privadas a través de denuncias de la comunidad y como vehículo para hacer llegar información contrastada a la ciudadanía. También se evalúa el impacto y la progresión del chatbot de WhatsApp desarrollado por Maldita.es en julio de 2020 y los avances que ha supuesto la automatización en comparación con el antiguo sistema manual de WhatsApp Business. Además, se profundiza en la utilidad del atributo "Reenviado Frecuentemente" (Frequently Forwarded) asociado a los contenidos para los verificadores mediante casos de estudio. Maldita.es es el primer verificador a nivel mundial en investigar su utilidad para la comunidad de fact-checking.

Además, nuestro chatbot de WhatsApp ha sido galardonado con el Premio Europeo de Periodismo en la categoría de innovación en 2021.

+34 644 22 93 19

SERVICIO AUTOMÁTICO DE WHATSAPP

02

El uso de Whatsapp en España: información y desinformación

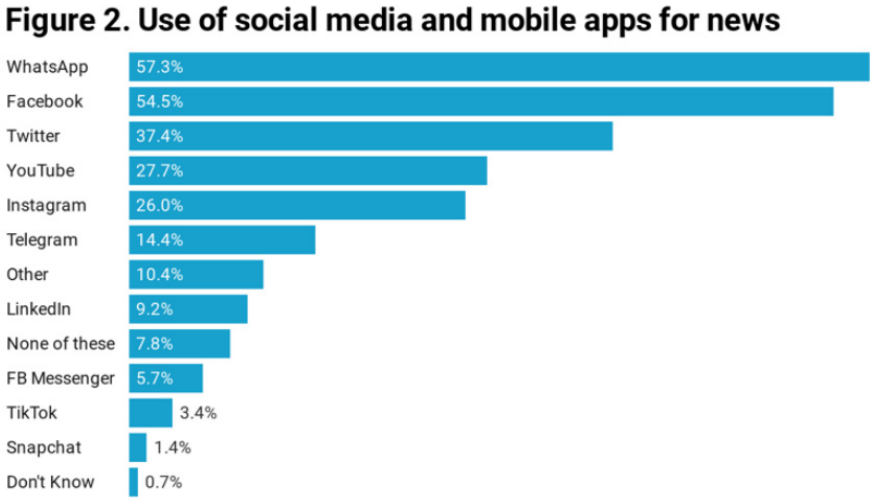
02

EL USO DE WHATSAPP EN ESPAÑA: INFORMACIÓN Y DESINFORMACIÓN

WhatsApp es el espejo de lo que se está hablando en cada momento en España. Las estadísticas indican que su uso es masivo: el 85% de los españoles utilizan la aplicación de mensajería a diario según el Estudio Anual de Redes Sociales de IAB Spain. Además, el uso de WhatsApp es transversal. Según el Digital News Report del Reuters Institute de 2020 la penetración de la aplicación es enorme en personas de todas las franjas de edad, todas las situaciones económicas y niveles de estudios. En todos esos grupos se registra un porcentaje de uso diario superior al 68%. Pero también es muy relevante para qué lo emplea la sociedad española. WhatsApp no sólo sirve para estar en contacto permanente con familiares, amigos y conocidos, también se ha convertido en una vía esencial para recibir información. El mismo Digital News Report señala que el 34% de los españoles empleo WhatsApp para informarse en 2020, lo que convierte a España en el país europeo en el que más se usa esta aplicación para recibir noticias de todos los analizados en el informe.

EL 34% DE LOS ESPAÑOLES EMPLEA WHATSAPP PARA INFORMARSE

También ha sido el canal al que se ha recurrido en busca de información durante la pandemia de COVID-19. El porcentaje de españoles que ha buscado información en la aplicación llegó a alcanzar el 57% en mayo de 2020, según este estudio de la Universitat Internacional de Catalunya (2020). De hecho, otro estudio de la Universidad Miguel Hernández (2020) señala que, durante los primeros meses de la pandemia, WhatsApp se convirtió en el medio más usado por todas las franjas de edad para compartir información sobre el coronavirus.



Uso de WhatsApp para informarse durante la pandemia. Fuente: UIC

Un año después del comienzo de la pandemia, WhatsApp continúa siendo una fuente de información sobre la COVID-19 para un 26% de la población española, según un informe del Reuters Institute de Oxford (2021). Es la mayor cifra entre los países europeos que aparecen en este estudio. Esta misma investigación detecta que en las aplicaciones de mensajería existe un problema con la desinformación. El 29% de los españoles reconoce que recientemente ha recibido a través de servicios de mensajes mucha desinformación sobre la COVID-19. Ese porcentaje es, una vez más, el mayor entre los países de Europa. Según este informe, estas aplicaciones son el tipo de plataforma digital que menos confianza genera a la hora de informarse sobre la pandemia en España. Sólo el 16% se fía de ellas, cinco puntos menos que de las redes sociales. Es importante señalar que esta investigación descubre una relación entre esas personas que se fían de los contenidos que les llegan por servicios de mensajería y la creencia en bulos. Cuanto más confía una persona en lo que recibe por estas aplicaciones, más probabilidades hay de que se crea desinformación sobre la vacuna del coronavirus. España es el tercer país tras Alemania y Estados Unidos en el que esa correlación es más fuerte.

Podemos concluir que WhatsApp es un pilar fundamental de la comunicación y de la información en España, pero que es muy sensible al impacto de la desinformación y por lo tanto un espacio que se convierte en peligroso. Además, como los mensajes están cifrados de extremo a extremo, es muy difícil para las instituciones y organizaciones de verificación localizar y desmentir los bulos que se difunden en esta plataforma.

2.1 Verificación en WhatsApp y la importancia de la comunidad

Cuando estrenamos nuestro servicio manual de WhatsApp Business en julio de 2018 lo hicimos siendo conscientes de que una gran parte de la desinformación que afectaba a la ciudadanía española lo hacía a través de esta plataforma conversacional. Ya entonces el consumo de información a través de WhatsApp incluía, según el Digital News Report de ese año, al 36% de la población española y nuestro primer número de WhatsApp nos sirvió para entender el fenómeno de la desinformación y diseñar la que sería nuestra estrategia para luchar contra la desinformación en los años siguientes.

El primer paso para luchar contra la desinformación es detectarla cuanto antes y para eso necesitas tener acceso al lugar en el que prolifera. En plataformas cifradas de extremo a extremo, como WhatsApp, sólo podemos saber qué es lo que se está viralizando en las conversaciones si los propios usuarios nos lo hacen saber. Por eso la estrategia de Maldita.es en los últimos años ha sido construir una comunidad que a día de hoy cuenta con más de 48.000 usuarios registrados que colaboran con nuestra redacción a la hora de detectar los

contenidos virales en redes sociales, pero también en sus conversaciones privadas. Personas comprometidas en la lucha contra la desinformación que nos avisan cuando ven contenidos sospechosos en WhatsApp, Telegram y otras plataformas.

48.000 USUARIOS REGISTRADOS QUE COLABORAN CON NUESTRA REDACCIÓN A LA HORA DE DETECTAR LOS CONTENIDOS VIRALES

Es un camino de doble sentido porque esas mismas personas que nos ayudan a identificar lo que se está volviendo viral en esas plataformas son también las que llevan el contenido verificado a sus conversaciones privadas y grupos. A los lugares a los que Maldita.es no puede llegar directamente, pero donde es más necesaria la información contrastada. El nuestro es un trabajo que no es posible sin nuestra comunidad.

Durante casi dos años utilizamos una cuenta de WhatsApp Business que, más allá de identificarte como persona jurídica, no tenía grandes diferencias con el WhatsApp de cualquier usuario: un número al que se podía acceder de manera individual y contestar uno a uno a los que habían escrito solicitando información. En una cuenta que recibía de media mensajes de entre 200 y 300 usuarios al día, eso suponía dedicar a un periodista a jornada completa sólo a contestar dichos mensajes. Era una tarea repetitiva y dura porque requería exponer a un redactor de manera constante a los bulos que nos llegaban, muy a menudo vinculados a discurso de odio.

Durante casi dos años utilizamos este sistema y llevamos a cabo un recuento manual de la cantidad de veces que nos llegaban las desinformaciones para poder medir su viralidad. Muchas veces nuestro WhatsApp se saturaba, incapaz de gestionar el aluvión de consultas sobre la veracidad de contenidos relacionados con una campaña electoral o un atentado en los que las denuncias podían alcanzar picos de hasta 800 usuarios en un día. Eso significaba no sólo que nosotros no podíamos acceder a los contenidos virales que nos enviaban los usuarios, sino también que esos usuarios se quedaban sin una respuesta a sus consultas.

03

El chatbot de Maldita.es: un diseño colectivo

03

EL CHATBOT DE MALDITA.ES: UN DISEÑO COLECTIVO

Con la llegada de la pandemia de COVID-19 el volumen de consultas se hizo imposible de gestionar: pasamos de 200-300 usuarios consultándonos diariamente a 2.000. Nuestro servicio de WhatsApp no cumplía ni su misión de alertarnos de los contenidos virales ni su misión de dotar de información contrastada a la ciudadanía. Fue entonces cuando conocimos a COVIDWarriors y al equipo de Wealize, y empezamos a trabajar en la creación de un chatbot automatizado que fuera capaz de recibir, almacenar e identificar los contenidos por los que nos preguntaban con el fin de responder de manera automática a las consultas ya resueltas.

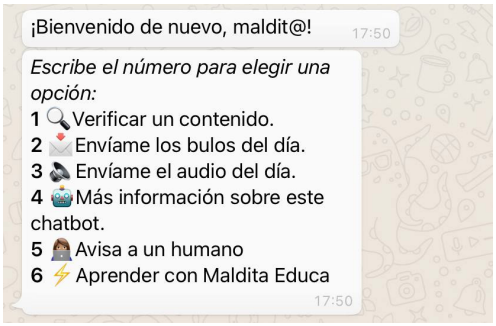
En julio de 2020 lanzamos nuestro chatbot de WhatsApp para automatizar parte de ese proceso y responder de forma inmediata a quienes se ponían en contacto con nosotros.

Después de casi un año en funcionamiento más de 26.000 personas han utilizado nuestro servicio automatizado, de las cuales el 61% lo ha hecho más de una vez. Nuestro chatbot ha enviado más de 400.000 mensajes, ha verificado 108.000 contenidos y ha facilitado 143.800 resúmenes de desmentidos diarios en formato de texto y 18.400 en formato de audio.

26.000 PERSONAS HAN UTILIZADO NUESTRO SERVICIO AUTOMATIZADO, DE LAS CUALES MÁS DE LA MITAD LO HA USADO MÁS DE UNA VEZ

Sólo en 2021, 16.200 usuarios nos han contactado, con una media de más de 950 conversaciones al día entre los que buscaban verificar contenidos y los que solicitaban otras opciones de nuestro chatbot.

Desde que estrenamos este servicio automatizado, cuando alguien manda un mensaje al chatbot recibe un menú con varias opciones entre las que puede elegir verificar un contenido, recibir los bulos desmentidos ese día por Maldita.es o aprender lecciones con Maldita Educa entre otras. Mediante un menú numérico los usuarios pueden elegir entre 6 opciones:



Menú general del chatbot

QR de acceso directo al chabot

1. Verificar un contenido

Si un usuario ha recibido un contenido que quiere verificar y pulsa la opción '1', el chatbot le pedirá que envíe la foto, el vídeo, el audio o la cadena en cuestión.

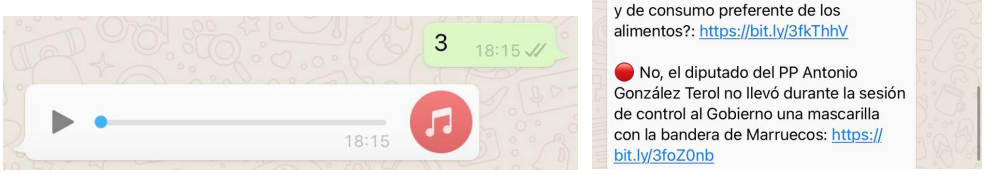
En el caso de que envíe un contenido que Maldita.es ya ha desmentido, el usuario recibe automáticamente el artículo relacionado. De lo contrario, el usuario recibirá un mensaje informándoles de que un redactor revisará el mensaje mandado por el usuario para verificarlo.

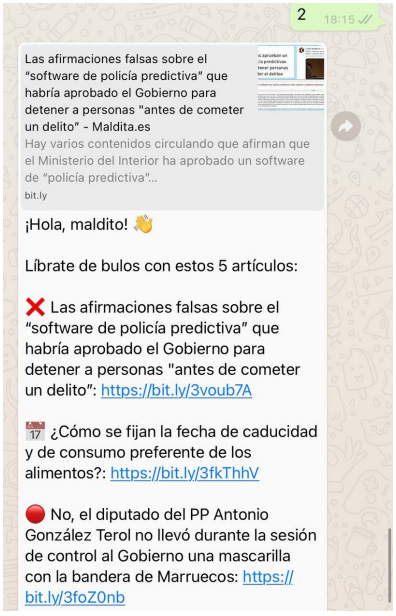
Además de que el usuario recibe automáticamente el desmentido de Maldita.es, esto nos permite saber qué bulos se están moviendo y cómo de virales

Mensajes de respuesta a las preguntas de los usuarios

2 y 3. Envíame los bulos y el audio del día

Para facilitar el acceso a los artículos de Maldita.es, el chatbot incluye la opción de pedir los artículos más importantes publicados ese día. También se ofrece un audio en el que un redactor narra los titulares de esos artículos.





El chabot ofrece los contenidos más importantes del día

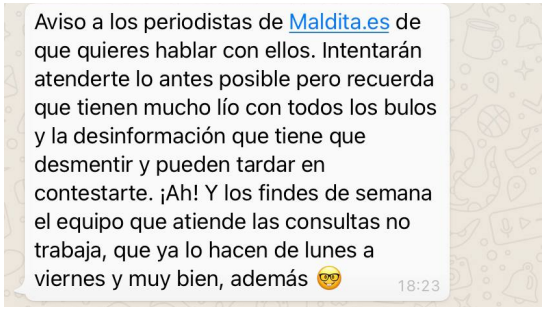
4. Más información sobre este chatbot

Dado que no todo el mundo está familiarizado con el concepto de 'chatbot', creemos que es importante ofrecer una opción para explicar qué es y qué hace. Cuando los usuarios pulsen el '4', recibirán un mensaje en el que se indica que es Maldita.es, en qué consiste este chatbot y su política de privacidad.

En el chatbot también hay información sobre el propio chatbot

5. Avisa a un humano

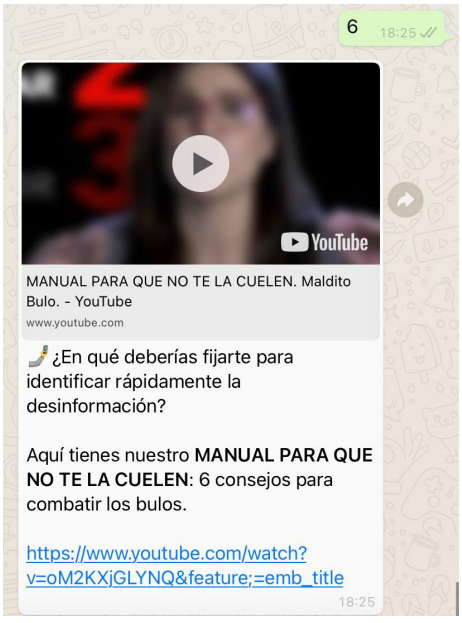
Es posible que el usuario no quede satisfecho con la respuesta a su consulta o simplemente que prefiera hablar con un humano. Al pulsar esa opción, el chatbot recoge dicha consulta para que un periodista de Maldita.es la revise.

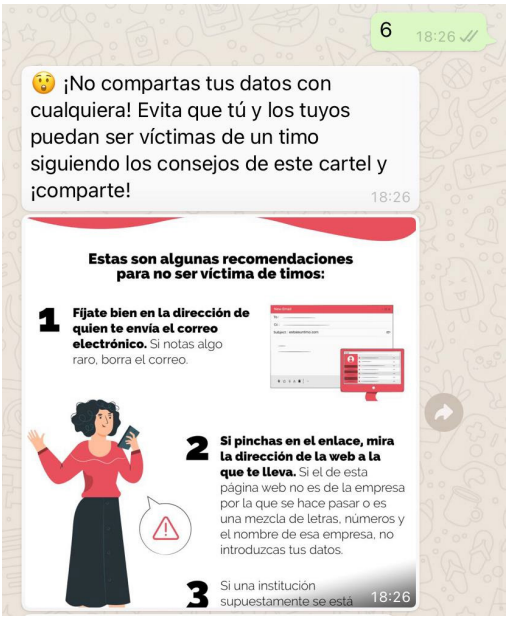


Si el chatbot no tiene respuesta, llama a un humano

6. Aprende con Maldita Educa

La alfabetización mediática y dar a la sociedad herramientas para que no se la cuelen es otro de los pilares de Maldita.es, por lo que también ofrecemos en nuestro chatbot la opción de que el usuario acceda a algunos de los muchos recursos que hemos preparado sobre este tema.





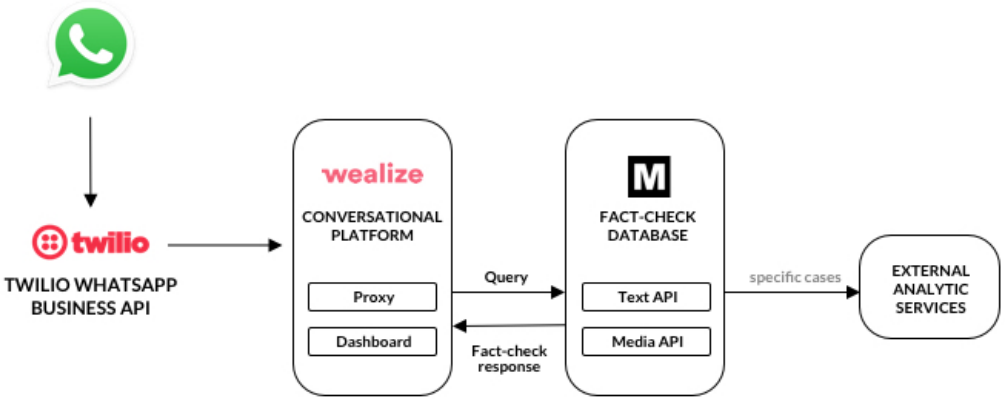
El chatbot también da acceso a contenidos educativos

Arquitectura del chatbot

El chatbot es la principal vía que tiene nuestra comunidad para alertarnos de posibles bulos en WhatsApp y se ha convertido en un elemento esencial de nuestro trabajo diario. No sólo procesa casi mil avisos de media cada día, también nos permite tener una visión de lo que se está viralizando en grupos cerrados de WhatsApp, donde no podríamos llegar si no fuera por las denuncias de nuestros usuarios debido a la encriptación de punta a punta que utiliza la plataforma, según ella misma.

La arquitectura se separa fundamentalmente en cuatro niveles diferentes:

- La conexión con WhatsApp vía Twilio
- La plataforma conversacional de Coloq.io,
- La base de datos de Maldita.es junto con el frontend de gestión
- Los servicios externos utilizados para procesar algunos tipos de desinformación como audios o imágenes.



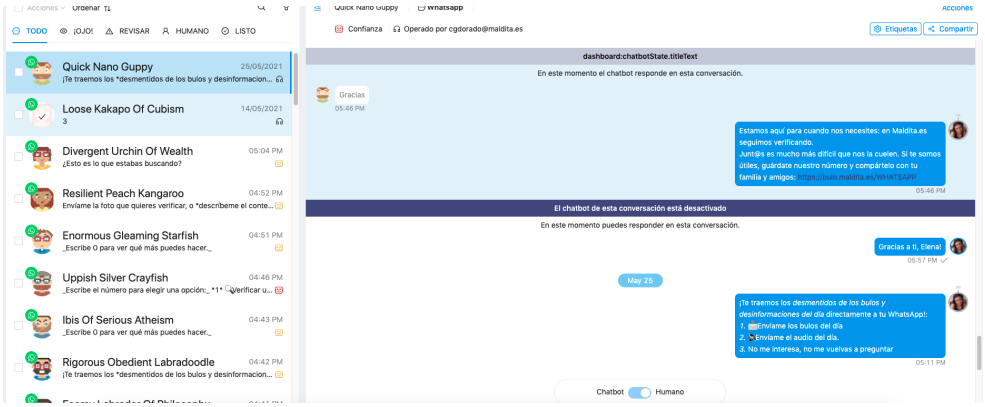
Arquitectura del chatbot

3.1 La plataforma conversacional de Coloq.io

El chatbot de Maldita en Whatsapp se construye sobre Coloq.io, una plataforma conversacional que permite conectar con usuarios en diferentes canales de conversación como Whatsapp a través de Twilio, pero también Webchat, Twitter, FB messenger y SMS, Google Business Messages; entender y responder a los mensajes mediante herramientas de inteligencia artificial conversacional, como IBM Watson Assistant, Google Dialogflow o NLTK; consultar fuentes de información e integrarse con procesos de trabajo; que todo el proceso esté supervisado por personas mediante reglas y alertas personalizadas; y obtener datos suficientes para poder generar informes como éste. Coloq.io es un producto de Wealize, una empresa de productos digitales especializada en blockchain e inteligencia artificial conversacional.

El usuario manda un mensaje por Whatsapp, se recibe por la API de Whatsapp Business en Twilio y se procesa con inteligencia artificial conversacional, manteniendo una conversación para verificar noticias, contestar preguntas u ofrecer más información.

Para cada conversación, se establecen una serie de reglas y alertas que permiten ofrecer la mejor información verificada a los usuarios. Además, Coloq.io permite que la conversación sea intervenida por una persona mejorando la experiencia del usuario, así como el envío de mensajes individuales o a varios usuarios según sea necesario.



Frontend de gestión conversacional de Coloq.io

3.2 La base de datos de Maldita.es

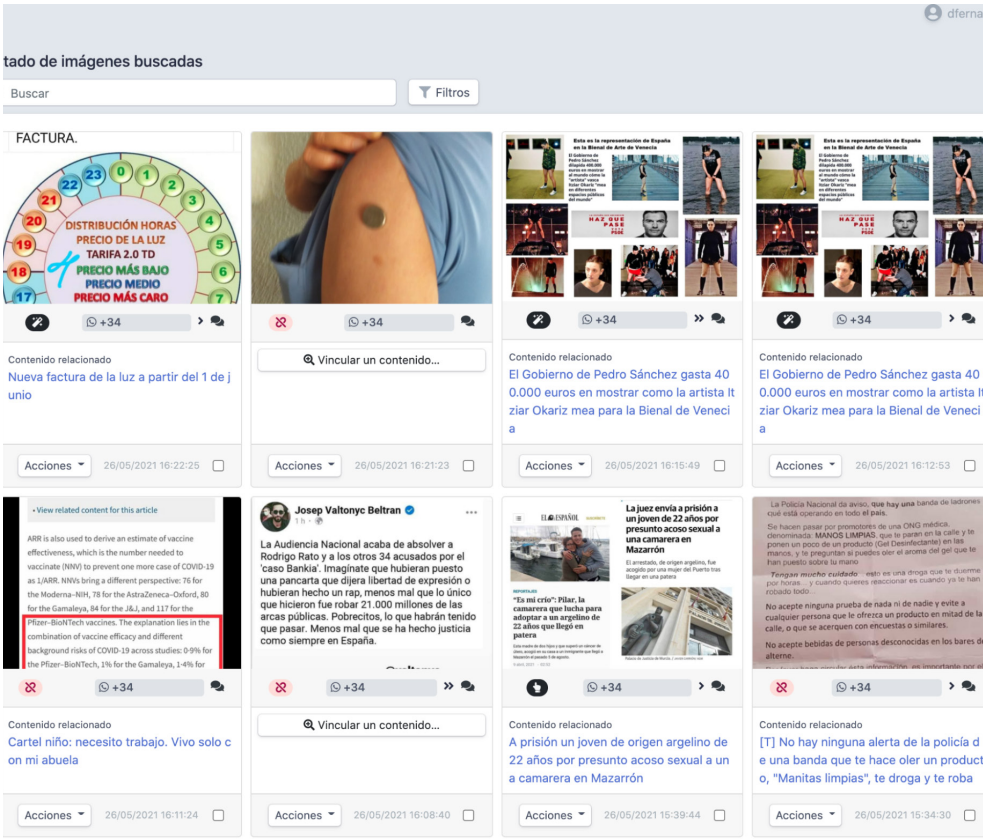
La base de datos de Maldita.es se compone de tres pilares diferenciados: la API, el sistema de búsqueda y el frontend de gestión.

Las consultas que se reciben a través de la API se manejan de forma distinta dependiendo de su formato:

- En el caso de imágenes, vídeos y audios son analizados tanto de forma interna como mediante servicios externos para extraer posibles coincidencias de la base de datos o cualquier elemento identificativo de texto que pueda permitir realizar búsquedas en la base de datos.

- En el caso de contenido de texto se buscan directamente mediante el sistema de búsqueda.

Los resultados de esa operación, junto con la búsqueda en sí, quedan registrados en la base de datos y pueden consultarse mediante el frontend de gestión.



Frontend de gestión de Maldita.es

Mediante este frontend pueden consultarse las búsquedas separadas por formato (texto, vídeo, audio o imágenes) permitiendo acceder a datos como la transcripción de un audio o una imagen, la interpretación del sistema de procesamiento de lenguaje del bot, la conversación en la plataforma de Coloq. io, el número de veces que nos ha llegado o los diferentes formatos en los que ha venido. Además, se utiliza para registrar nuevos contenidos de potencial desinformación para hacerles seguimiento y que a partir de ahí los contenidos idénticos vayan agregándose para medir la viralidad de los mismos.

Una vez que el contenido esté asociado a un desmentido, el chatbot envía automáticamente ese artículo a todos los usuarios que han preguntado por esa cuestión y, además, a partir de ese momento contestará automáticamente a todas las personas que planteen esa consulta nuevamente.

Además, la recolección de la frecuencia con la que nos llegan los avisos nos da pistas importantes sobre la evolución del bulo (cuándo nace, la intensidad con la que se está difundiendo, el momento en que empieza a decaer, etc..). Cada nuevo dato que recibimos sobre un bulo nos da la posibilidad de conocerlo

mejor y de darle una respuesta más eficaz.

Para entender bien la importancia del sistema de alertas a través del chatbot de WhatsApp, lo primero es conocer el proceso de trabajo de Maldita.es. El objetivo de la herramienta es resolver cuanto antes las dudas de los usuarios que nos consultan para aclararles si el contenido sobre el que dudan es un bulo o no, pero también que nos permita saber qué es lo que se está viralizando en cada momento en las conversaciones privadas de los usuarios a las que no podemos tener acceso porque están encriptadas de punto a punto y por lo tanto necesitamos que sean los propios usuarios los que nos las hagan llegar.

1. Lo primero es analizar cada recurrencia que nos llega. Son cientos de ellas cada día y hemos ido automatizando el proceso, aunque todavía hay una parte que se hace manualmente.

2. Si detectamos que la alerta se refiere a un bulo que ya tenemos identificado y desmentido, el chatbot automáticamente responde con nuestra verificación.

3. Si no lo hemos desmentido ni identificado se crea un acceso de contenido para ir midiendo su viralidad que, junto con su peligrosidad, serán las dos variables a tener en cuenta a la hora de decidir si se investiga o no. Tenemos recursos limitados así que centramos nuestros esfuerzos en los posibles bulos que se están haciendo muy virales y en los que consideramos que pueden ser un peligro real para los ciudadanos si se propagan.

04

Análisis: cómo puede el atributo "reenviado frecuentemente" ayudar a los verificadores

04

ANÁLISIS: CÓMO PUEDE EL ATRIBUTO "REENVIADO FRECUENTEMENTE" AYUDAR A LOS VERIFICADORES

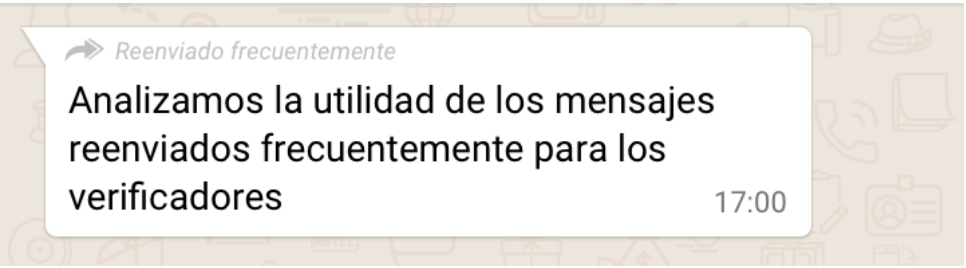
A partir del mes de marzo de 2021, los mensajes que llegan a nuestro chatbot de WhatsApp empezaron a contener una información adicional.

Dependiendo de la cantidad de veces que habían sido reenviados, llevaban un atributo diferente facilitado por la API de WhatsApp. Desde entonces, en nuestra base de datos hay registrados tres tipos de mensajes.

- ENVIADOS: son aquellos mensajes que sólo se han mandado una vez y no llevan ninguna marca especial.

- REENVIADOS: son aquellos que se han mandado más de una vez y menos de cinco veces. Aparecen marcados con una flecha.

- REENVIADOS FRECUENTEMENTE (a partir de ahora FF, de frequently forwarded): se trata de los mensajes que han tenido cinco reenvíos o más. Aparecen marcados con una doble flecha. Estos mensajes, según explica WhatsApp, no pueden seguir mandándose a varios grupos a la vez, pero pueden continuar circulando si se envían a los siguientes destinatarios de uno en uno. Este atributo también se conoce como 'Highly Forwarded Messages' (HFMs), mensaje altamente reenviado en castellano.



Aunque todas las organizaciones que tienen acceso a la API de WhatsApp tienen acceso al atributo FF, Maldita.es es el primer verificador en investigar su utilidad para la comunidad de fact-checking.

Este atributo otorga a los verificadores una información muy relevante sobre la viralidad de un contenido y, por tanto, nos sirven para mejorar nuestro trabajo a la hora de elegir en base a criterios objetivos qué contenidos verificar.

MALDITA. ES EL PRIMER VERIFICADOR A NIVEL MUNDIAL EN INVESTIGAR LA UTILIDAD DEL ATRIBUTO FF PARA EL TRABAJO DE FACT-CHECKING

Hemos analizado un mes del funcionamiento del chatbot entre el 5 de marzo a las 00:00h y el 4 de abril a las 23:59h y los datos prueban la importancia que tiene para el trabajo de Maldita.es. Durante ese mes redactamos y publicamos 207 artículos relacionados con desinformaciones y contenidos que requerían contexto: en 163 de ellos las alertas o dudas nos llegaron a través del chatbot. Es decir, en el 78,74% de los desmentidos había algún aviso que nos había llegado al chatbot de WhatsApp.

EN EL 78% DE LOS DESMENTIDOS EL CHATBOT DE WHATSAPP NOS HABÍA ALERTADO DE SU VIRALIDAD

En ese periodo de tiempo hemos recibido en nuestro chatbot un total de 8.014 alertas y 1.162 de ellas estaban marcadas con el atributo Frequently forwarded. Esto supone un 14,5% del total. Es importante señalar que nuestro chatbot no recibe mensajes de texto con más de 1.600 caracteres debido a las limitaciones de Twilio.

No a todas esas 8.014 alertas se les ha hecho un seguimiento porque los recursos del equipo de Maldita.es son limitados y porque además entre las alertas de texto o audio hay mucho "ruido" que debe ser descontado. Se seleccionan para ver su evolución los que a primera vista parecen falsos; los que tienen un alto índice de peligrosidad y los que pueden ser un tema disputado y han llegado de manera significativa a lo largo del tiempo. Como ya hemos explicado en la descripción del proceso que seguimos en Maldita.es, se realiza un filtrado automático y otro manual para tratar de darles significado a los mensajes que nos llegan y vincularlos a alguna de las investigaciones que hacemos para confirmar si se trata de desinformación.

De todas esas alertas, se les hizo un seguimiento para monitorizar su evolución con respecto a la viralidad tras considerar que podían contener desinformación a 2.089, es decir, el 26% del total. Si nos fijamos sólo en los contenidos asociados al atributo Frequently forwarded, de los 1162 contenidos que nos llegaron a lo largo de ese mes se les hizo seguimiento a 860: es decir, al 74%. Un

contenido asociado a este atributo es por lo tanto tres veces más probabilidad de ser monitorizado por nuestro equipo en base a su criterio periodístico y metodológico que uno que no lo tiene.

¿Todos los contenidos a los que les hace seguimiento Maldita.es terminan asociados a un artículo explicativo como bulo o desinformación? No. Para empezar porque tenemos recursos limitados y llegamos a donde llegamos. Además hay que tener en cuenta que algunas de las cosas que monitorizamos pueden no ser verificables de manera independiente o incluso ser reales.

En nuestro análisis hemos constatado que el 78,72% de los contenidos asociados al atributo Frequently forwarded a los que les hacemos seguimiento, es decir, alertas, terminan señalando bulos o desinformaciones.

EL 78% DE LAS ALERTAS CON FF TERMINARON SIENDO BULOS O DESINFORMACIONES

Nuestro análisis nos ha permitido certificar que no es lo mismo recibir una alerta normal que una asociada al atributo de Frequently Forwarded (FF). Estas últimas nos dan una idea más precisa del grado de peligrosidad y de viralidad de cada uno de los bulos sobre los que recibimos una alerta. Concretamente hemos visto que las alertas con FF ayudan a detectar dos tipos de fenómenos:

- Sirven de alerta temprana para detectar que un bulo va a tener previsiblemente una alta tasa de viralidad.
- Señalan claramente los que llamamos bulos zombies, los que los desinformadores utilizan periódicamente y van apareciendo y desapareciendo de nuestro sistema de alertas.

Además, hemos visto que los FF no sólo nos proporcionan pistas sobre cómo van evolucionando bulos concretos. El análisis también arroja conclusiones interesantes relacionadas con la temática de los desinformadores que necesitan ser investigadas en el futuro. Hemos detectado que hay asuntos sensibles donde se concentra gran cantidad de alertas con FF. Por ejemplo, las grandes controversias políticas o los temas relacionados con fraudes. También hemos encontrado una vinculación entre los mensajes FF y las campañas organizadas para criminalizar a colectivos vulnerables como los migrantes.

A continuación hacemos un análisis detallado de los Frequently Forwarded identificados por nuestro chatbot durante el primer mes y exploramos, con casos de estudio, las posibles utilidades que los Frequently Forwarded tienen para los verificadores:

1. FF Y FORMATOS: DATOS Y ANÁLISIS
2. FF Y CATEGORIZACIÓN DE DESINFORMACIÓN
3. FF: INDICATIVO DE ALTA TASA DE VIRALIDAD
4. FF Y BULOS ZOMBIES: LA RESURRECCIÓN ANUNCIADA
5. FF EN CASOS DE ESTUDIO

4.1 FF Y FORMATOS: DATOS Y ANÁLISIS

Nuestro chatbot tiene capacidad para recibir, recopilar e identificar alertas en cuatro formatos diferentes: texto, vídeo, imagen y audio. Durante el análisis de las alertas recibidas en el periodo de estudio hemos encontrado diferencias significativas entre los cuatro formatos asociados al atributo FF. Cambia mucho el número de alertas totales, cuántas son relevantes o qué cantidad de esas alertas está marcada con FF. Además, tenemos comprobado que lo habitual en Whatsapp es que cada bulo se mueva en un solo formato. De modo que desglosar los datos entre los cuatro formatos nos da información valiosa sobre cómo funciona la desinformación en WhatsApp.

Formato	Contenido total	FF%
Texto	5.532	9,9%
Imagen	1.651	16,5%
Video	736	42,4%
Audio	95	29,4%

Distribución de alertas con atributo FF por formatos

La primera gran diferencia que encontramos es el porcentaje de avisos con FF que nos llega en cada formato:

- En texto recibimos muchos inputs pero menos del 10% está marcado con el FF. Esto sucede porque la interacción de los usuarios con el chatbot va dejando rastro en la base de datos y genera un porcentaje de registros que no son relevantes.

- En imagen el porcentaje también es bajo pero en este caso el dato creemos que no se debe a ese "ruido" en la comunicación. Nuestra hipótesis para que el porcentaje de FF de imágenes sea muy bajo en comparación con el del vídeo o audio es que se debe a que muchas veces la descarga de imágenes automática a la galería en WhatsApp sí viene predeterminada y es más sencillo compartirla directamente desde el móvil que cuando se trata de vídeos o audios que se puede configurar que no se almacenen para no ocupar espacio.

Cuando nos fijamos sólo en los datos de los contenidos a los que les hemos hecho seguimiento también podemos sacar algunas conclusiones interesantes. Vemos que en texto el porcentaje de alertas vinculadas es muy bajo, el 12%, y el de audios con el 36% también es significativamente menor que en vídeo e imagen. Esto sucede porque en texto y en audio el chatbot recibe una gran cantidad de "ruido" al que no se le hace seguimiento: respuestas al chatbot,

saludos, preguntas mal formuladas o ininteligibles, audios personales planteando preguntas específicas...

Formato	Total alertas	Alertas con match	% vinculado
Texto	5.532	666	12%
Imagen	1.651	971	58,81%
Video	736	417	56,65%
Audio	95	35	36,84%
TOTAL	8.104	2.089	26%

Distribución de alertas con atributo FF por formatos

4.2 FF Y CATEGORIZACIÓN DE DESINFORMACIÓN

Cuando nos fijamos únicamente en los mensajes a los que les hemos hecho seguimiento que además tienen asociado el atributo FF los datos nos demuestran que se trata de alertas mucho más significativas: el 74% del total de avisos asociados con FF, independientemente del formato, coincide con un contenido identificado para hacerle seguimiento porque nuestro equipo cree que tiene altas probabilidades de ser falso. En audio y en texto el porcentaje de mensajes marcados con FF es muy alto: en texto roza el 85% y en audio llega al 75%. Además, si comparamos estos porcentajes con el de las alertas a las que les hemos hecho seguimiento y que, sin embargo, no tiene el atributo FF vemos que en todos los formatos la diferencia es muy grande. Es especialmente importante en los formatos de texto y audio pero incluso en el formato de vídeo no llega al 50%.

Formato	Alertas con FF vinculadas	Alertas sin FF vinculadas
Texto	84,54%	4,03%
Vídeo	67,62%	48,58%
Imagen	63,60%	17,69%
Audio	75%	20,89%
TOTAL	74,01%	9,7%

Comparativa entre alertas con FF vinculadas y alertas sin FF vinculadas

El atributo FF es una pista sólida para señalar posibles contenidos con desinformación. De acuerdo con los datos analizados este mes, hay una gran diferencia entre el porcentaje de alertas que analizamos como posibles bulos cuando llevan FF y cuando no lo llevan. Cuando tienen FF el 74,01% aparece vinculada a algún contenido que estamos analizando como posible desinformación. Cuando no lo llevan ese porcentaje cae hasta el 9,7%. Además, esta diferencia a la hora de marcar contenidos sospechosos se da en todos los formatos. Incluso en vídeo, que es el formato con un porcentaje más alto de alertas sin FF vinculadas, hay una diferencia de casi 20 puntos entre llevar FF y no llevarlo. Todos estos datos sostienen nuestra teoría de que el atributo FF es un indicador robusto de que el contenido que señalan tiene probabilidades de contener desinformación.

Hay otro dato que muestra la fuerte relación entre los FF y la desinformación. No todos los contenidos que investigamos terminan recibiendo la calificación de bulo. Hay algunos que comprobamos que son ciertos, otros son irrefutables y otros, siguiendo la metodología de Maldita.es, consideramos que no tenemos pruebas suficientes para decir que son desinformación. Con los datos del mes analizado vemos que hay 860 alertas vinculadas a alguna de estas investigaciones y que 677, el 78,72%, señalaban contenidos que Maldita.es acabó calificando de bulo o desinformación.

Formato	Alertas con FF que era desinformación
Texto	62,3%
Vídeo	22,2%
Imagen	13%
Audio	2,5%
TOTAL	**100%**

Porcentaje de alertas con FF que apuntaban a un bulo

También aquí hay importantes diferencias dependiendo de los formatos. De esos 677 avisos con FF que alertaban de un bulo confirmado por Maldita.es, el 62,3% llegó al chatbot como una cadena de texto. El segundo formato más habitual es el vídeo, con el 22,2% seguido de la imagen con el 13% y, finalmente, el audio que representa el 2,5% del total.

4. 3. FF: INDICATIVO DE ALTA TASA DE VIRALIDAD

En el análisis de este mes hemos comprobado cómo las alertas con FF señalan claramente los grandes focos de desinformación. Los contenidos sobre los que más avisos repetidos nos han llegado coinciden con contenidos asociados al atributo FF; es decir, el FF marca de manera clara los contenidos más virales.

EL ATRIBUTO FF MARCA EN EL 80% DE LOS CASOS LOS CONTENIDOS MÁS VIRALES

Para verificar esta afirmación en base a los datos hemos analizado dos tipos de contenidos:

- El Top 10 de los bulos con más alertas a lo largo del periodo analizado: los 10 bulos y desinformaciones sobre los que más nos han preguntado ese mes.
- El Top 10 de desinformaciones con más alertas asociadas al atributo FF en el periodo analizado

Los avisos con FF se concentran de manera muy intensa en los 10 bulos con más recurrencias en este mes. Un tercio de todos los avisos marcados con FF del mes aparecen en esta decena de bulos y desinformaciones. Concretamente 387 de los 1.162 que nos han llegado este mes se concentran en el Top 10.

Pero además la densidad de las alertas con FF en estos bulos también es muy alta. En total entre los 10 suman 803 avisos y la mitad de ellos son Frequently Forwarded. Ese porcentaje más que triplica el 14,5% de media que observamos en el total de la base de datos durante este mes.

De hecho, la lista del Top 10 de bulos sobre los que hemos recibido más alertas durante este mes es muy similar a la del Top 10 de bulos con más alertas con FF. Ocho de los bulos aparecen en las dos. Los dos que reciben muchos FF pero no tanta cantidad de alertas son en formato texto que ya hemos visto que favorece que aparezca una alta densidad de avisos con el atributo FF.

Estos datos muestran que los avisos con FF aparecen en mayor cantidad y en un porcentaje muy superior en los grandes bulos del mes, aquellos que han sido más virales y, por lo tanto, entrañan un mayor peligro.

Top 10 de los bulos con más alertas

Bulo	Alertas	Alertas FF	% Alertas FF	Formato
Volcán de Chillán en Chile	188	119	62,29%	Texto
Cese de Pablo Iglesias	119	29	24,36%	Imagen
Parlamento Dinamarca sobre Cataluña	111	54	48,64%	Vídeo
Begoña Gómez excluida Complutense	101	73	72,27%	Imagen y texto
Troyano: No puedo creer que seas tú	54	4	7,4%	Imagen
Parlamento Dinamarca monarquía	51	16	31,37%	Vídeo
Nueva regla de WhatsApp	51	49	96,07%	Texto
Empresa fantasma	45	22	48,88%	Imagen
Llamada del 626634795	43	32	74,41%	Imagen y texto
30% de los vacunados morirán	40	4	10%	Imagen
TOTAL	**803**	**402**	**50%**	

Top 10 de bulos con más alertas asociadas al atributo FF

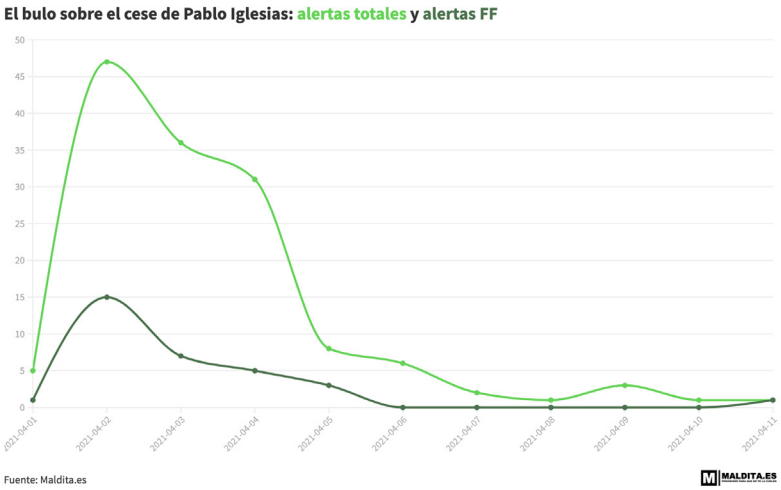
Bulo	Alertas	Alertas FF	% Alertas FF	Formato
Volcán de Chillán en Chile	188	119	62,29%	Texto
Begoña Gómez excluida Complutense	101	73	72,27%	Imagen y texto
Parlamento Dinamarca sobre Cataluña	111	54	48,64%	Vídeo
Nueva regla de WhatsApp	51	49	96,07%	Texto
Llamada del 626634795	43	32	74,41%	Imagen y texto
Cese de Pablo Iglesias	119	29	24,36%	Imagen
Rebaja de las pensiones	26	26	100%	Texto
Empresa fantasma	45	22	48,88%	Imagen
Despedidos Príncipe Felipe	34	19	55,88%	Texto
Parlamento Dinamarca monarquía	51	16	31,37%	Vídeo
TOTAL	**769**	**439**	**57,08%**	

El bulo sobre el cese del vicepresidente Pablo Iglesias

Este bulo es uno de los casos de estudio que ejemplifican cómo el atributo FF antecede a una gran viralización. Se trata de un bulo en formato de imagen que es la desinformación desmentida de nueva creación sobre la que más alertas recibimos en el periodo analizado. En solo 10 días recibimos 146 avisos en total. En el mensaje acusan falsamente al vicepresidente del Gobierno de cobrar una indemnización muy elevada al dejar su cargo y además haciendo una triquiñuela legal, según la desinformación. Para "demostrarlo" adjuntan una publicación del Boletín Oficial del Estado del día 31 de marzo, en el que dimite oficialmente el vicepresidente, sólo 15 horas antes de que la primera alerta llegue a nuestro chatbot.

Imagen del bulo sobre el cese del vicepresidente Pablo Iglesias

La primera alerta que nos llega el 1 de abril a las 15:36 y esa misma noche ya recibimos la primera advertencia con FF, que se repite a primera hora de la mañana del día 2. Dos FF seguidos poco después de aparecer en nuestro chatbot ya nos dan una idea de que se trata de un bulo potencialmente muy viral. Durante el día 2 de abril nos alertan 47 veces de que ese bulo estaba circulando por WhatsApp. En la tarde de ese día crece la densidad de avisos con FF. Seis de las siete alertas que recibimos entre las 17:01 y las 21:13 son con FF. La difusión durante los dos días siguientes es todavía muy intensa y empieza a decaer a partir del día 5 abril.



El bulo sobre el cese de Pablo Iglesias. Alertas totales vs alertas FF

Hay que tener en cuenta que se trata de un bulo en formato imagen, en el que detectamos menor porcentaje de avisos con FF, seguramente porque el proceso para reenviar las imágenes hace que se pierdan algunas de esas marcas. Aún así vemos una cantidad muy considerable de FF que, además, marcan de una manera muy nítida que era un bulo que estaba logrando un alto grado de viralidad. La concentración de avisos con FF en la tarde del día 2 de abril son un signo claro de que es un bulo muy difundido y ante el que contra el que es importante reaccionar rápido.

4. 4. FF Y BULOS ZOMBIES: LA RESURRECCIÓN ANUNCIADA

Al igual que un zombi es un ser resucitado, hay bulos que vuelven cada cierto tiempo. Son desinformaciones que, pese a haber sido desmentidas en el pasado, reaparecen después de unos meses.

En base a nuestro análisis hemos detectado que los bulos zombies a menudo vuelven asociados al atributo Frequently Forwarded, y eso es útil para los verificadores ya que aquellos contenidos que vuelvan en el tiempo de manera recurrente es altamente probable que lo vayan a seguir haciendo, por lo que es conveniente trabajar en un desmentido.

Este tipo de bulos llevan en nuestra base de datos desde hace años pero no acaban de desaparecer y cada cierto tiempo se vuelven a poner en circulación. En este estudio nos hemos fijado en aquellos los detectamos por primera vez hace meses, acumulan más de 100 recurrencias en nuestra base de datos y nos volvieron a avisar sobre ellos en el mes estudiado.

En total, hay 14 de esos grandes bulos zombies que cumplen estas tres condiciones. Entre todos suman 503 recurrencias de las que 298, casi el 60%, está marcada con FF. Este porcentaje nos indica que los contenidos antiguos, los que llevan tiempo circulando por WhatsApp, es fácil que acumulen reenvíos y nos lleguen con la marca del FF. En la tabla vemos que cuando hay una cantidad alta de alertas, es decir, cuando el bulo está volviendo a circular con fuerza, el porcentaje de avisos con FF tiende a ser alto. En la parte baja de la lista aparecen otros bulos que hemos visto difundirse con fuerza pero que en el mes que hemos estudiado permanecen en estado latente. Un estudio más profundo del comportamiento de los bulos zombies a lo largo del tiempo podría confirmar si existe una relación estrecha entre las alertas con FF y la reaparición de bulos zombies.

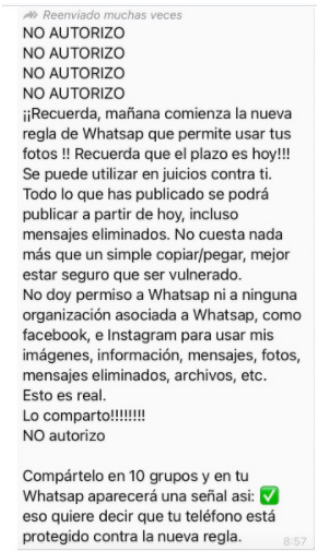
Bulos zombies con más alertas en este mes

Bulo	Alertas totales	Alertas este mes	Alertas este mes con FF	% Alertas con FF este mes
Volcán de Chillán en Chile	307	188	119	62,29%
Troyano: No puedo creer que seas tú	337	54	4	7,4%
Nueva regla de WhatsApp	413	51	49	96,07%
Empresa fantasma	352	45	22	48,88%
Llamada del 626634795	151	43	32	74,41%
Correo de la Dian	232	28	16	57,14%
Laboratorio de Wuhan	234	27	26	96,29%
Rebaja de las pensiones	155	26	26	100%
Parlamento Europeo comunismo	178	18	1	5,55%
"Manitas limpias"	969	14	0	0
Coches oficiales desguace	128	4	2	50%
Alimentos alcalinos y coronavirus	145	3	1	33,33%
Envío de dinero Ria	187	1	0	0
Esponjas para robar	166	1	0	0
TOTAL		**503**	**298**	59,24%

Los patrones que hemos detectado en la llegada de alertas con FF nos dan pistas muy valiosas para localizar la desinformación más grave que está circulando en cada momento por WhatsApp y decidir así, en base a los datos, a qué vamos a dedicar nuestros recursos limitados a la hora de verificar. Ya hemos comprobado cómo este tipo de alertas se concentra en los bulos más virales o que tienen capacidad de reaparecer en el tiempo y seguir desinformando. Vamos a analizar ejemplos concretos que muestran cómo las FF son un indicador de que un bulo se está empezando a viralizar y que ayuda a detectar rápidamente cuándo una desinformación que ya ha tenido mucho recorrido en el pasado está de vuelta.

La cadena de WhatsApp sobre privacidad que siempre vuelve

Uno de los bulos zombies más destacados que hemos visto en el periodo analizado nos llegó en formato de texto y habla de una supuesta nueva norma de WhatsApp que permitiría a la plataforma utilizar las fotos que enviemos y que todo el contenido compartido se haga público.

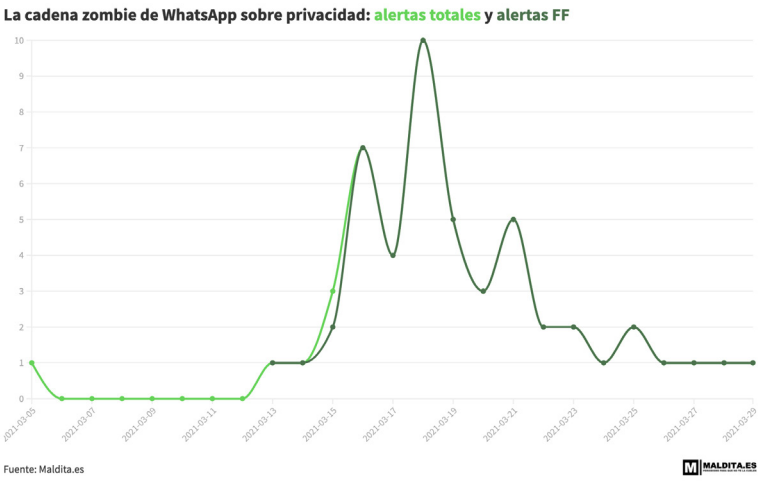

Reenviado muchas veces
NO AUTORIZO
NO AUTORIZO
NO AUTORIZO
NO AUTORIZO
¡¡Recuerda, mañana comienza la nueva regla de Whatsap que permite usar tus fotos !! Recuerda que el plazo es hoy!!! Se puede utilizar en juicios contra ti. Todo lo que has publicado se podrá publicar a partir de hoy, incluso mensajes 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

La falsa nueva regla de WhatsApp, ejemplo de bulo zombie

Sobre este bulo hay 425 avisos registrados en nuestra base de datos. Aparece por primera vez el 14 de enero de 2021 a las 20:16 y su desarrollo es explosivo. Sólo el día 15 de enero nos llegan 301 avisos sobre este bulo a nuestro chatbot, más de 1 cada 5 minutos de media. Entonces no teníamos activado el sistema de alertas con FF y por lo tanto desconocemos si lo eran. A partir del día 16 de enero empieza a bajar el ritmo y se nota que su difusión va decayendo hasta el fin de enero. En febrero apenas aparece pero vuelve con fuerza en marzo, sobre todo a partir del día 13. Entonces ya teníamos activadas las alertas con FF y se ve de una manera muy nítida cómo este tipo de cadenas de mensajes de texto quedan marcadas con esta herramienta.

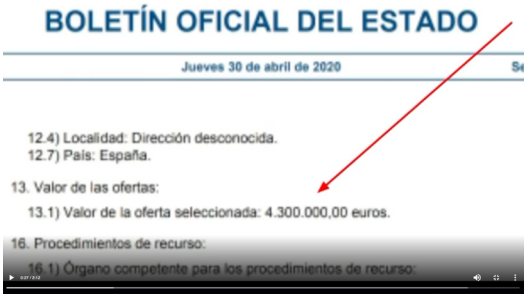
De las 51 alertas que recibimos sobre este bulo en el periodo de estudio, 49 llevan la marca de la FF, el 96%.



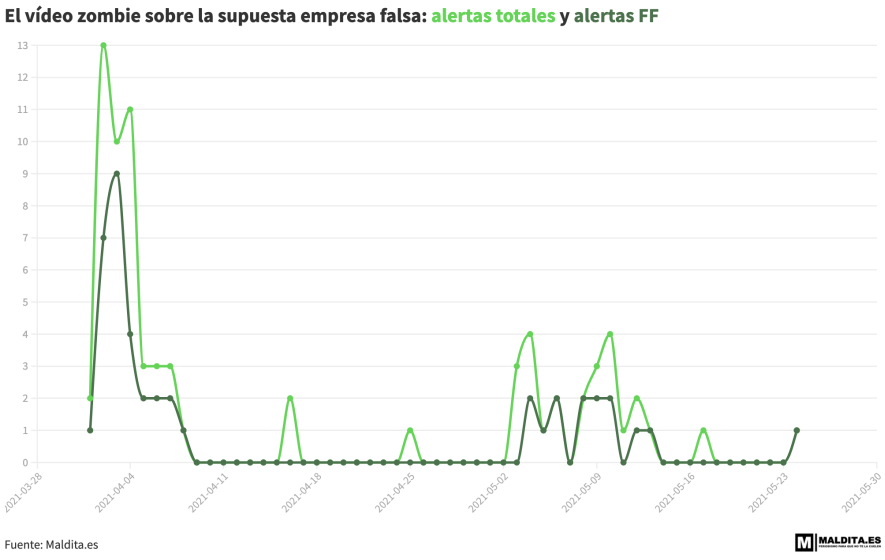
La cadena zombie de WhatsApp sobre privacidad. Alertas totales vs alertas FF

El vídeo zombie sobre la supuesta empresa fantasma

Éste es un caso de estudio de un contenido zombie en formato de vídeo que necesita contexto. En él se acusa al gobierno español de hacer negocios con una empresa fantasma para comprar material para detectar la COVID-19. En la base de datos de Maldita.es aparecen 380 avisos a lo largo del tiempo.



El primero nos llegó el 3 de mayo de 2020 cuando todavía no teníamos activado nuestro chatbot y hacíamos una recolección manual de los datos: sólo ese día recibimos 168 alertas en nuestro servicio de WhatsApp. El día 4, aunque baja la intensidad a 70 avisos, sigue muy activo y a partir de entonces decae. Los desinformadores suelen tratar de aprovechar al máximo los bulos que tienen éxito y consiguen una difusión tan intensa como éste. Nuestra hipótesis es que ésta es la razón por la que el 1 de abril, después de mes y medio sin recibir ninguna alerta sobre este bulo, volvemos a verlo además con un grado importante de viralidad. En tres días y medio recibimos 36 alertas y casi la mitad con FF. Además vemos que las alertas con FF nos empiezan a llegar muy pronto. 3 de las 6 primeras llevan la marca. Esto es un síntoma de que no se trata de un contenido nuevo, si no de un vídeo que ya había estado circulando previamente por WhatsApp y, por lo tanto, es más probable que acumule reenvíos, pero también nos puede alertar de que se trata de un contenido recurrente que conviene tener verificado.



La cadena zombie de WhatsApp sobre privacidad. Alertas totales vs alertas FF

4. 5. FF SEGÚN TEMÁTICA: CASOS DE ESTUDIO

En el análisis detallado de los bulos de este mes hay una serie de casos de estudio que plantean hipótesis para futuras investigaciones en relación a alta presencia del atributo FF y unos patrones específicos de aparición de estas alertas.

Bulos políticos: duelo en el Parlamento

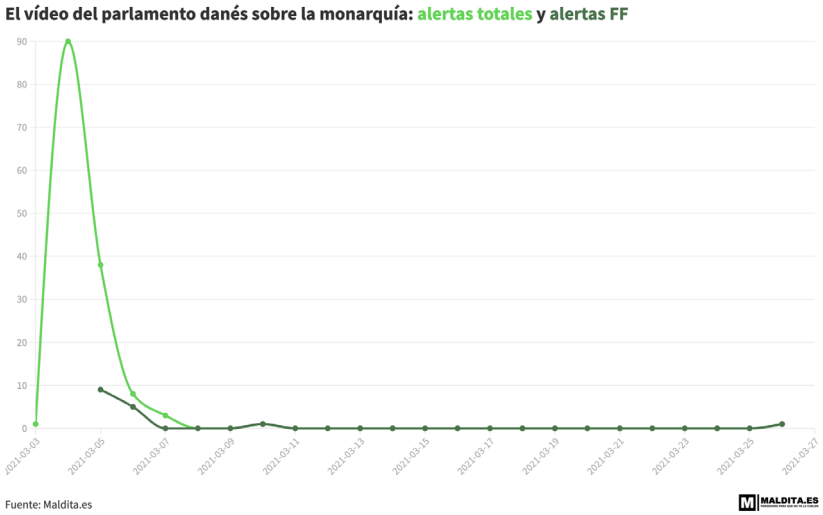
Una de las motivaciones más importantes que identificamos en base a nuestra experiencia a la hora de desmontar desinformación tiene que ver con razones políticas. Son bulos que buscan modificar la percepción de una realidad para favorecer una determinada ideología: gran parte de los bulos que detectamos tienen un trasfondo ideológico.

En el periodo de estudio hemos podido ver un duelo de desinformación entre dos tendencias políticas que manipularon el mismo vídeo pero con un contenido diferente. El vídeo que utilizaron era de una sesión del Parlamento de Dinamarca en la que los diputados empezaban a reír durante la alocución de la Primera Ministra sobre la compra de los elefantes de un circo por parte del gobierno con el fin de darles una vida digna. Como hablan en danés, un idioma poco conocido, los desinformadores manipulaban los subtítulos con textos que no tenían nada que ver con lo que se decía realmente, una práctica habitual en la desinformación. Primero se difundió un vídeo en el que se simulaba que los diputados se reían de la monarquía española. A eso se respondió desde el otro lado del espectro político con otra versión en la que lo que se simulaba es que las risas eran por el proceso independentista catalán.

Este duelo de vídeos satíricos con contenido político son los más destacados en este formato en el periodo que hemos investigado. De hecho entre los dos suman 162 avisos, un 37% del total de las alertas en formato vídeo vinculadas a algún contenido. De esos avisos, 72 poseen el atributo FF: el 23% de los 312 FF en formato vídeo que hemos recibido durante este mes.

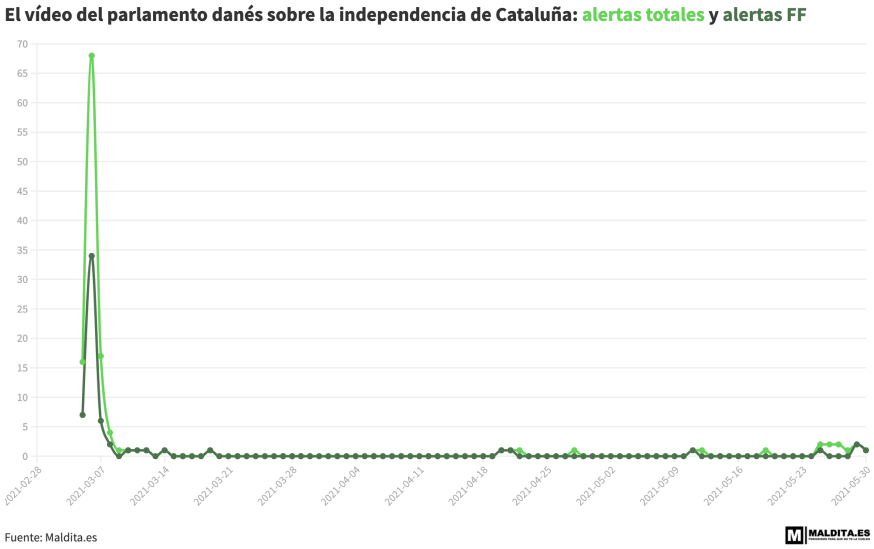
El mismo vídeo se usa contra la monarquía (izquierda) y contra el independentismo catalán (derecha)

El bulo contra la monarquía comienza el 3 de marzo, y se mueve muchísimo los días 4 y 5. La mañana de ese 5 de marzo se activó el sistema de alerta sobre los Frequently Forwarded en nuestro chatbot y muy poco después nos llega ya la primera alerta con FF sobre este bulo, a las 13:24. Cuando empiezan a registrar los FF este bulo ya está en su fase de caída. Sigue apareciendo, aunque cada vez con menos intensidad, hasta el día 7 de marzo. Aún así, en ese tiempo recibimos 16 alertas con FF. El número no es muy alto pero si nos fijamos en la estructura de la línea de avisos completa de este bulo lo lógico es pensar que en los dos días anteriores a la activación del sistema también había una cantidad importante de FF que todavía no se registraban así en nuestra base de datos.



El vídeo del parlamento danés sobre la monarquía. Alertas totales vs alertas FF

Cuando el bulo contra la monarquía empezaba a decaer se puso en marcha la respuesta, el montaje que satirizaba al independentismo catalán. Entra en escena el día 5 de marzo por la tarde, recibimos el primer aviso a las 19:12, y llega con mucha fuerza. En este caso, como el sistema de FF ya estaba activo, podemos ver la evolución completa del bulo. El quinto aviso que recibimos sobre este bulo ya es un FF y a partir de entonces prácticamente la mitad de las alertas que nos llegan están marcadas. En 72 horas recibimos 99 avisos, 49 con FF, pero la mayoría de estos avisos se concentran el día 6 de marzo. Sólo ese día recibimos 68 avisos, una media en el día de un aviso cada 21 minutos. Hay momentos de especial intensidad. El mismo día 6, entre las 20:16 y las 21: 26, recibimos 9 avisos, todos con FF, un indicativo claro de que en esos momentos el vídeo circulaba por WhatsApp a toda velocidad.



El vídeo del parlamento danés sobre la independencia de Cataluña. Alertas totales vs alertas FF

Los dos casos tienen un dibujo similar, el de bulos intensamente virales. Consiguen una difusión muy fuerte poco después de nacer y eso hace que se disparen las consultas sobre ese bulo en las horas siguientes. Después de un pico en el que casi monopolizan el chatbot, empiezan a decaer hasta que desaparecen. Esta estructura es habitual en los grandes bulos políticos. Siguen la actualidad y cuentan con muchos actores dispuestos a difundirlos por redes sociales. Cuando alguna noticia les da pie a poner en circulación el bulo, toda la maquinaria que apoya el grupo ideológico beneficiado por el bulo trata de moverlo lo máximo posible. Por eso es habitual que nuestro chatbot reciba una avalancha de consultas en poco tiempo sobre ese bulo.

Las alertas del bulo sobre la monarquía y el bulo sobre Cataluña en el tiempo

Falsas alertas de hackeo

Hay un tipo de desinformaciones que no están ligadas a la actualidad, pero vuelven periódicamente y lo hacen con mucha fuerza. Son las cadenas de texto que alertan sobre supuestos intentos de hackeo. Consiguen mucha viralidad aprovechándose en ocasiones de las buenas intenciones de los usuarios que reenvían los mensajes para alertar a sus familiares y amigos del supuesto peligro que se denuncia en el mensaje.

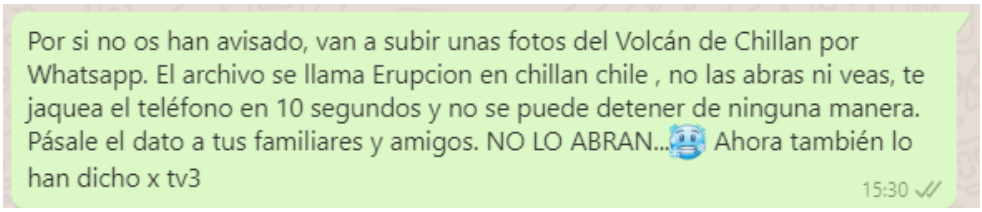
En estos casos vemos que se viraliza durante años el mismo texto. Los registros que nos llegan al chatbot muestran avalanchas de avisos sobre estos bulos que duran pocos días pero que adquieren mucha intensidad. Estos mensajes concentran muchos FF por varias razones:

- Las cadenas de texto es fácil que acumulen reenvíos por la facilidad para hacerlos y de copiar y pegar textos largos.
- Por la alta intensidad con la que circulan.
- Porque son mensajes antiguos que vuelven a la vida y esa característica también está relacionada con un alto número de FF.

En el mes que hemos analizado tenemos varios ejemplos de esto:

El volcán Chillán de Chile

En nuestra base de datos figuran 321 alertas en total sobre un supuesto hackeo del móvil que se activa al abrir una foto de un volcán chileno llamado Chillán que circula por WhatsApp: es el contenido con un mayor número de alertas en el periodo analizado aunque es, además, un bulo zombie.



Cadena de Whatsapp sobre el virus del volcán Chillán de Chile

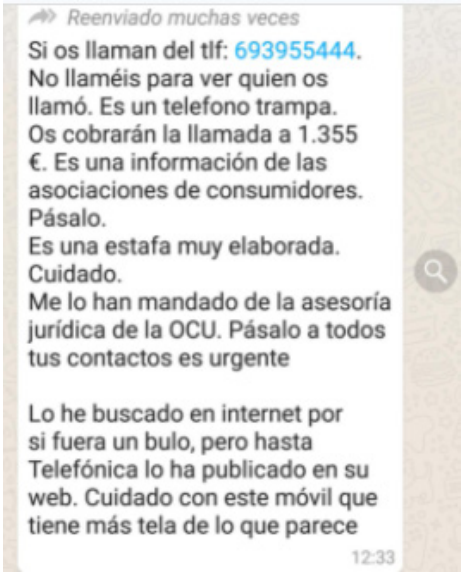
La primera que nos llegó fue el 9 de octubre de 2019. Se movió mucho entre octubre y noviembre de ese año, apareció con menos intensidad hasta febrero de 2020 y entonces desapareció, pero los desinformadores lo volvieron a poner en circulación el 6 de marzo de 2021. Volvió con mucha fuerza, sobre todo entre el 8 y el 11 de marzo. En esos 4 días este bulo prácticamente monopolizó las alertas en formato texto que recibimos en nuestro chatbot. De 186 que nos llegaron, 163 eran sobre el volcán chileno. Como era previsible, el porcentaje de FF también era muy alto. El 65% de las alertas en esos cuatro días llevaban FF.

La Cadena de Whatsapp sobre el virus del Volcán Chillán. Alertas totales vs alertas FF

Este bulo es el número uno del mes, es la desinformación sobre la que nuestra comunidad nos ha alertado más veces en ese periodo. En total son 188 avisos que se concentraron en sólo 12 días, 119 de ellos con FF.

Llamadas muy caras

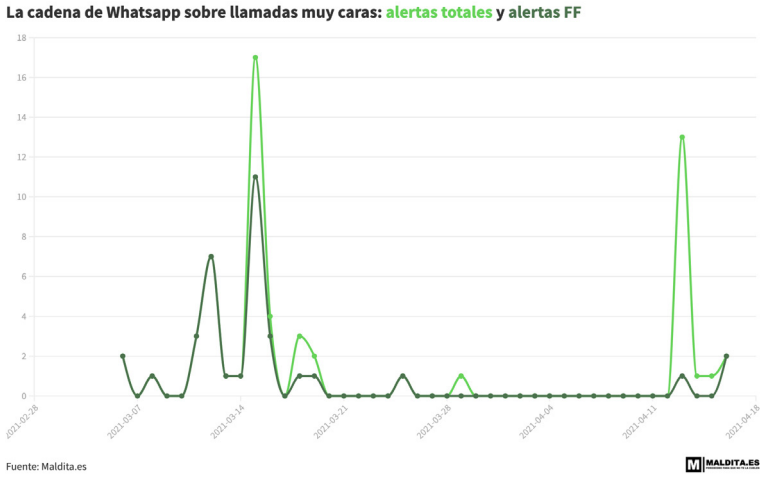
Vemos un comportamiento muy parecido en este otro bulo con una temática muy similar. Aquí los desinformadores muestran dos números de teléfono y alertan de te van a cobrar una elevada cantidad de dinero si respondes a sus llamadas Es falso y trata de hacerse viral incitando a los usuarios a compartirlo con sus contactos para evitar que caigan en un supuesto timo.



El bulo de la llamada figura en la base de datos de Maldita.es desde 2019

En nuestra base de datos hay 161 avisos sobre este bulo. Lo vimos por primera vez en septiembre de 2019 y desaparece en enero de 2020. Después de un año sin que tengamos noticias de él, lo volvemos a ver en enero de 2021 cuando recibimos 5 avisos, pero es en marzo, dentro del mes que hemos analizado, cuando vuelve a hacerse viral. Las 19 primeras alertas que recibimos en el mes de marzo son todas con FF. Es decir, en cuanto tuvimos activo el sistema de alertas con FF quedó claro que este bulo era altamente viral. Además, este

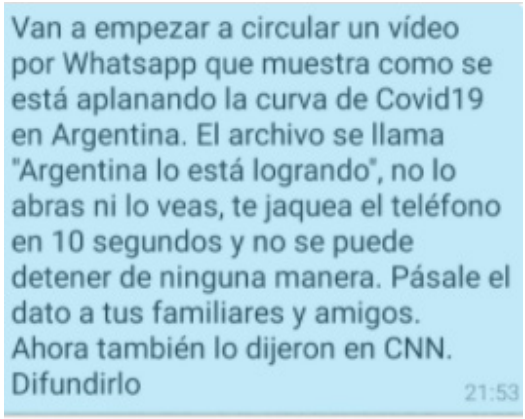
bulo tiene otra característica peculiar. A partir del 14 de marzo detectamos que los desinformadores están intentando ampliar el público al que hacer llegar este bulo y nos empiezan a avisar de que también está circulando una versión traducida al catalán. En total, durante el mes recibimos en el chatbot 43 avisos sobre este bulo, 32 con FF, es decir el 74%.



La Cadena de Whatsapp sobre llamadas muy caras. Alertas totales vs alertas FF

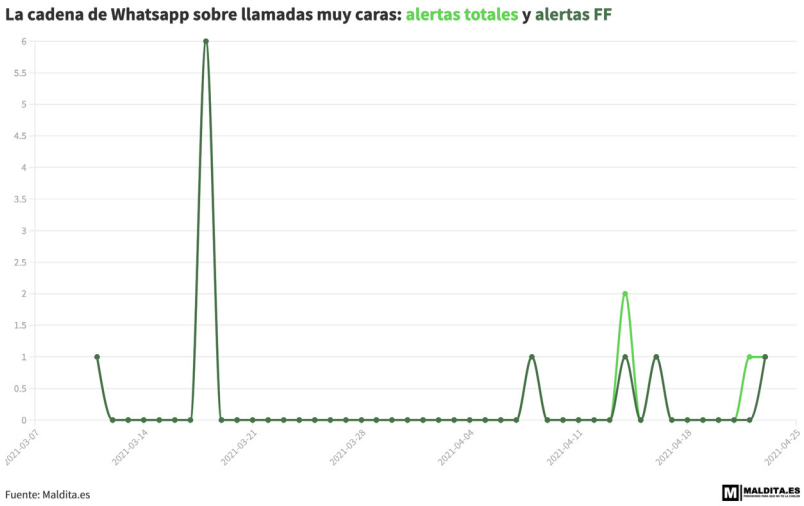
El bulo viajero

Es otro ejemplo de cadena de texto que circula por WhatsApp y que atrae un porcentaje muy alto de avisos con FF. En este caso también alerta de un supuesto hackeo del móvil a través de un vídeo titulado "Argentina lo está logrando". Es un bulo viejo que detectamos por primera vez en abril de 2020 y lo volvemos a ver en el mes de 2021 que hemos estado estudiando: marzo.



Ejemplo de bulo detectado en Latinoamérica y en España

Pero merece la pena comentarlo en este informe porque tiene una característica que aumenta sus probabilidades de atraer FF. Se trata de un bulo que también ha circulado por Latinoamérica. Por ejemplo lo ha desmentido también la organización de verificación argentina Chequeado o la boliviana Bolivia Verifica. Al tratarse de un mensaje que se mueve de un país a otro, aumentan las posibilidades de acumular reenvíos. En nuestra base de datos, durante el periodo de estudios, nos llegaron 7 avisos, 6 de ellos con FF.



La Cadena de Whatsapp sobre el virus "Argentina lo está logrando". Alertas totales vs alertas FF

Posible campaña de desinformación contra migrantes

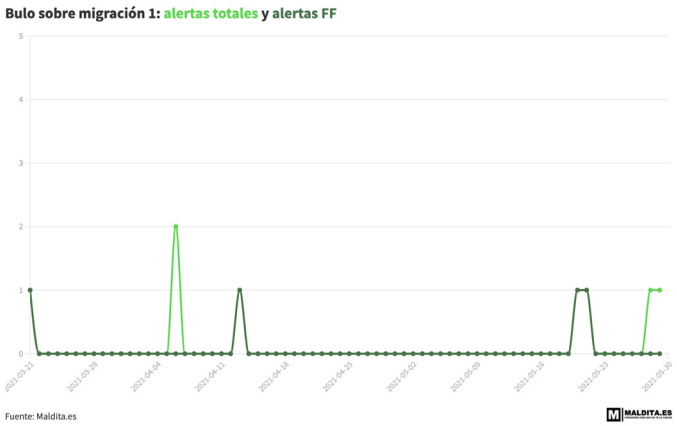
En el análisis de los Frequently Forwarded también hemos detectado indicios que podrían apuntar a acciones coordinadas de desinformación. En el mes que hemos estudiado descubrimos que entre el 21 y el 22 de marzo nos habían avisado de cuatro desinformaciones diferentes que trataban de criminalizar a los inmigrantes. Se trataba de cuatro bulos antiguos sobre los que nos avisaron en un corto espacio de tiempo y que, aunque no tuvieron mucho seguimiento, nos permiten plantear la hipótesis de que se estaba tratando de poner en marcha una campaña de desinformación contra la población inmigrante.

Los bulos a los que nos referimos los detectamos en 24 horas, entre el 21 de marzo a las 22:08 y el 22 de marzo a las 22:23, y se trata de cuatro vídeos. Todos atacan a los inmigrantes y tres de ellos les acusan de recibir cuantiosas ayudas del estado de manera irregular. Los cuatro están en nuestra base de datos desde hace meses y los avisos con FF marcan que han tenido un recorrido anterior porque en el periodo que estudiamos no consiguen hacerse virales según nuestros registros. Los vemos uno a uno.

Bulo sobre migración 1: La mujer que rechaza un trabajo para seguir cobrando ayudas

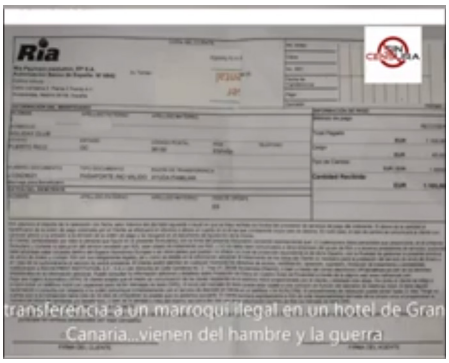
El primero de los bulos asegura que una mujer musulmana rechaza un trabajo para poder seguir cobrando ayudas. Nos denunciaron esta mentira por primera vez en marzo de 2019. Sigue apareciendo esporádicamente durante 2019 y 2020. El 21 de marzo a las 22:08 lo detectamos otra vez. Era la primera vez que

nos llegaba con el sistema de FF activo y la alerta que nos llega está marcada con Frequently Forwarded. En abril y mayo lo hemos visto otras 4 veces, 2 de ellas con FF.

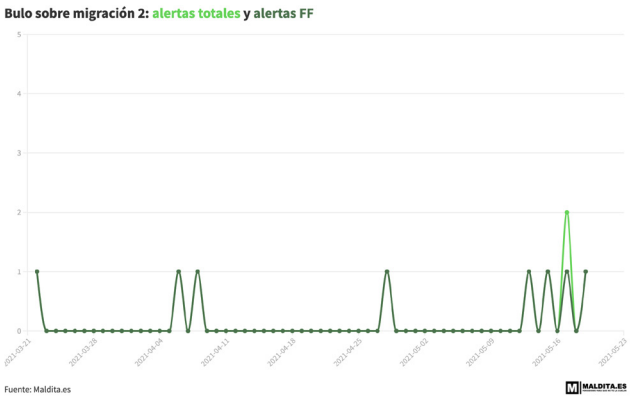


Bulo sobre migración 1. Alertas totales vs alertas FF

Bulo sobre migración 2: el irregular que cobra ayudas

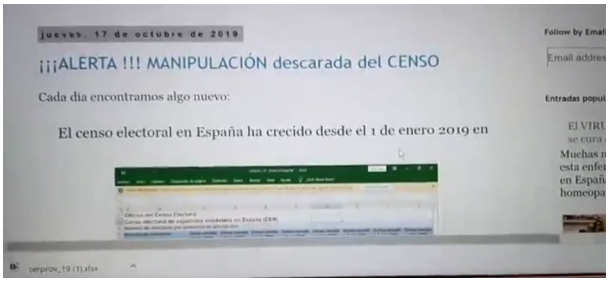


Otro bulo que reaparece en esos días habla de un inmigrante indocumentado al que acusan de cobrar ayudas irregulares. Esta desinformación la vimos mucho en diciembre de 2020 y sigue apareciendo con regularidad hasta el 22 de febrero. Luego desaparece durante un mes y vuelve el 22 de marzo a la 1:23. Fuera del periodo de estudio, más allá del 4 de abril, encontramos pruebas de que es un vídeo que utilizan habitualmente los desinformadores que atacan a los inmigrantes. Entre abril y mayo nos han alertado sobre este bulo 8 veces más, 5 de ellas marcadas con FF.

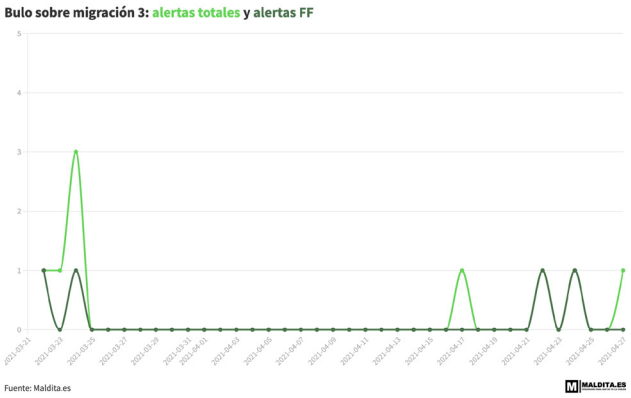


Bulo sobre migración 2. Alertas totales vs alertas FF

Bulo sobre migración 3: el censo electoral inflado



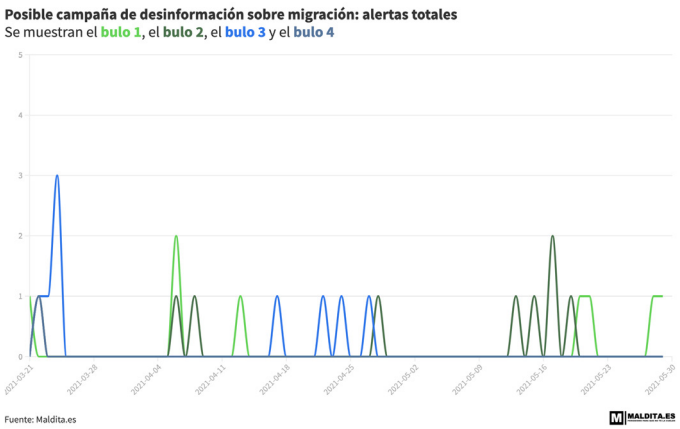
El tercer bulo en vídeo contra los inmigrantes llega el 22 de marzo a las 20:47 y se refiere a una falsa manipulación del censo electoral para dejar votar a 200.000 inmigrantes irregulares. Este también apareció en 2019. Lo volvemos a ver de manera intermitente los meses siguientes. En 2021 nos alertaron sobre este bulo el 14 de febrero y estuvo un mes desaparecido hasta el 22 de marzo, cuando nos llegó una alerta marcada con FF. Desde entonces nos han llegado ocho alertas más sobre este bulo, 3 de ellas con FF.



Bulo sobre migración 3. Alertas totales vs alertas FF

Bulo sobre migración 4: el traficante irregular que cobra ayudas públicas

El último vídeo también intenta relacionar a un inmigrante que supuestamente se dedica al tráfico de drogas con ayudas irregulares del Estado. Nos alertan sobre él a las 22:23 del día 22 de marzo. Este otro bulo que llevaba un mes sin aparecer aunque entre el 31 de enero y 9 de febrero de 2021 habíamos recibido 20 avisos sobre esta desinformación.



Estos cuatro bulos manifiestan un intento, marcado con alertas asociadas al FF, de resurgir todos a la vez en 24 horas, lo que nos hace pensar que podría tratarse de una campaña orquestada y que por lo tanto en el futuro debemos investigar si este atributo puede ayudarnos a identificar campañas de desinformación organizadas.

Hay otro dato que apoya la sospecha de que estos cuatro vídeos formaban parte de un intento de lanzar una campaña. A partir del 18 de mayo de 2021 detectamos una oleada de desinformación contra los inmigrantes a raíz de la entrada irregular de varios miles de personas procedentes de Marruecos en la ciudad española de Ceuta. En Maldita.es detectamos dos decenas de bulos creados a partir de ese día. En ese contexto, en los días 19 y 20 de mayo nos volvieron a alertar sobre dos de esos vídeos. Concretamente recibimos avisos sobre el bulo de la mujer musulmana que rechaza un trabajo y el del inmigrante negro al que acusan de cobrar ayudas ilegales.

Pese a todos estos elementos, el periodo que hemos analizado es corto y la cantidad de ejemplos que hemos detectado es pequeña. Sería necesario un estudio más profundo de la base de datos de Maldita.es para confirmar si estos patrones que hemos visto a pequeña escala se repiten de forma habitual y se confirma la hipótesis de que los FF pueden ser un indicio sólido de una acción coordinada en un tema específico como es la inmigración.

06

Conclusiones

06

CONCLUSIONES

En países en los que el consumo de información a través de WhatsApp es importante, un servicio de Whatsapp como el chatbot de Maldita.es es una de las principales herramientas de trabajo para un verificador a la hora de detectar posibles bulos: es una mirilla a la desinformación que está afectando a las conversaciones que tiene la ciudadanía en entornos cerrados. Gracias a la colaboración de nuestra comunidad, podemos saber qué está pasando dentro de WhatsApp a pesar de ser una red encriptada de punta a punta según sus creadores: sus denuncias nos permiten saber qué es lo que se está viralizando en cada momento. El 78% de los contenidos desmentidos por Maldita.es en el periodo analizado habían llegado a través de WhatsApp.

Para que un servicio de WhatsApp tenga utilidad es necesario que el verificador tenga una comunidad creada dispuesta a colaborar denunciando posibles desinformaciones y viralizando en sus propias conversaciones los desmentidos. Un chatbot automatizado permite que los usuarios de esa comunidad puedan escalar sin colapsar ni el sistema ni al equipo de periodistas encargados de la verificación.

Además, el atributo FF asociado a los contenidos puede ayudar a los verificadores a identificar y priorizar los contenidos que están siendo más virales y por lo tanto están teniendo más impacto. Nuestro análisis ha identificado que en el 80% de los casos los contenidos más virales son también los que tiene más alertas con FF asociadas. Además, cuando los verificadores aplican la metodología eligiendo qué contenido investigar según viralidad y peligrosidad, las alertas asociadas al FF tienen tres veces más probabilidades de ser identificadas como posible desinformación por los periodistas que las que no lo están.

Por último, es importante señalar que en el mes analizado el 78% de las alertas con FF investigadas terminaron siendo bulos o desinformaciones. Por lo tanto el atributo FF señala los contenidos más virales a los que además habitualmente los verificadores les harían seguimiento según su metodología y que terminan siendo calificados como desinformación por el trabajo periodístico. Este atributo puede ser, por lo tanto, de utilidad para los verificadores a la hora de escalar su trabajo y decidir qué investigar primero en una situación de recursos limitados.

En definitiva el chatbot de WhatsApp y las alertas con FF nos ayudan a articular un sistema de alertas tempranas que nos puede permitir reaccionar rápido para detener la viralidad de un bulo que se empieza a difundir, neutralizar nada más reaparecer bulos que vuelven periódicamente y detectar muy pronto campañas de desinformación organizadas con motivaciones específicas, si bien esto necesita un estudio más profundo.