Comportamiento malicioso en la Web

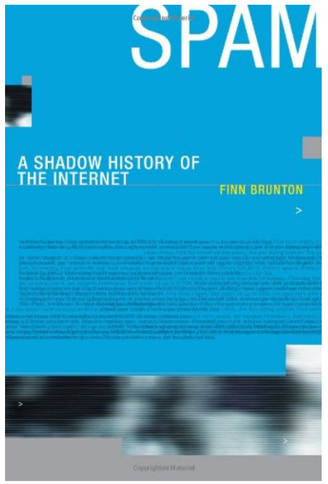
Daniel Gayo Avello

Última modificación: Mon, 15 May 2023 09:36:52 GMT

Tabla de contenidos

La Web es un entorno adversarial

- Ya hemos señalado en varias ocasiones que la Web es un entorno **adversarial**:
  - Por un lado hay una **competición** entre los distintos sitios web por la **atención** de los usuarios. Competición que en ocasiones lleva a algunos sitios a recurrir a subterfugios.
  - Por otro existen **agentes maliciosos** que buscan explotar el ecosistema de la Web (sitios, usuarios, buscadores, etc.) con objetivos muy diferentes pero raramente benignos.
- A lo largo de las próximas sesiones exploraremos dos ideas principales:
  - Los usuarios maliciosos: **troles, bots, títeres** y vándalos.
  - **La desinformación y la información errónea**.
- En principio no cubriremos el ***spam***, quien tenga interés puede consultar el libro *"Spam: A Shadow History of the Internet"*.
- Por otro lado, debido a su gran importancia en la sociedad moderna la mayor parte de ejemplos se realizarán en el contexto de los **medios sociales**.



¿Qué son los medios sociales?

- Hay casi tantas **definiciones** de medios sociales como investigadores en el área...
- La primera definición que se hizo en Wikipedia (en 2006) fue la siguiente:

> *Social Media is the term used to describe media which are formed mainly by the public as a group, in a social way, rather than media produced by journalists, editors and media conglomerates.*

- Una definición más reciente (2015)...

> *Social media represent a set of communication practices that can typically be described as 'many-to-many.' In contrast to broadcast media, consumers are typically also producers. In contrast to in-person communication, audiences are often ambiguous or underspecified.*

¿Qué son los medios sociales?

- En ambos casos se menciona el concepto de **contenido generado por usuarios** pero, aparte de eso, ambas definiciones son bastante diferentes. Además, parecen incompletas sin añadir una definición ostensiva...
- Así, en la segunda versión de la definición wikipédica se incluían como ejemplos:

> *blogs, Podcasts, Vlogs, Wikis, MySpace, YouTube, Second Life, Digg, Memeorandum, del.icio.us, Reddit, Flickr, Tailrank and Newsvine.*

- La definición de Hogan y Melville menciona

> *collaborative encyclopedias such as Wikipedia, social network sites (SNSs) Facebook and Twitter, photo-sharing sites Instagram, and social news site Reddit.*

- El problema con las definiciones ostensivas es que envejecen mal. En este momento Second Life y MySpace están virtualmente muertos, del.icio.us ha desaparecido y quién se acuerda de Memeorandum, Tailrank o Newsvine. En un futuro lo mismo sucederá con Wikipedia, YouTube, Facebook, Twitter, Instagram o Reddit.

¿Qué son los medios sociales?

- A pesar de eso los ejemplos dejan clara una segunda característica de los medios sociales: requieren una **red de telecomunicaciones** (ahora Internet, antes las BBS solo necesitaban líneas telefónicas).
- Sin embargo, el aspecto más crucial de los medios sociales es, precisamente, su componente **social**.
- No sería hasta la quinta revisión de la entrada de la Wikipedia que se dan algunas pistas al respecto:

> *people create and share with each other [opinions, insights, experiences and perspectives]*

- Hogan y Melville, por su parte, argumentan que los medios sociales se caracterizan por una comunicación de **muchos-a-muchos** (en comparación con el uno-a-muchos de las retransmisiones o el muchos-a-uno de los sistemas de peticiones online) y que **carece de una audiencia específica** (*"all Facebook friends, Twitter followers, readers of a bulletin board, and so forth"*).
- Dicho de otro modo, la parte social de los medios sociales viene dada por las relaciones que se establecen entre los usuarios cuando crean y consumen contenidos. **Cualquiera puede ser un creador con su propia audiencia al mismo tiempo que es parte de la audiencia de una multitud de usuarios.**

¿Qué son los medios sociales?

¿Qué son los medios sociales?

- La postura de Baym no es aislada pero tampoco es habitual.
- En esa línea es recomendable leer a Jodi Dean, p.ej. *"Blog Theory: Feedback and Capture in the Circuits of Drive"* donde acuña el término **capitalismo comunicativo**.
- El objetivo del capitalismo comunicativo es

> *capture their users in intensive and extensive networks of enjoyment, production, and surveillance.*

- Bajo ese marco el mensaje, su autoría o su audiencia son irrelevantes siempre y cuando contribuyan a un flujo continuo de comunicación que permita el lucro de los dueños de la plataforma.
- **Ese modelo de negocio es lo que ha permitido que todo tipo de usuarios y prácticas maliciosas prosperen puesto que no hay incentivo económico para controlarlas.**

¿Qué son los medios sociales?

- En realidad, la postura tampoco es nueva y críticas muy similares ya fueron hechas por Carmen Hermosillo *humdog* **¡en 1994!**

> *[cyberspace] is a black hole; it absorbs energy and personality and then re-presents it as* ***spectacle****. […] it is fashionable to suggest that cyberspace is some kind of _island of the blessed_ where people are free to indulge and express their Individuality. […] in reality, this is not true. […] i have seen many people spill their guts on-line, and i did so myself until, at last, i began to see that i had* ***commodified myself****. commodification means that you turn something into a product which has a money-value. […] i created my interior thoughts as a means of production for the corporation that owned the board i was posting to, and that commodity was being sold to other commodity/consumer entities as* ***entertainment****. […] many cyber-communities are businesses that rely upon the* ***commodification of human interaction****.*


Pope Francis
@Pontifex · Follow
We must adapt our socio-economic models so they have a human face, because many models have lost it. Thinking about these situations, in God's name I want to ask:
6:00 PM · Oct 16, 2021
8.2K
Reply
Copy link
Read 170 replies


Pope Francis
@Pontifex · Follow
Technology giants to stop preying on human weakness, people's vulnerability, in order to make a profit.
6:06 PM · Oct 16, 2021
17.3K
Reply
Copy link
Read 380 replies

¿Qué son los medios sociales?

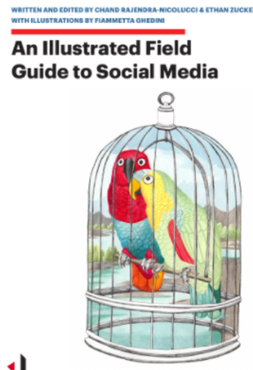
- Los **medios sociales** son herramientas digitales de **coste irrisorio** y fácilmente accesibles que permiten a cualquiera **publicar, compartir y acceder** a contenidos multimedia destinados a audiencias inespecíficas, **colaborar** en acciones colectivas y **entablar relaciones personales**. Aunque no es un requisito, lo cierto es que las plataformas más populares en la actualidad están centralizadas y son de propiedad privada.
- Esta definición cubre a sitios como Facebook o Twitter pero también blogs, listas de distribución o chats.
- **Todos ellos están afectados por los comportamientos maliciosos que vamos a estudiar.**

Para saber más...

Web Information Systems / Social media and digital platforms

Social media and digital platforms

Add comments Made with Milanote



https://www.tandfonline.com/doi/abs/10.1080/

Social Media: Defining, Developing, and Divining

What is a social medium, and how may one moderate, isolate, and influence communicative processes within? Although scholars assume an inherent understanding of social media based on extant technology, there is no commonly accepted definition of what social media *are*, both functionally and theoretically, within communication studies. Given this lack of understanding, cogent theorizing regarding the uses and effects of social media has been limited. This work first draws on extant definitions of social media and subcategories (e.g., social network sites) from public relations, information technology, and management scholarship, as well as the popular press, to develop a definition of social media precise enough to embody these technologies yet robust enough to remain applicable in 2035. It then broadly explores emerging developments in the features, uses, and users of social media for which future theories will need to account. Finally, it divines and prioritizes challenges that may not yet be apparent to theorizing communication processes with and in mercurial social media. We address how social media may uniquely isolate and test communicative principles to advance our understanding of human–human and human–computer interaction. In all, this article provides a common framework to ground and facilitate future communication scholarship and beyond.

Prehistory

3 cards

https://arstechnica.com/information-technc

A 1986 bulletin board system has brought the old Web back to life in 2017

Limited to an anachronistic 1200 bits per second, it took several moments for the green-phosphor ASCII art to scroll from the bottom to the top of the screen. A login prompt and a blinking cursor invited me to continue deeper: Enter GUEST for a quick look around.)

http://sfhqbbs.org/telnet-dura.php

Home Page - STar Fleet HQ - Atari Telnet BBS

fTelnet Client I do not run this BBS. I provide this connection because I really like this BBS and want to help make it easier for you to contact them too. Site design © 2017 - Andrew Klenotic newdatejust.com Some site content is copyrighted by their respective owners

https://twitter.com/TwitterSupport/status/1195

Twitter Support on Twitter

We've heard you on the impact that this would have on the accounts of the **deceased**. This was a miss on our part. We will not be removing any inactive accounts until we create a new way for people to memorialize accounts.

https://www.theguardian.com/commentisfree/

Reports of social media's influence on voters are greatly exaggerated

You know the joke: one dark night, a policeman comes on a drunk rootling around under a street lamp. When asked what he's doing, the guy says that he's looking for his car keys. "Is this where you dropped them?" asks the cop. "No," comes the reply.

https://twitter.com/sivavaid/status/126638544

SIVA VAIDHYANATHAN on Twitter

For every American Twitter user there are seven American Facebook users. For every Twitter user worldwide there are eight Facebook users. So what Twitter does matters very little.

https://www.tandfonline.com/doi/full/10.1080/1

What is platform governance?

Following a host of high-profile scandals, the political influence of platform companies (the global corporations that that operate online 'platforms' such as Facebook, WhatsApp, YouTube, and many other online services) is slowly being re-evaluated.

0 Unsorted

https://twitter.co

Paul Ford on Twitt

Web nerds in 2005: Yo
hear so many voices y
we should get rid of th
Social media: Got it, g
comments section.

https://www.w3.c

Socialwg

Specifications that the
produced. Some use c
by multiple specificati
approaches, so trying
can be a bit confusing
overview of these pro

Oportunidades que (en teoría) ofrece la minería de medios sociales

- En teoría, puesto que **cualquiera** puede compartir **cualquier** contenido en medios sociales, la minería Web en los mismos abriría la puerta a:
  - **Comprender mejor el comportamiento individual y social**. [+]
  - **Conocer la opinión pública en tiempo real**. De interés para gobiernos, administraciones públicas, compañías y marcas.
- No, no y no.
- Las novelas de Asimov nos dan una pista...

Oportunidades que (en teoría) ofrece la minería de medios sociales

*La **psicohistoria** es el nombre de una ciencia ficticia en el universo de la Saga de la Fundación de Isaac Asimov, que es una combinación de historia, psicología y estadística matemática para calcular el comportamiento estadístico de poblaciones extremadamente grandes, como la del Imperio Galáctico.*

*"Implicit in all these definitios is the **assumption** that the **human conglomerate** being dealt with is **sufficiently large** for valid statistical treatment.*

*A further necessary **assumption** is that the human conglomerate be itself **unaware of psychohistoric analysis** in order that its reactions be truly random."*

*"The **unstated axiom**: that there is **only one species of intelligence** in the Galaxy and that it is Homo Sapiens. If there were 'something new:' if there were other species of intelligence widely different in nature, then their behavior would not be described accurately by the mathematics of psychohistory and Seldon's Plan would have no meaning."*

Oportunidades que (en teoría) ofrece la minería de medios sociales

- Trasladando la psicohistoria al terreno de la minería de medios sociales estamos suponiendo lo siguiente:
  1. Los **usuarios de medios sociales** son una **muestra** razonablemente **representativa** de la población en su conjunto y, además, se comportan y expresan *online* de manera análoga a como lo hacen en el mundo físico.
  2. Los **usuarios** de medios sociales **no conocen** las predicciones **ni son conscientes** del análisis al que son sometidos.
  3. Los **únicos usuarios** de medios sociales son **seres humanos**.
- **Todas esas suposiciones son falsas:**
  - los medios sociales no son representativos de la sociedad (⚠ pero la prensa los interpreta como si lo fueran [1]);
  - el comportamiento *online* no siempre refleja el comportamiento *offline* y en muchas ocasiones es manifiestamente falso;
  - los usuarios son muy conscientes de que sus acciones colectivas pueden cambiar la percepción sobre un evento desde el exterior;
  - los bots y títeres abundan en medios sociales (anécdota, ahora..., incluso hay bots con depresión y ansiedad: aquí y aquí).
- La realidad es que la **manipulación *online*** está a la orden del día (en particular en política): p.ej. en Brasil, México, España (incluida Cataluña), Canadá, Polonia o Bolivia.
- **Los problemas de representatividad son salvables**, p.ej. con "paneles virtuales".
- **Los problemas de manipulación y desinformación son un reto** y, por eso, debemos conocer la forma en que operan los distintos agentes maliciosos.
- Libros interesantes (no técnicos): *"Compromised Data - From Social Media to Big Data"* y *"Meaning in the Age of Social Media"*.
- Más recursos en el *moodboard*.

Agentes maliciosos en la Web

- Los usuarios (o agentes) maliciosos son aquellos que **explotan la Web con la intención de causar daño a otros usuarios o que causan daño en la persecución de su beneficio**.
- Existen tres tipos básicos de usuarios maliciosos en la Web:
  - **Troles:** Usuarios que intentan **sembrar discordia** en una comunidad online.
  - **Bots y títeres:** Los **bots son cuentas automatizadas** en redes sociales que son capaces de publicar contenidos e interaccionar con otros usuarios; se consideran maliciosos cuando fingen ser usuarios reales con opiniones reales. Los **títeres** son cuentas que también fingen ser un usuario real pero que en realidad son **identidades falsas controladas por un tercero**, normalmente con la intención de desinformar.
  - **Vándalos:** Usuarios que alteran el contenido de sitios web con fines humorísticos, políticos o manipulativos. Por ejemplo, el vandalismo en la Wikipedia es muy habitual pues no requiere ninguna habilidad técnica.
- De los anteriores prestaremos atención únicamente a los dos primeros por su impacto en el terreno de la discusión política.
- El vandalismo no solo es problemático por sus acciones sino que, en ocasiones, puede llevar a que información incorrecta, falsa o meramente paródica acabe siendo aceptada como veraz [1].
- Si dicha "veracidad" es refrendada por la prensa se da la circunstancia de que ya existe una "fuente" externa que puede usarse para dar soporte a la información incorrecta en Wikipedia. [2]

Agentes maliciosos en la Web

Troles

- Algunas definiciones...

> *Un trol es una persona que publica mensajes provocadores, irrelevantes u off-topic en una comunidad online (foros, chats, comentarios en periódicos o blogs, etc.) para iniciar discusiones, molestar o alterar emocionalmente a los otros miembros de la comunidad, o para desbaratar la conversación normal en la comunidad. Sus motivos son muy diversos pudiendo hacerlo, no para obtener una ganancia específica, sino por mera diversión.*

> *Un trol es alguien que se comporta sin seguir el comportamiento entendido como aceptable en una comunidad online.*

Agentes maliciosos en la Web

Troles

- Otra definición...

> *A troller is a CMC user who constructs the identity of sincerely wishing to be part of the group in question, including professing, or conveying pseudo-sincere intentions, but whose real intention(s) is/are to cause disruption and/or to trigger or exacerbate conflict for the purposes of their own amusement. Just like malicious impoliteness, trolling can (1) be **frustrated** if users correctly interpret an intent to troll, but are not provoked into responding, (2) be **thwarted**, if users correctly interpret an intent to troll, but counter in such a way as to curtail or neutralize the success of the troller, (3) **fail**, if users do not correctly interpret an intent to troll and are not provoked by the troller, or, (4) **succeed**, if users are deceived into believing the troller's pseudo-intention(s), and are provoked into responding sincerely. Finally, users can **mock troll**. That is, they may undertake what appears to be trolling with the aim of enhancing or increasing affect, or group cohesion.*

Agentes maliciosos en la Web

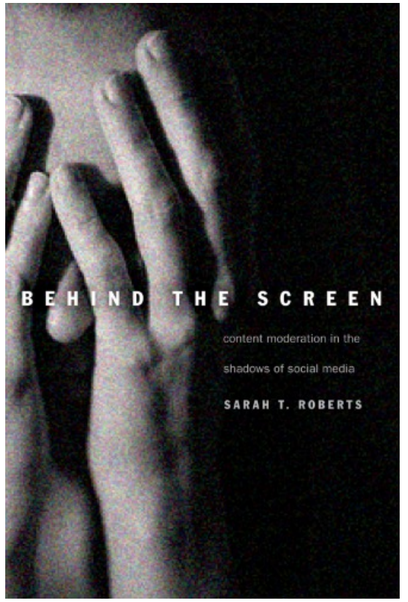
Troles

- Hay áreas de la Web donde los troles campan a sus anchas como las secciones de comentarios de periódicos, blogs y otros sitios web (p.ej. YouTube):
  - *The Worst Parts Of Humanity Live In The Comments Section*
  - *Comment sections are poison: handle with care or remove them*
  - *Are Comment Sections Worth It?*
  - *Why are YouTube comments the worst on the internet?*

Agentes maliciosos en la Web

Troles

- La presencia continua de troles tiene un impacto palpable: o bien se elimina por completo la participación de los usuarios (cosa imposible en un foro, p.ej.) o bien se deben dedicar esfuerzos enormes a **moderar los contenidos antes de su publicación**.
- En muchos casos la moderación es llevada a cabo por **voluntarios** dentro de cada comunidad siendo ellos mismos quienes establecen las normas de lo aceptable y no inaceptable (véase [1], [2], [3] y [4]). Un caso paradigmático de esto es Reddit.
- En otros casos los propietarios de la plataforma establecen las reglas de lo aceptable (p.ej. Facebook, o Twitch: aquí y aquí) y cuentan con **empleados** que son quienes criban los contenidos de acuerdo a las mismas. El libro *"Behind the Screen"* de Sarah T. Roberts es una obra muy reciente que expone en toda su crudeza la realidad detrás de ese trabajo. En esa línea también es muy recomendable el documental *"The Cleaners*.
- La primera consecuencia de la moderación es la posibilidad (real o imaginada) de **censurar contenidos y usuarios bajo el razonamiento de que incumplen las reglas de la comunidad**, o a instancias de distintos gobiernos.
- La segunda es la posibilidad de **explotar la moderación o el discurso alrededor de la moderación como un arma contra los adversarios**. Así, es posible coordinar acciones de reporte contra un usuario al que se desea bloquear (un ejemplo, otro ejemplo).
- Por otro lado, es habitual entre usuarios, partidos y organizaciones conservadoras afirmar que los medios sociales tienen un sesgo liberal y les censuran (quizás por artículos como éste o por usar un vocabulario específico que ha sido apropiado por un solo extremo del espectro), llevando a situaciones como [1] y [2, 3] aunque, en realidad, **el sesgo ideológico en las plataformas es, más bien, hacia la derecha** [4][5].



Agentes maliciosos en la Web

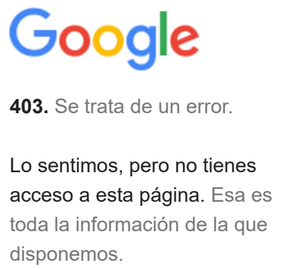
Troles

- Existen puntos de vista contradictorios sobre la naturaleza de los troles:
  - Por un lado se ha afirmado que los troles lo son por naturaleza y que exhiben rasgos como el narcisismo, el sadismo, la psicopatía o el maquiavelismo ([1], [2] y [3]).
  - Por otro, se ha encontrado que cualquier usuario bajo las circunstancias adecuadas (estado de ánimo o un contexto previo de trolling) puede convertirse en un trol ([4]) o puede que no... ([5])
- Esta situación implica que...
  - en el caso de los troles "innatos" bastaría con detectarlos y expulsarlos de la comunidad.
  - en aquellos casos en los que alguien se convierte en un trol no habría solución sencilla; sobre todo cuando los comentarios negativos de la comunidad hacia un usuario sólo hacen empeorar la situación y pueden llevar a que haya aún más trolling [6].

Agentes maliciosos en la Web

Troles

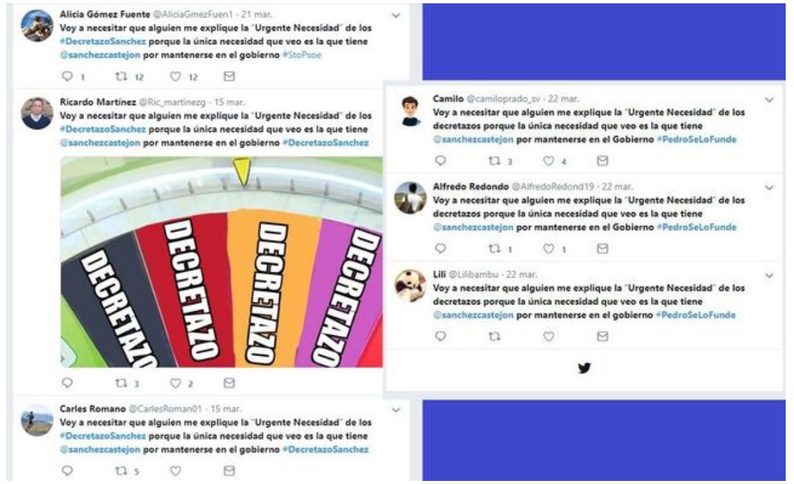
- Por otro lado, no todos los troleos son producto de troles "de verdad":
  - Existen las **"granjas de troles"** en las que personas, normalmente pagadas por un gobierno, utilizan técnicas de troleo para desbaratar comunidades *online* de naturaleza política.
  - En algunas ocasiones se ha hablado de **troles "buenos"**. Por ejemplo, cuando *kpopers* trolearon la campaña de Trump del 2020 o las cuentas de VOX en España...
- Para saber más sobre el espectro de los troles y las distintas técnicas que usan os puede interesar el artículo *"Black Hat Trolling, White Hat Trolling, and Hacking the Attention Landscape"* (disponible en el *Perusall* de la asignatura).



Agentes maliciosos en la Web

Bots

- Los *bots* son cuentas totalmente automatizadas que publican contenidos e interaccionan (p.ej. mediante likes o retuits) con usuarios reales en medios sociales.
- Por supuesto no todos los bots son maliciosos (ejemplo) pero aquellos que hacen creer a otros usuarios que son personas reales con opiniones auténticas pueden considerarse como tales (ejemplo de su uso durante la campaña del Brexit y ejemplo de su impacto negativo en el debate político).
- Una ventaja para los creadores de este tipo de cuentas es que son muy sencillas de gestionar, su mayor inconveniente es que son relativamente sencillas de detectar puesto que publican los mismos mensajes con los mismos *timestamps* (ejemplo).
- Aprovechando esas características, el grupo OSoMe de la Universidad de Indiana desarrolló la herramienta Botometer para detectar *bots* en Twitter (muy recomendables sus publicaciones al respecto). Otras herramientas similares son botcheck.me y Bot Sentinel.
- ⚠ De manera reciente se han señalado algunos aspectos problemáticos y preocupantes en la forma en que estas herramientas (en particular Botometer) detectan bots arrojando muchas dudas sobre la investigación realizada hasta el momento y sobre la supuesta prevalencia de bots en medios sociales. **Véase.**
- Para saber un poco más sobre bots y desinformación en política os puede interesar este post: ¿Qué posiciones ideológicas usan más bulos y bots?



Agentes maliciosos en la Web

Títeres

- Los títeres (*sockpuppets*) deben considerarse siempre maliciosos puesto que son cuentas operadas por humanos que finjen ser una persona diferente al operador. No confundir estas acciones con suplantación de identidad, o la creación de cuentas para *lurking* o para un círculo más íntimo (p.ej. las cuentas *finsta* en Instagram) .
- Generalmente los títeres tienden a distribuir contenido erróneo o manifiestamente falso y aunque pueden tener un impacto grave publicando reseñas falsas la mayor preocupación actual es por su uso en el terreno político.
- Uno de los primeros casos que se discutieron públicamente ocurrió en 2011 cuando salió a la luz la solicitud por parte de la Fuerza Aérea de los EE.UU. de un "software de gestión de personas" para su uso en Iraq y Afganistán.

> *Software will allow 10 personas per user, replete with background , history, supporting details, and cyber presences that are technically, culturally and geographacilly consistent. Individual applications will enable an operator to exercise a number of different online persons from the same workstation and without fear of being discovered by sophisticated adversaries. Personas must be able to appear to originate in nearly any part of the world and can interact through conventional online services and social media platforms. The service includes a user friendly application environment to maximize the user's situational awareness by displaying real-time local information.*

Agentes maliciosos en la Web

Títeres

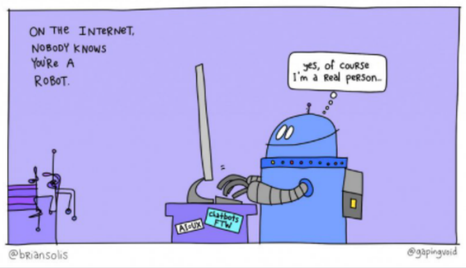
- Posteriormente, en 2015, ganó notoriedad la "Agencia de Investigación de Internet" rusa, una "granja de troles" cuyos objetivos fueron opositores rusos pero también Ucrania y las elecciones en EE.UU. de 2016.
- Esos ejemplos, o los de los internautas chinos pagados por su gobierno para publicar propaganda, tienen una motivación política clara pero no es la única. Recordemos que los medios sociales son un instrumento del capitalismo comunicativo y que el *engagement* supone oportunidad de lucro. En consecuencia, no es sorprendente que surjan granjas de troles por motivos meramente económicos (ejemplo).
- A pesar de su espectacularidad **se ha puesto en duda el impacto real de estas actividades en términos generales** [1] aunque pueden tener un **impacto apreciable en las comunidades más polarizadas** [2].
- El mayor problema con los títeres es que al estar controlados por seres humanos son menos predecibles y son, en consecuencia, **más difíciles de detectar**. Aún así se han hecho ciertos avances en el área [3][4][5][6].

Para saber más...

Web Information Systems / Bots, sockpuppets and trolls

Bots, sockpuppets and trolls

Add comments Made with Milanote



http://amultiverse.com/comic/2013/11/06/trollcore/

https://cacm.acm.org/magazines/2016/7/204C

The Rise of Social Bots

Bots (short for software robots) have been around since the early days of computers. One compelling example of bots is chatbots, algorithms designed to hold a conversation with a human, as envisioned by Alan Turing in the 1950s.

https://apnews.com/4086949d878336f8ea6d

Cyborgs, trolls and bots: A guide to online misinformation

NEW YORK (AP) - Cyborgs, trolls and bots can fill the internet with lies and half-truths. Understanding them is key to learning how misinformation spreads online. As the 2016 election showed, social media is increasingly used to amplify false claims and divide Americans over hot-button issues including race and immigration.

https://www.internetmythen.de/en/?mythen=n

Myth #30: Digital rights campaigns are run by bots, not real activists. ~ 50 Myths of the Internet

Myth: Like many other digital rights campaigns, the 2018/2019 protests against the EU Copyright Directive were not an expression of massive civic concern for digital rights. The protest campaign, the largest of its kind in recent years, was in fact a prime example of disinformation activities.

https://www.businessinsider.com/trolls-bots-fl

Trolls and bots are flooding social media with disinformation encouraging states to end quarantine

Protests have popped up around the US calling for states to end quarantine and reopen businesses. An analysis from Bot Sentinel, a bot tracking platform, found that bots and trolls have been stoking sentiments online that have fueled the protests, using hashtags like #ReopenAmericaNow and #StopTheMadness.

0 Unsorted

Trolls
59 cards

Bots
59 cards

Sockpuppets
2 cards


Papa Francisco
@Pontifex_es · Follow
Todos somos responsables de la comunicación que hacemos, de las informaciones que damos, del control que juntos podemos ejercer sobre las noticias falsas, desenmascarándolas. Todos estamos llamados a ser testigos de la verdad: a ir, ver y compartir. #JMCS
9:00 AM · May 16, 2021
27.5K
Reply
Copy link
Read 581 replies

Información errónea y desinformación

- **Una premisa habitual de los usuarios de la Web es que la información que encuentran en ella es verdadera** [1].
- Ni que decir tiene, **una porción sustancial de dicha información es falsa**. De hecho, una parte importante de los contenidos, usuarios y acciones que tienen lugar en la Web son falsos (véase).
- Además, **no toda la información falsa es igual**:
  - Por un lado hay información falsa **basada en opiniones** (p.ej. reseñas falsas y *astroturfing*) o en **hechos** (p.ej. rumores y bulos) [2].
  - Por otro lado hay información falsa que se crea *ex profeso* para **intoxicar** (desinformación, del ruso дезинформация) e **información errónea** que se distribuye sin mala intención o incluso de buena fe.

- Dependiendo entonces de su naturaleza, la problemática de la información falsa puede enfrentarse de distintos modos:
  - Existe más de una década de trabajo sobre **reseñas falsas** en la Web puesto que tienen un impacto económico inmediato [3] [4] [5] [6] [7] [8]
  - También existe bastante literatura sobre la **difusión de rumores** (hechos falsos que se propagan sin mala intención) [9] [10] [11] [12] [13] [14] [15] [16]. En este sentido, la herramienta *Twitter**Trails*** desarrollada en Wellesley College resulta muy interesante (historias recientes, historias archivadas). Hoaxy del grupo OSoMe persigue objetivos similares.
  - Se ha trabajado muy intensamente en detección de ***astroturfing*** (simular un falso apoyo popular de algo o alguien).
  - Por último, hay múltiples trabajos sobre **desinformación en medios sociales** (información falsa que se difunde con el objeto de dañar a alguien).
- Por su interés reciente y relación con los agentes maliciosos que ya hemos estudiado **nos centraremos en la detección de campañas de *astroturfing* y desinformación en medios sociales.**

Información errónea y desinformación

Astroturfing

- En inglés el término *grassroots* hace referencia a movimientos (normalmente políticos) que **crecen desde las bases y evolucionan de una manera autoorganizada y no jerárquica**. Un ejemplo en España de este tipo de movimientos sería el 15M.
- Puesto que, literalmente, *grassroots* significa "las raíces de la hierba" se comenzó a usar el término ***astroturfing*** para referirse a **movimientos orquestados pero que fingen ser espontáneos y de base** ya que *Astro Turf* es un conocido fabricante de cesped artificial.
- Por supuesto, el *astroturfing* no requiere de la Web ni de medios sociales pero es ahí donde ha alcanzado proporciones enormes al utilizar bots o títeres.
- Un ejemplo relativamente reciente ocurrió en EE.UU. cuando, durante el proceso para revertir las políticas de Obama sobre neutralidad de red, la FCC recibió casi medio millón de comentarios falsos de "personas" que afirmaban estar contra dichas políticas (ver aquí, aquí y aquí).
- Otro ejemplo bien conocido de *astroturfing* en Twitter tuvo lugar durante las elecciones presidenciales mexicanas de 2012 con los "peñabots".
- Entre los primeros sistemas de detección automática de *astroturfing* en medios sociales podemos citar *Truthy* (del grupo OSoMe de la Universidad de Indiana) [1][2] y el desarrollado por Kyumin Lee *et al.* en la Universidad de Texas A&M. Un ejemplo más reciente es el descrito por Onur Varol *et al.*, de nuevo de Indiana.
- Rizando el rizo, *Truthy* y sus creadores fueron objeto de una campaña de desinformación en 2014 según la cual estaban desarrollando para el gobierno de Obama una herramienta para limitar la libertad de expresión de sus oponentes (tuits de ejemplo aquí y aquí).

Información errónea y desinformación

Desinformación

- La **desinformación y las campañas de desprestigio** son, quizás, los mejores ejemplos del uso de la Web para causar el mayor daño posible.
- Su objetivo no es otro que **dañar al oponente mediante la difusión masiva de información falsa sobre el mismo**.
- Entre los primeros ejemplos podemos encontrar:
  - La campaña de *link bombing* que instigó en 2006 el blog progresista MyDD.com contra los Republicanos que concurrían a elecciones al Congreso de EE.UU.; el objetivo era alterar los resultados de los buscadores para hacer llegar al público historias negativas sobre dichos políticos (ver aquí, aquí y aquí).
  - Una campaña de desprestigio organizada en Twitter en 2008 contra la candidata demócrata al Senado de EE.UU Martha Coakley (véase).
- Tenemos ejemplos más recientes en las elecciones presidenciales de EE.UU. de 2016 [1], el referendum del Brexit también de 2016 [2] o las presidenciales francesas de 2017 [3].
- Al igual que con el *astroturfing* estas campañas dependen enormemente de **bots y títeres** y están fuertemente vinculadas con las ya mencionadas **"granjas de troles"**.
- Como ya sucedió con los títeres, **no hay suficientes datos sobre el impacto real de este tipo de información** aunque parece que quienes tienen más posibilidades de verse expuestos a la misma son quienes ya están más polarizados y que, además, los efectos persuasivos son muy limitados.

Información errónea y desinformación

Fake text

- Desde hace algún tiempo es posible generar grandes cantidades de texto en lenguaje natural con una apariencia superficial de verosimilitud mediante los denonimados *Large Language Models* (p.ej., GPT, GPT-2 o GPT-3). **Demo**
- La posibilidad de generar textos razonablemente realistas de manera automática abre la puerta a su abuso.

> *An AI fake text generator that can write paragraphs in a style based on just a sentence has raised concerns about its potential to spread false information.*
> *It was developed by two students at the University of Toronto in Canada and the results are currently making the rounds on Twitter.*
> *But while fake news can be a problem in the real world, and fake news accounts for around 1.4% of social media, writing fake news out of whole cloth, as researchers have found, is an art that's harder to pull off.*
> *At first glance, their AI text generator is an impressive creation. It turns basic sentence structure into impressive texts without even having to write a word of text of its own.*
> *Their "writing language" is based on a simple rule, they say: "Only one noun character can be used per sentence and not more than once per sentence". And with that in place, the generator generates standard, boring, computer-generated texts.*

- Del texto anterior sólo el primer párrafo fue escrito por un humano...
- Se están tratando de desarrollar técnicas para determinar si un texto ha sido generado automáticamente o ha sido escrito por una persona (véase).
- Para ocultar algo ya no hay que censurarlo, sólo hay que inundar a la población con otra información. Artículo reciente de investigadores chinos sobre cómo generar comentarios a noticias de forma automática (hilo en Twitter criticando la motivación ética de esta práctica).

Información errónea y desinformación

Verificación de hechos y autocontrol de las plataformas

- Habría dos modos no automáticos de responder ante la desinformación en medios sociales:
  - Por un lado, agentes externos podrían verificar los hechos que se difunden (aka *fact checking*).
  - Por otro, las propias plataformas podrían aplicar políticas que evitasen la publicación de desinformación.
- Ambas aproximaciones tienen un obvio problema de escalabilidad pues requieren en última instancia de personal humano.
- La segunda, además, llevaría la moderación de contenidos un paso más allá dedicarse empresas privadas a dirimir qué es verdad y qué no lo es (aún así hay quien se lo está pidiendo). Además, dichas políticas podrían llevar en casos extremos al *deplatforming* [1][2][3] de usuarios o comunidades concretas por difundir desinformación de manera reiterada.
- Hay que tener en cuenta, además, que las plataformas no están necesariamente interesadas en ninguna de dichas soluciones (recordemos, capitalismo comunicativo):
  - Durante 2017 y 2018 Snopes llevó a cabo verificación de hechos en Facebook para terminar el acuerdo sin dar demasidas explicaciones en febrero de 2019.
  - La Unión Europea comenzó a desarrollar desde 2018 un Código de buenas prácticas en materia de desinformación al que las principales plataformas se fueron adhiriendo. A pesar de eso, Facebook tolera abiertamente que los políticos publiquen anuncios con información falsa (aquí, aquí y aquí).
- A pesar de todo, se siguen dedicando esfuerzos a mejorar los sistemas de verificación de hechos, p.ej. EUfactcheck.eu o *SOMA Disinfobservatory* y se están depositando bastantes esperanzas en la posibilidad de automatizar la verificación de hechos.

Información errónea y desinformación

Para saber más...

- **Os puede resultar entretenido hacer este *quiz*...**
- Literatura sobre desinformación, propaganda y *fake news*:
  - *The spread of true and false news online*
  - *The science of fake news*
  - *Social media, political polarization, and political disinformation: A review of the scientific literature*
  - *Weaponizing the Digital Influence Machine: The Political Perils of Online Ad Tech*
  - *Data Voids: Where Missing Data Can Easily Be Exploited*
  - *PHEME: Computing Veracity – the Fourth Challenge of Big Data*

Para saber más...

Conclusiones

- La Web es desde sus inicios un entorno adversarial, esa situación solo empeora cuanta más gente usa la Web para obtener información para más actividades.
- El *spam* es un problema común pero no es el más acuciante para la sociedad.
- Los usuarios maliciosos y la información falsa pueden ser mucho más graves.
- En los medios sociales abundan las cuentas automatizadas total o parcialmente que fingen ser personas reales.
- Esas cuentas se utilizan tanto para promocionar ciertas ideas como para intoxicar con información falsa.
- Hay más anécdotas sobre el impacto de estos comportamientos que datos sólidos.
- Existe bastante literatura sobre la detección automática de este tipo de comportamientos pero hay que reflexionar sobre esa investigación pues es probable que se haya sobreestimado la prevalencia de los bots y, además, aún queda mucho por hacer, en particular en relación a la automatización de la verificación de hechos.