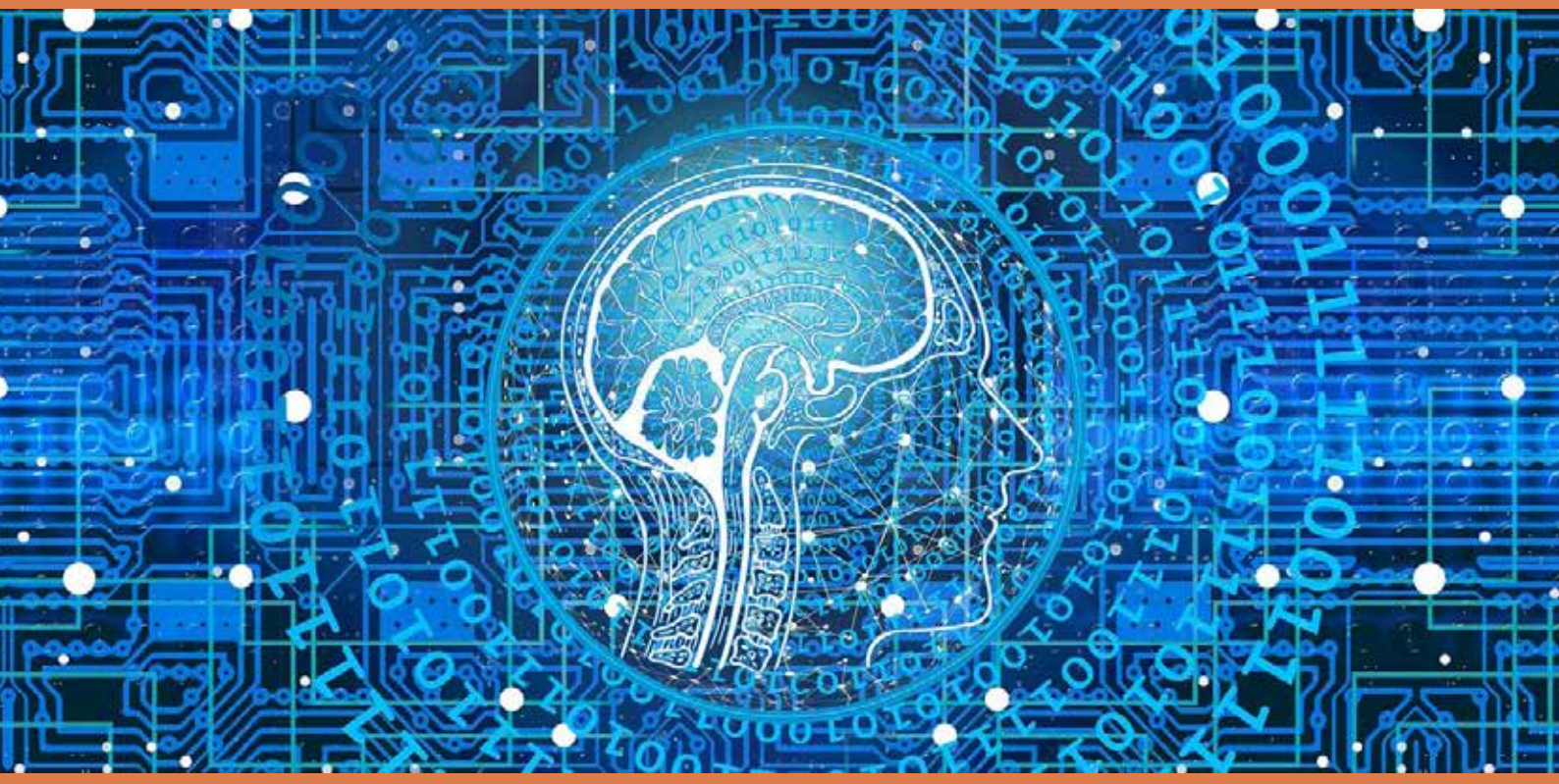
Detección de noticias a través de aplicaciones de inteligencia artificial



La inteligencia artificial aplicada a informativos 2019-2020

OI2 Observatorio para la Innovación de los Informativos en la Sociedad Digital

rtve + UAB Universitat Autònoma de Barcelona

Este informe se ha desarrollado bajo la investigación "Inteligencia artificial aplicada al periodismo 2019-2020" por la Cátedra RTVE-UAB Innovación en los informativos en la sociedad digital.

Equipo de investigación

Por parte de RTVE: José Juan Ruiz, Pere Vila, David Corral, Carmen Pérez, Esteban Crespo, Esteban Mayoral, Miguel Ángel Martín, Pedro Cánovas

Por parte de la UAB: José Manuel Pérez Tornero, Cristina Pulido, Santiago Tejedor, Laura Cervi, Diana Sanjinés, Wei Zhang, Sally Tayie.

Con la colaborción de

red INNO NEWS
CSO2017-90819-REDT

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

27 de noviembre del 2019, Barcelona.

Índice

Introducción

Este informe - intitulado *Detección de noticias a través de aplicaciones de inteligencia artificial*- se inscribe en el marco de un proceso de investigación centrado en la innovación del periodismo audiovisual, que Radio Televisión Española (RTVE) y la Universidad Autónoma de Barcelona (UAB) pusieron en marcha en el año 2015, año en el que se constituye el Observatorio para la Innovación de las Informativos en la Sociedad de la información (OI2) y se crea una cátedra conjunta sobre el mismo tema. Como tal informe, es el primero de una serie relacionada con la línea de investigación desarrollada por OI2 que se ocupa de la relación entre periodismo e inteligencia artificial.

Desde 2015, Observatorio OI2 y la Cátedra RTVE-UAB han desarrollado una serie continuada de actividades -jornadas, debates, seminarios, investigación y experimentos-, destinadas a explorar, e implementar -en la medida de lo posible-, las capacidades de innovación, que tanto la tecnología como el avance del conocimiento en general, están abriendo en los últimos tiempos. El objetivo fundamental de todas estas actividades es descubrir las tendencias de futuro en el sector y avanzarse -mediante estrategias de desarrollo y formación- a los cambios previsibles.

En este marco, los primeros estudios y actividades, desde el 2015 al 2018, se centraron en un objeto de estudio: la transformación de los sistemas de producción y difusión de la información televisiva en el nuevo marco de la sociedad de la información. Dentro de este enfoque, la atención se puso en cuestiones esenciales:

1. Las innovaciones que afectaban a la información audiovisual en general y sus consecuencias en el funcionamiento y la estructura de los canales de información continua (de 24 horas).

2. Las nuevas posibilidades y formas de participación de la audiencia en la producción de información audiovisual que las redes sociales y las tecnologías ligeras permiten.

El análisis de la innovación en materia de información audiovisual y las estrategias de los canales de 24 horas se inició con la celebración de las Primeras Jornadas de OI2, el 3 de noviembre de 2015. En ellas, más de 25 profesionales y expertos, de diversos países europeos -con presencia de las grandes cadenas europeas[1]- tuvieron la oportunidad de compartir experiencias y reflexionar conjuntamente sobre el nuevo rumbo que la sociedad digital ha impuesto a los sistemas de producción de noticias de los servicios públicos audiovisuales europeos.

Posteriormente, en mayo de 2016, dentro del marco de un seminario celebrado en la Universidad CEU San Pablo de Madrid, se estudiaron las grandes líneas de transformación que afectaban a la información audiovisual y se empezó a dibujar una nueva hoja de ruta para la innovación tecnológica y de programación. Como conclusión de todos estos trabajos, se elabora y se publica un primer informe del Observatorio en la línea de innovación de informativos: *El reto de la innovación de los informativos en la era digital*[2].

Entre 2016 y 2017, el periodismo móvil y las nuevas

---

1 En estas primeras jornadas también participaron representantes de las televisiones públicas: la francesa de France 24, la italiana RAI News 24 y la alemana ZDF.

2 http://www.gabinetecomunicacionyeducacion.com/sites/default/files/field/publicacion-adjuntos/oi2_el_reto_de_la_innovacion_de_los_informativos_en_la_era_digital.pdf

posibilidades de participación que -junto a las redes sociales- fueron los grandes temas de estudio e investigación en OI2 y la Cátedra RTVE-UAB.

En junio de 2016, dentro de un proyecto europeo, Y-Nex, tiene lugar un taller de formación en materia de periodismo móvil destinado a profesionales de diversas televisiones europeas. El taller no solo sirvió como una oportunidad de reciclaje profesional, sino que permitió experimentar con las nuevas tecnologías ligeras e intercambiar experiencias entre diversos servicios públicos audiovisuales.

Posteriormente, en el mismo 2016 y en 2017, dentro del Salón Manga de la Feria de Barcelona, se desarrollaron diversas experiencias de aplicación de tecnologías de periodismo móvil en la producción y difusión de noticias. Allí se realizaron actividades de experimentación y se adquirió experiencia para desarrollar ciertas tecnologías de cara a su integración en las rutinas de producción de RTVE.

En noviembre de 2016, en la Universidad Autónoma de Barcelona, se desarrollan las II Jornadas de OI2, centradas ya en el periodismo móvil y en su capacidad de innovación. En ellas se presentaron las diversas experiencias realizadas por OI2 y la Cátedra RTVE-UAB a lo largo del año, y se analizaron las tendencias y aplicaciones de las herramientas de periodismo móvil en los diversos procesos de producción audiovisual.

Como conclusión de todas estas actividades se inicia la elaboración del segundo informe del *OI2 MOJO*. Manual de periodismo móvil[3] que se publica en 2018. El manual presenta una puesta al día, desde el punto de vista teórico y práctico, del periodismo móvil y de las tecnologías disponibles. Recoge todas las experiencias llevadas a cabo por OI2 en esta materia y trata de establecer un mapa de tendencias en el sector. La utilidad de este manual se pone a prueba de inmediato en diversos cursos realizados por el IORTVE en diversas sedes de RTVE -Madrid, Barcelona, Castilla la Mancha, etc.-, y en dos talleres experimentales realizados con ocasión del Festival de Teatro clásico de Almagro (en 2018 y 2019).

En febrero de 2018, se celebraron las terceras jornadas de OI2, en la sede de la Facultad de Comunicación de la Universidad CEU San Pablo, Madrid, que se titularon *De espectadores pasivos a ciudadanos activos*. En ellas, se plantearon los retos que la participación ciudadana representaba para la innovación de la información televisiva, tanto desde el punto de vista profesional como social y tecnológico. De esta forma, se cierra el primer ciclo de investigación de OI2 centrado en considerar las posibilidades genéricas de innovación en materia de informativos con especial énfasis en el periodismo móvil y la participación.

En el siguiente ciclo, el *leitmotiv* de las diversas actividades será el de la inteligencia artificial aplicada al periodismo.

La inteligencia artificial y el periodismo

Las IV Jornadas de OI2, que tuvieron lugar en Barcelona el 21 de noviembre de 2018, marcan, pues, un punto de inflexión. Se sigue focalizando la innovación periodística en relación con la evolución tecnológica, pero se selecciona ya lo que se considera un vector esencial de esta evolución: el desarrollo de la inteligencia artificial.

Las IV jornadas sirven para poner en relación periodistas, expertos en inteligencia artificial, e investigadores, enfrentados, todos, a los retos que la inteligencia artificial plantea al periodismo audiovisual. Se trata de diseñar y consensuar un mínimo mapa conceptual que permita identificar

3 (2018) Madrid, IORTVE.

las grandes áreas de transformación en el citado campo, líneas que habrían de orientar la investigación futura.

Fruto de estas jornadas, RTVE y la UAB logran establecer un nuevo plan de actividades -para el Observatorio y la Cátedra- para los años 2019-2020 en materia de inteligencia artificial y periodismo. Dicho plan se manifiesta en la renovación de los acuerdos entre RTVE y la UAB para un nuevo período de colaboración que llegará hasta el curso 2020-2021.

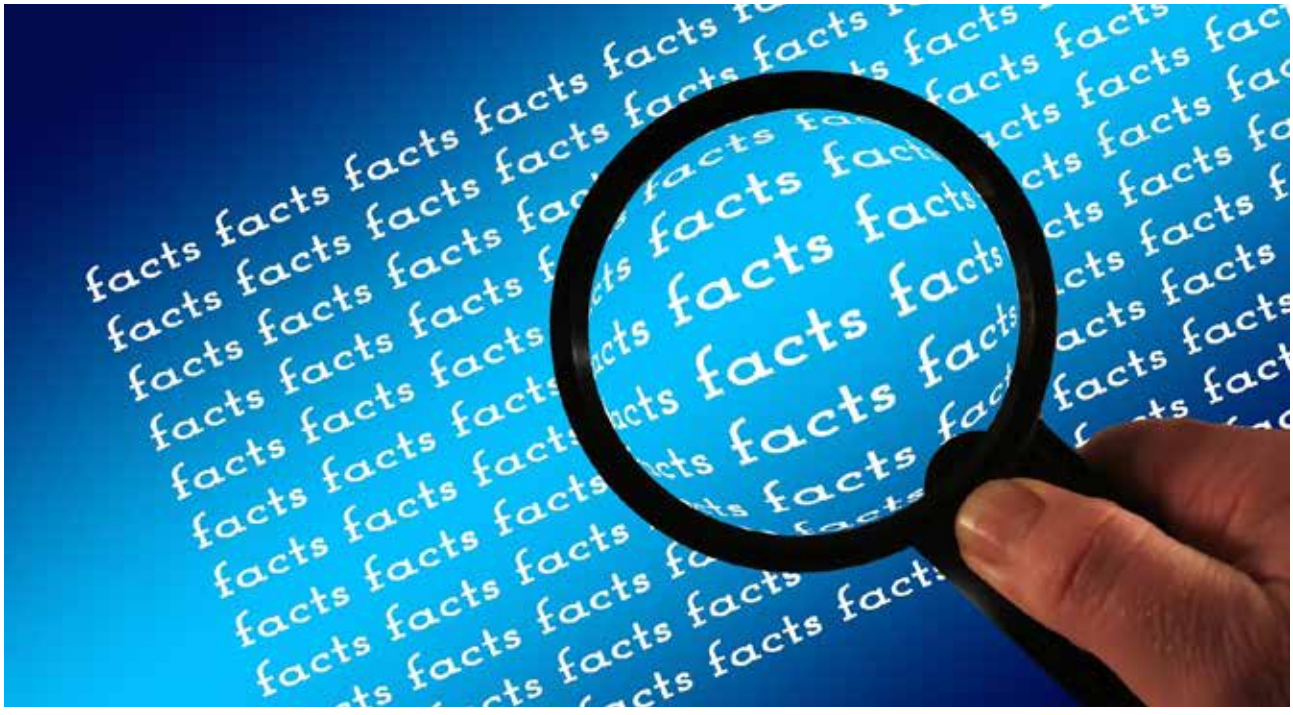
El nuevo plan de investigación diseñado sobre periodismo e inteligencia artificial tiene los siguientes objetivos:

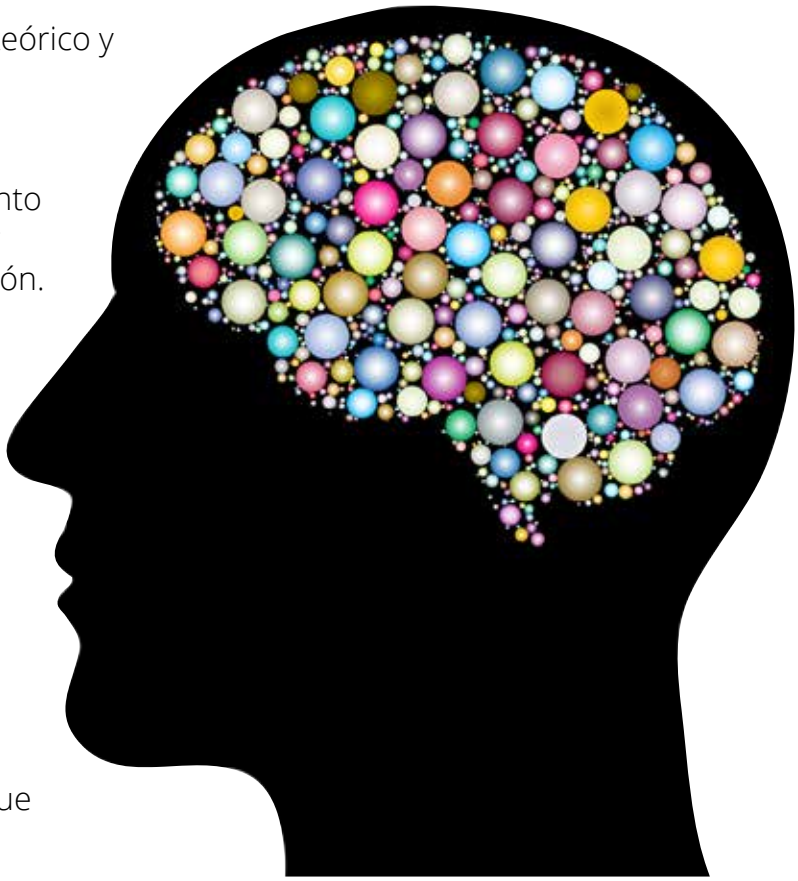
1. Conocer, describir y analizar el panorama actual de las diversas aplicaciones de la inteligencia artificial en el periodismo.

2. Identificar el modo más eficiente cómo estas aplicaciones pueden incorporarse, con éxito, a las redacciones audiovisuales.

3. Estudiar el impacto de la inteligencia artificial en las redacciones periodísticas, reconociendo tanto las mejoras como los inconvenientes.

4. Finalmente, establecer un mapa sobre las tendencias de futuro en el sector que permita su comprensión y el desarrollo de tareas de innovación y mejora.





En este marco y desde el punto de vista teórico y práctico, los resultados esperados son:

A. La adopción de un mínimo cuadro conceptual que permita un conocimiento adecuado del sector y permita diseñar estrategias de innovación y de formación.

B. La identificación y caracterización de las grandes áreas de aplicación empírica de la inteligencia artificial en cada uno de las áreas del periodismo y el análisis de su implementación en estrategias de innovación.

C. La identificación de las principales herramientas tecnológicas disponibles en el mercado, así como el diseño de un cuadro de criterios e indicadores que nos permita valorarlas.

D. El estudio de casos concreto en las que las herramientas de inteligencia artificial son usadas con éxito dentro de las organizaciones periodísticas.

E. El diseño de proceso de experimentación y desarrollo de tecnologías de inteligencia artificial dentro de RTVE y, en general, dentro de los servicios públicos audiovisuales.

El plan de investigación sobre periodismo e inteligencia artificial parte del reconocimiento de las diversas fases de producción y difusión de la información televisiva: desde el momento que se produce un hecho noticiable, hasta el instante en que este -en forma de texto audiovisual- se publica o se difunde en un determinado canal o medio; y, posteriormente, el tiempo en que dicha información relativa se archiva de cara a su reutilización -tanto por parte del periodista como el espectador-.

Así, el plan de investigación previsto distingue las siguientes fases o etapas. En cada una de ellas se abordan diferentes temas, en función de su propia naturaleza; pero en todas ellas la atención se pone en la utilización de la inteligencia artificial y en su impacto en las organizaciones periodísticas. Las fases son las que siguen:

1. **Fase de detección del hecho noticioso**: desde que se produce un hecho noticiable hasta que el periodista o la redacción es consciente de lo sucedido y decide cubrir la noticia.

2. **Fase de recopilación de la información**: momento en que el periodista -o la organización periodística- ante un hecho noticioso ya identificado, inicia las tareas de recopilación de información y la obtención de imágenes, audios o testimonios directos.

3. **Fase de escritura y elaboración de la información audiovisual**: fase de la elaboración de la noticia en sí, en la que el periodista utilizando la información audiovisual disponible elabora su pieza periodística.

4. **Fase de publicación y difusión**: momento en que las piezas periodísticas se publican en diferentes canales y soportes.

5. **Fase de archivo**: el momento en que , una vez finalizado todo el proceso comunicativo, la información se archiva para posteriores usos.

6. **Fase de recepción y uso**: el estadio en el que el usuario utiliza o reutiliza -y valora- la información audiovisual producida, publicada y archivada. Es en esta fase donde se puede estudiar cuestiones como la actitud y valoración de la audiencia sobre los contenidos y las diferentes formas en las que puede participar en la producción o recreación de tales contenidos.

Insistimos, en cada una de estas seis fases, se trata siempre de describir el tipo de ayuda y el empoderamiento que -tanto al periodista como a los espectadores- puede prestar la inteligencia artificial. Y en todas y cada una de ellas, se trata de poner de relieve los siguientes aspectos:

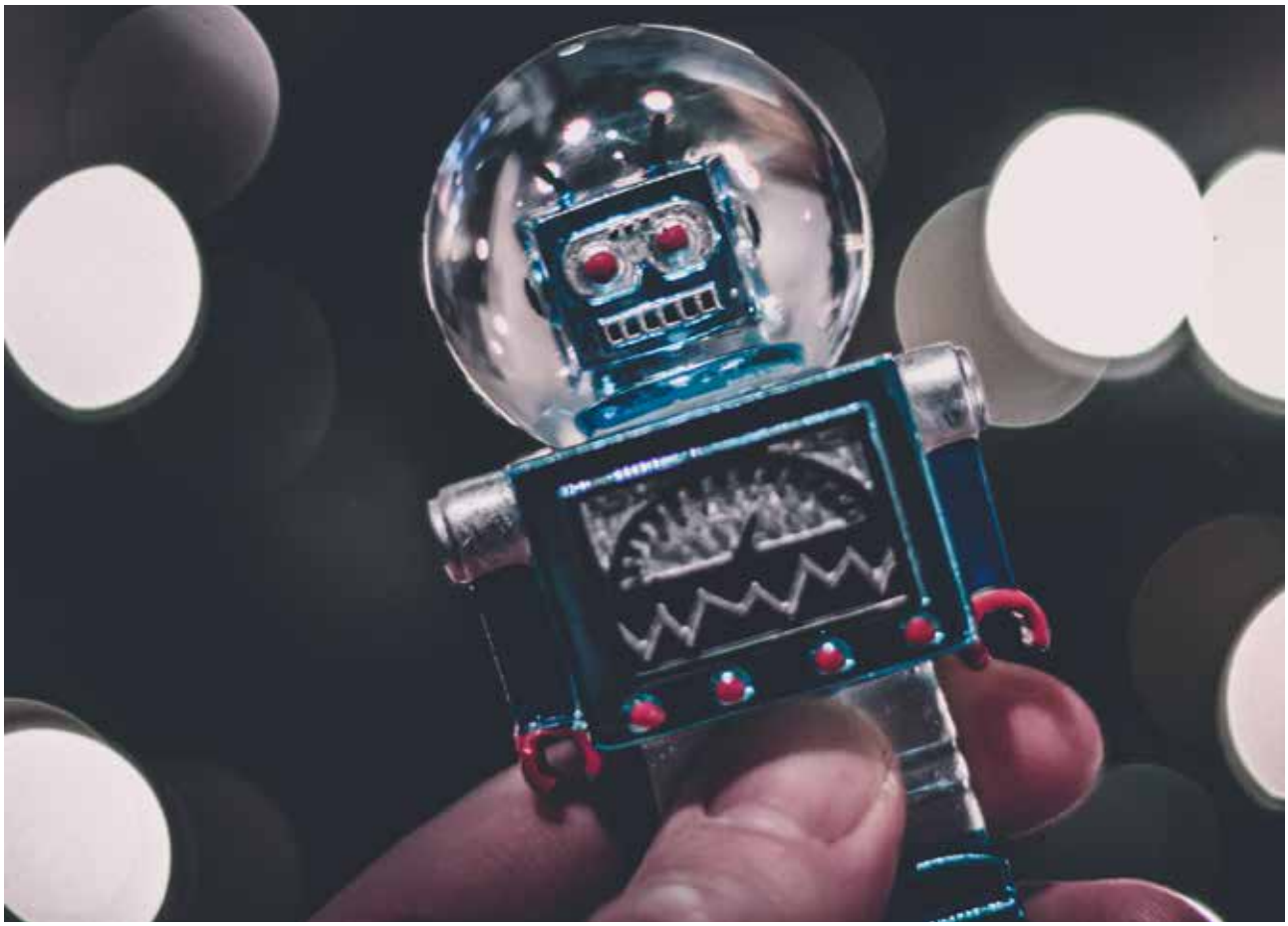
- El avance del conocimiento científico sobre la cuestión
- El desarrollo industrial existente
- Las tendencias y perspectivas de futuro, tanto desde el punto de vista tecnológico como profesional
- Las consecuencias periodísticas y sociales que la aplicación de la inteligencia artificial puede tener en el futuro inmediato

Detección de noticias

El presente informe recoge los primeros avances realizados sobre la fase 1, es decir, la fase de detección del hecho noticioso. Corresponde con el momento en el que una organización periodística o un periodista explora su entorno y utiliza sus fuentes de información para la búsqueda de hechos y acontecimientos que -previstos o no previstos- son susceptibles de convertirse en piezas (textos de naturaleza diversa) que pueden interesar a su público. En esta fase, los medios periodísticos disponen de cauces habituales que les proporcionan información, de alguna manera previsible, pero, también, deben estar atentos a su entorno para poder captar aquellos otros acontecimientos que tiene un carácter tan singular que no pueden preverse y atenderse con precisión.

La pregunta de investigación que motiva este informe es la siguiente: ¿puede ser útil la inteligencia artificial a la hora de proporcionar a los periodistas indicios e informaciones sobre hechos singulares que no están previstos de antemano? O, dicho de otra manera, ¿pueden las máquinas y los ordenadores, con su enorme capacidad de procesar muchos datos a mucha velocidad, proporcionar información fiable a los periodistas que les permita a estos identificar hechos noticiosos?

Este informe trata de responder a estas preguntas.



La inteligencia artificial en el periodismo

Antes de iniciarnos en el tema específico que nos ocupa en este informe, es necesario recopilar algunos datos clave sobre la inteligencia artificial aplicada en el periodismo, e introducir algunos de los términos más empleados en el campo. A esto dedicaremos este apartado.

Future Today Institute (FTI, en adelante)[4] anotaba en su informe que la inteligencia artificial no se debe considerar, a

4 Newman,N. (2019). 2019 Industry Trends: Journalism, Media, Technology. Acceso: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-01/Newman_Predictions_2019_FINAL_2.pdf

estas alturas, como una simple tendencia tecnológica sino que debe considerarse como la forma que adopta la tercera era de la computación. Es decir, como un desarrollo insoslayable que afectará a todos os campos de la actividad humana, tal y como ha afectado ya a la digitalización y a la informática.

Y, sin embargo, tal y como anotan los autores del informe, sigue existiendo un enorme debate, cargado tanto de optimismo como de temor, sobre lo que la inteligencia artificial (IA, en adelante) puede aportar. Pero es un debate en el que no acaba de aceptarse que la inteligencia artificial ha venido ya para quedarse, y que, por tanto, hemos de contar siempre con ella.

Por lo que hace al campo del periodismo, el mismo informe, así como otros muchos, afirman que el uso de la inteligencia artificial tiene impactos tanto positivos como negativos[5].

Por un lado, la IA ofrece y ofrecerá, herramientas que permiten a los periodistas realizar su trabajo de manera más eficiente, y generar nuevos conocimientos a partir de la búsqueda y análisis de datos; que, a su vez, sirven para la personalización de contenidos y para la adaptación a las necesidades de los diversos públicos.

Pero, por otro lado, se reconoce también que la inteligencia artificial puede ser, una herramienta al servicio de la desinformación, e incluso, para la producción y la difusión de información sesgada y falseada. Lo cual, como se sabe, representa uno de los riesgos más preocupantes para la salud de la esfera pública actual.

5 Bollier, D. (2017). Artificial Intelligence: *The Promise and Challenge of Integrating AI Into Cars , Healthcare and Journalism*. Maryland: The Aspen Institute.

En este sentido, la investigación desarrollada en esta fase de OI2 pretenden, como se decía al principio, contribuir a avanzar en el conocimiento del trabajo en esta específica área, y contribuir a prevenir los riesgos y a explotar las oportunidades.

Los bots

Uno de los indicios de la intersección entre periodismo e inteligencia artificial es la aparición del término *bot*.

La palabra *bot* forma ya parte de nuestro vocabulario, según Webb (2018) y está referida a aplicaciones muy básicas, que consisten en automatizar una sola tarea, así como a aplicaciones complejas en los que las máquinas desarrollan tareas mucho más sofisticadas.

Tal y como señala FTI, hay dos tipos de bots: los aplicados a la producción de noticias y los aplicados a aumentar la productividad.

Los primeros ayudan a agregar y alertar automáticamente a un usuario sobre un evento específico, tema que, en este primer informe, vamos a trabajar más a fondo.

Por otra parte, los bots de productividad son herramientas que las organizaciones de periodismo deberían usar para ayudar a automatizar y optimizar sus operaciones diarias. El riesgo depende particularmente de quién lo diseña, cómo y para qué. Por ello, la transparencia de estos procesos, y el diálogo entre profesionales de los diferente perfiles académicos y profesionales es necesario.

En este sentido, en la Universidad de Texas se está ayudando tanto a estudiantes como a profesionales a aprender a diseñar bots que se adapten a las necesidades que poseen. En el informe que presentamos, se incluye un estudio de caso específico sobre cómo desde RTVE conjuntamente con la Universidad Carlos III han desarrollado un sistema de alertas propio “Social Media Radar” para detectar hechos noticiosos.

Otro de los datos de interés señala que los usuarios más activos de noticias son 2.5 veces más proclives a utilizar sistemas de alertas en los móviles que en otros medios, y que, por tanto los medios deben incorporar de una forma más estratégica el usos de las alertas más allá de las breaking news, según el informe de Reuters Institute Digital News Report[6]. Como último dato, rescatamos uno de los esquemas visuales realizados por dos autores[7], un humano y un bot trabajando juntos en la elaboración de una guía de recomendaciones a incorporar en las redacciones ya mixtas donde humanos y bots trabajan juntos para ofrecer un servicio informativo de calidad.

Figura 1. Flujo de trabajo tradicional y con la integración de IA (Marconi & Journalist, 2017)

6 Newman, N. (2019). Journalism, Media and Technology Trends and Predictions 2019. Disponible en: http://www.digitalnewsreport.org/publications/2019/journalism-media-technology-trends-predictions-2019/

7 Marconi Alex, Journalist, Machine, F. S. (2017). A guide for newsrooms in the age of smart machines. *The Future of Augmented Jounralism*. Retrieved from https://insights.ap.org/uploads/images/the-future-of-augmented-journalism_ap-report.pdf

Sin lugar a duda, el poder analizar cómo se está utilizando paso a paso la inteligencia artificial en cada una de las fases claves de elaboración, publicación, difusión y archivo de noticias, ayudará a vislumbrar mejor los beneficios y las desventajas para diseñar las estrategias a incorporar en las redacciones u otros equipos de medios profesionales.

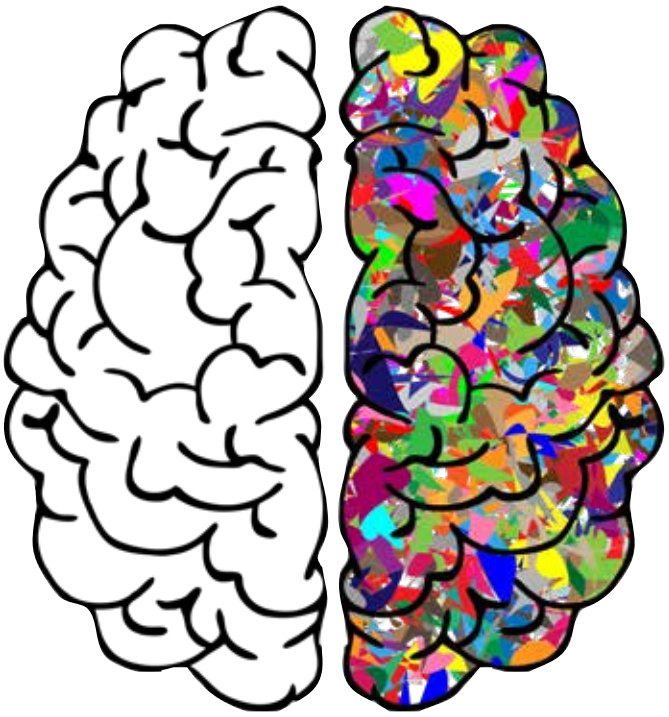
No es posible disponer de un modelo general que defina lo que puede considerarse noticia o no.

La detección e identificación de hechos noticiosos en periodismo

Lo que se llama detección o identificación de noticias en periodismo no es un hecho simple: involucra infinidad de actores y tareas. Y no solo responde a un proceso o protocolo único, sino que tiene que ver con una casuística singular y contextual.

En este sentido, no es posible disponer de un modelo general que defina lo que puede considerarse noticia o no. En cada situación y en cada momento, un periodista o un medio periodístico puede utilizar procedimientos y criterios diferentes para seleccionar los hechos noticiosos y mostrarlos a su público.

No obstante, la existencia de una comunidad periodística, de una tradición casi secular de trabajo periodístico, nos permiten establecer un cierto modelo canónico o teórico para describir el proceso de selección de la información.



Trataremos pues, de exponer, en primer lugar, en qué consiste este proceso y qué elementos participan en él. Intentaremos, también, distinguir sus fases y sub-tareas, además de señalar los conceptos y términos que nos pueden ayudar a analizar y explicar el conjunto de actividades y elementos que intervienen en el citado proceso.

Desde el punto de vista de las ciencias de la información y del periodismo, la identificación de noticias se concibe como una actividad propia del trabajo periodístico -individual, colectivo o institucional- de observación, identificación y selección de hechos noticiables. Por tanto, es una tarea que involucra acciones físicas, pero sobre todo cognitivas, intelectuales.

En esta actividad y siempre desde el punto de vista del periodismo, se combinan agentes humanos, observaciones directas sobre la realidad, fuentes informativas de diversa naturaleza –testimonios, agencias de información, gabinetes de comunicación, bases de datos, instituciones, etc.-, e, incluso recientemente, procedimientos automatizados –búsquedas web, consulta a bases de datos, fuentes estadísticas, etc.

Los términos y conceptos más habituales que, dentro de la lógica de trabajo periodístico, se utiliza –aún sin demasiada precisión ni formalización- son los siguientes: hecho noticioso, fuentes informativas, testimonios, contraste de información, verificación, titular, noticia, *lead* o resumen de la noticia, interés periodístico, construcción de la notica, noticia publicable, actualidad, *breaking news*, etc.

A partir de ellos y tratando de encontrar un criterio más preciso de formalización, proponemos distinguir las siguientes fases y procesos:

a) Identificación de hechos noticiables entre un conjunto de hechos observados –directa o indirectamente- y aplicando criterios de noticiabilidad. Esta fase incluye: 1) El registro de datos en bruto (de la realidad o de fuentes informativas). 2) La selección –o filtrado- según la noticiabilidad. Estos valores, aunque responden a una cierta generalidad, dependen del medio y de la ocasión.

La identificación de noticias es vista como una actividad propia del trabajo periodístico –individual, colectivo o institucional- de observación, identificación y selección de hechos noticiables.

b) Selección o filtrado según los criterios y valores de publicación. Estos son más estrictos y restringidos que los criterios de noticiabilidad, porque en ellos entran un sinfín de consideraciones añadidas a los de noticiabilidad, tales como es espacio disponible, el alcance del medio, la disponibilidad de documentos y su tipo, factores contextuales y temporales, etc.

c) Procesamiento narrativo del hecho seleccionado como noticiable o publicable. Aunque esta fase pueda incluirse en lo que consideramos fas de escritura periodística, hay que asumir que una cierta escritura es inevitable y forma parte consustancial del proceso de detección de hechos noticiosos.

La identificación de eventos en informática

Desde el punto de vista de la ingeniería informática, el concepto de identificación de noticias no se utiliza habitualmente. Sin embargo, existen términos muy próximos que pueden hacer operativo el concepto de noticia.

De hecho, en lo que se conoce con el nombre de minería de datos, los informáticos distinguen entre **datos** –o registros- y **patrones** (tendencias, relaciones, reglas que permiten describir el comportamiento de los datos). En cuanto a ámbitos como la web y redes sociales, se usa el término de detección de eventos para hacer referencia a la tarea de descubrir patrones (Capdevila *et al.* 2017). A partir de la detección de eventos se pueden identificar **eventos de la vida real.**

En este sentido, una noticia sería, desde el punto de vista informático, el resultado de procesar y ordenar, en lenguaje natural, algunos datos que se relacionarían con eventos de la vida real.

El concepto de evento -siempre en el campo de la minería de datos- se aplica a una agrupación de datos suficientemente significativa como para considerarla con autonomía propia. Cuando se habla de evento de la vida real se está suponiendo:
a) Que es un hecho que se ha dado en el plano de la realidad cuya existencia podemos suponer -como un *a priori* o, en otro caso, a partir de algunos datos que consideramos iniciales-.
b) Que es un hecho de que, por definición, no disponemos de todos los datos o, como mínimo, del cual no disponemos de una estructura de datos que nos permita reconocerlo como hecho unívocamente –aunque esta situación pueda ser temporal-.

Así pues, los elementos funcionales para la identificación de eventos en la perspectiva de la informática son, los datos –su disponibilidad y su carácter-, las reglas de selección y agrupación de estos datos –en función de atributos y categorías- los ordenadores o procesadores de información, los programas, y, en su caso, los sujetos humanos implicados –en la programación o en el entrenamiento de las máquinas, etc.-.

Por otro lado, estas entidades –en la perspectiva de la informática- organizan su actividad según los siguientes procesos:

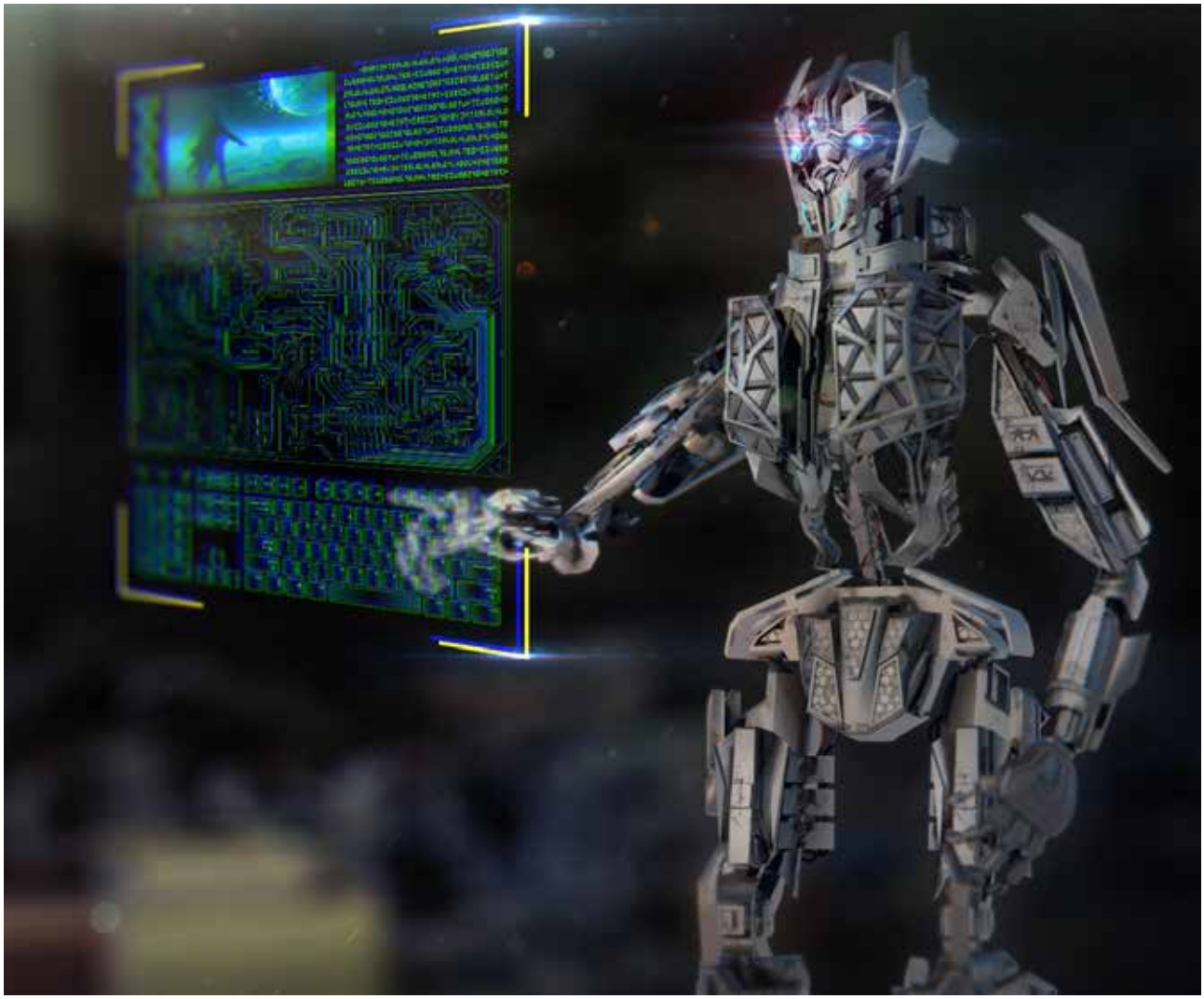
a) Recogida de datos.

b) Caracterización de los datos según categorías predeterminadas.

c) Selección de los datos considerados pertinentes.

d) Organización, clasificación y agrupación de tales datos según criterios establecidos.

e) Procesamiento de los resultados obtenidos para la programación de nuevas tareas.



Ingeniería de datos e identificación de hechos noticiosos

¿De qué manera puede ayudar la ingeniería o minería de datos a la tarea periodística de detección de hechos noticiosos y publicables?

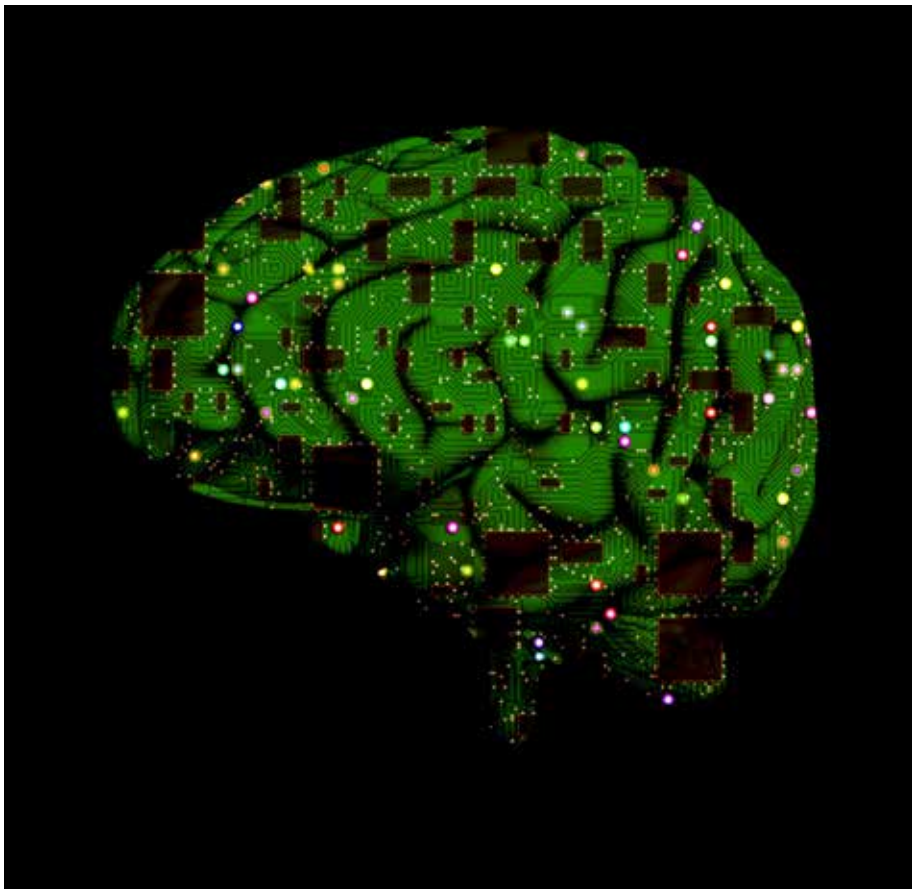
Para responder a esta pregunta, deberíamos tratar de hacer corresponder, con precisión, los conceptos utilizados en el trabajo periodístico con los propios de la ingeniería de datos. Y a partir de esta correspondencia se podrían crear programas y diseñar procesos que faciliten algunas tareas de los periodistas.

Desde nuestro punto de vista, el siguiente cuadro pone en relación los modelos conceptuales del periodismo y la ingeniería de datos en relación con la detección de hechos y permite un trabajo conjunto.

Periodismo/ hechos noticiosos	Ingeniería de datos/ eventos
Identificación de hechos noticiosos: • Obtención (directa e indirecta de datos) • Uso de Categorías • Uso de Criterios de selección: – de noticiabilidad – de publicación Identificación de noticias publicables • Uso de Categorías • Uso de Criterios de selección: – de noticiabilidad – de publicación	Recogida de datos (no estructurados) Caracterización de los datos (según categorías) Selección según categorías (de noticiabilidad y de publicación). Pueden preverse selecciones sucesivas y recursivas. Agrupación estructural (categorías y clústeres). Pueden preverse agrupaciones sucesivas y recursivas.
Escritura periodística	Procesamiento narrativo

A partir de este cuadro, pueden estudiarse y diseñarse diferentes estrategias de aplicación de la ingeniería de datos y de la inteligencia artificial a la detección de hechos noticiosos.

En los apartados siguientes estudiaremos lo que se ha investigado sobre el tema y mostraremos algunos casos concretos de aplicación, así como algunos ejemplos de las herramientas que se comercializan en este campo.



La investigación académica sobre detección automática de hechos noticiosos

Ya hemos señalado, que la investigación científica y tecnológica de la inteligencia artificial al periodismo es aún incipiente. Pero lo es mucho más cuando se trata de la relación entre la IA y los procesos de identificación y detección de hechos noticiosos.

Hemos tratado de identificar, con las bases de datos a nuestro alcance (publicaciones indexadas en la Web of Science y en Scopus) lo que se ha experimentado y publicado sobre la cuestión. Hemos aplicado diversos criterios de búsqueda y en ellos han destacado las siguientes palabras clave: *Artificial Intelligence*; *Journalism*; *Machine learning; Deep Learning; Bots; Topic extraction; Events detection;* entre otras.

De este modo, hemos logrado identificar, en conjunto, no más de unas cien publicaciones generales sobre la cuestión y –sobre el tema de la detección de hechos noticiosos- solo una media docena de publicaciones[8], pero ninguna de ellas directamente relacionadas con el campo del

8 Gu, Y. et al. Detecting Hot Events from Web Search Logs. Conference Proceedings - Web-Age Information Management (2010).

Lee, C. H., Chien, T. F. & Yang, H. C. An automatic topic ranking approach for event detection on microblogging messages. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*. 1358–1363 (2011).

Guille, A. & Favre, C. Event detection, tracking, and visualization in Twitter: a mention-anomaly-based approach. *Social Network Analysis and Mining*. 5, 1–18 (2015).

Hua, T., Chen, F., Zhao, L., Lu, C. T. & Ramakrishnan, N. Automatic targeted-domain spatiotemporal event detection in twitter. *Geoinformatica*. 20, 765–795 (2016).

Dashdorj, Z., Tsogtbaatar, B., Tumurchudur, A. & Altangerel, E. High Level Event Identification in Social Media. *Proceedings - 12th International Conference on Semantics, Knowledge and Grids, SKG* (2016).

Capdevila, J., Cerquides, J. & Torres, J. Event Detection in Location-Based Social Networks. Book - Data Science and Big *Data: An Environment of Computational Intelligence* (2017).

periodismo. Ninguna de ellas habla, expresamente, de noticias periodísticas o de hechos noticiosos susceptibles de ser recogidos por la información periodística. No obstante, la media docena de publicaciones seleccionadas o bien, indirectamente, aborda en parte el tema de las noticias, o bien nos proporcionan conceptos o metodologías que pueden ayudar a desarrollar nuestro campo de investigación. Es en este sentido en el que la incorporamos a nuestra reflexión.

El sistema de detección de eventos

Las experiencias recogidas en las publicaciones seleccionadas tienen en común algunos rasgos interesantes:

Los datos son considerados como indicios del mundo real.

1. Todas ellas utilizan la noción de *evento* para referirse a un acontecimiento que tiene lugar en las redes sociales, por tanto, a un hecho discursivo, no físico.

2. Algunas de ellas utilizan, por otro lado, el término *evento del mundo real* para remitirnos al plano de los hechos cuya existencia, con ciertas cautelas, los datos recogidos pueden permitir suponer. En este sentido, los datos son considerados como indicios del mundo real (na pista o una prueba de un acontecimiento real, eso sí referido por el discurso de las redes sociales).

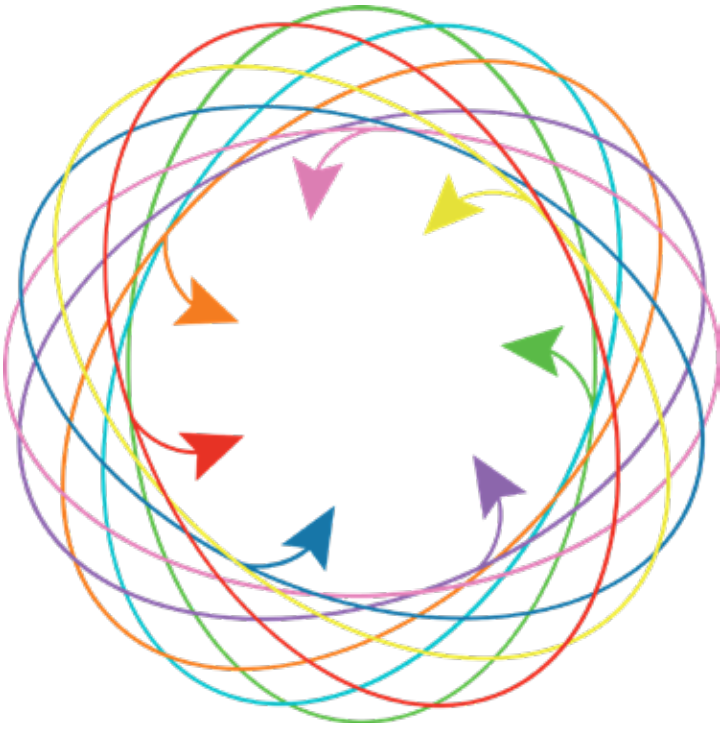
3. Finalmente, todas ellas tratan de establecer métodos, categorías y protocolos para organizar la información en bruto o datos no estructurados que proporcionan las búsquedas realizadas a) en la web –y con más precisión en algunos archivos log, o la proporcionada por las redes sociales y de que registran los *clicks* a la red efectuados por los usuarios-. B) Los microblogs, especialmente Twitter; siempre con el objeto de poder identificar automáticamente eventos significativos.

En general, pues, estas investigaciones adoptan un enfoque meta-discursivo –un lenguaje que se refiere a otro lenguaje-. El discurso objeto es, generalmente, proporcionado por las búsquedas en la web o por los mensajes de las redes sociales. Por tanto, parten de la idea o del presupuesto de que los hechos noticiosos pueden ser conocidos -al menos algunos de ellos- mediante la recolección y el análisis de datos provenientes de la Red.

En cuanto al meta-discurso, estas investigaciones utilizan categorías y algoritmos que pueden ser aplicados a algunos de los datos que utilizan. Pero hay que advertir que estas categorizaciones no han tenido en cuenta ni las prácticas periodísticas, ni la terminología usada en el trabajo periodístico. En general incorporan categorías provenientes del lenguaje común, sin demasiada formalización. Lo mismo puede decirse, como derivada, del uso de criterios. Muchos de ellos se basan en axiomas o presuposiciones de pensamiento corriente, sin que hayan sido formalizadas a través de un procedimiento analítico expreso.

Por tanto, cuando en estos experimentos –y hay que advertir de ello- se aplica el procesamiento automático, no hay que esperar otra cosa que resultados condicionados por tales categorías y criterios y por el modo en que se han aplicado.

Sin embargo, nada de ello impide, desde nuestro punto de vista, obtener enseñanzas interesantes de cara a la creación e investigación de sistemas informáticos orientados a la detección de hechos noticiosos. Lo cual es tanto como admitir que se podrían aplicar metodologías y procesos parecidos a los aquí expuestos a la detección de eventos periodísticos. En todo caso, y dado el campo de investigación, sí que se advierte la necesidad de trabajar más específicamente a la hora de usar categorías, criterios, algoritmos, la necesidad de confrontación directa con el lenguaje y la práctica periodística.



Un método para detectar eventos significativos a través del análisis de las búsquedas en la Web

Gu *et al.* (2010) tratan de detectar noticias y eventos a partir de la web.

Los autores reconocen, a partir del análisis de publicaciones previas, tres modos en los que puede utilizarse la información proporcionada por la web:

- **Análisis de contenido textual de las páginas web.** Para lo cual se utilizan técnicas de procesamiento del lenguaje natural.
- **Análisis de datos de estructura** –o estructurales- de la Web. Lo que se analiza entonces es la propia estructura y links de las páginas webs analizadas.
- **Análisis del registro de log**[9], y, especialmente, de los *click-trough* contenidos en tales registros.

9 Un *log* es un documento que registra las actividades de comunicación o de transacción de un usuario mientras utiliza un determinado programa. En el caso de los buscadores web se genera un registro de las búsquedas realizadas durante una tarea de búsqueda.

En este caso, la opción que eligen los autores es estudiar exclusivamente los registros generados en las operaciones de búsqueda, y, más concretamente, los links utilizados en la navegación a lo largo de esa búsqueda. La razón que dan para esta decisión es que consideran que su procedimiento tiene varias ventajas:

a) Pueden partir de una información sencilla de obtener (en este caso, del servicio de buscador MSN).

b) Además, dicha información parece ser muy reveladora del interés y de la motivación de los usuarios a la hora de buscar información.

El primer argumento es indiscutible y, por tanto, puede ser aceptado sin reservas en el proceso de experimentación. Ahora bien, no se puede dejar de poner de relieve las dependencias y lagunas que esta información puede tener, desde el inicio. Como también habría que tomar en cuenta que cualquier sistema que trabaje con este tipo de información se halla en estrecha correlación con las decisiones adoptadas por el proveedor de los datos. Lo cual nos está indicando que sería lógico –en caso de incluir este tipo de desarrollos en el campo del periodismo- que se necesitara elementos de contraste y de verificación.

Por otra parte, y en lo que se refiere ya a aspectos teóricos y no meramente metodológicos, los autores sitúan su investigación en el marco de lo que ellos mismos denominan *Event Detection*. Y, dentro de este campo, en la orientación a la que denominan como *Topic Detection.*

Los autores sitúan su investigación en el marco de lo que ellos mismos denominan Event Detection.

Además, incorporan una distinción conceptual que debe tomarse en cuenta. Señalan la diferencia que existe entre a) la detección retrospectiva de eventos –hechos ya realizados- y b) nuevos eventos. Y optan por situarse en el primero de ambos campos.

Por tanto, por lo que se refiere a nuestro campo de interés, el periodismo, hemos de considerar que esta experiencia no serviría para mostrarnos un hecho noticioso recién producido, inminente, o en gestación. En este sentido, no nos valdría como sistema de alerta temprana o de detección rápida de hechos noticiosos. Sin embargo, la investigación tiene interés porque puede servir –en paralelo al uso de otras fuentes- para reforzar nuestra valoración de un hecho –en un sentido u otro-, o para verificar alguna valoración ya hecha por otro procedimiento. Y, en todo caso, sí que nos permite obtener conceptos y perspectivas metodológicas que serían –al menos aparentemente- trasladables a la detección temprana de hechos noticiables.

Categorías y criterios

Los autores proponen tres criterios que deben cumplir su sistema:

1. La eficacia (o efectividad);

2. La eficiencia y

3. La organización de los eventos detectados.

Para que el sistema resulte eficaz, tiene que conectar, según los autores con un hecho de la vida real que resulte interesante.

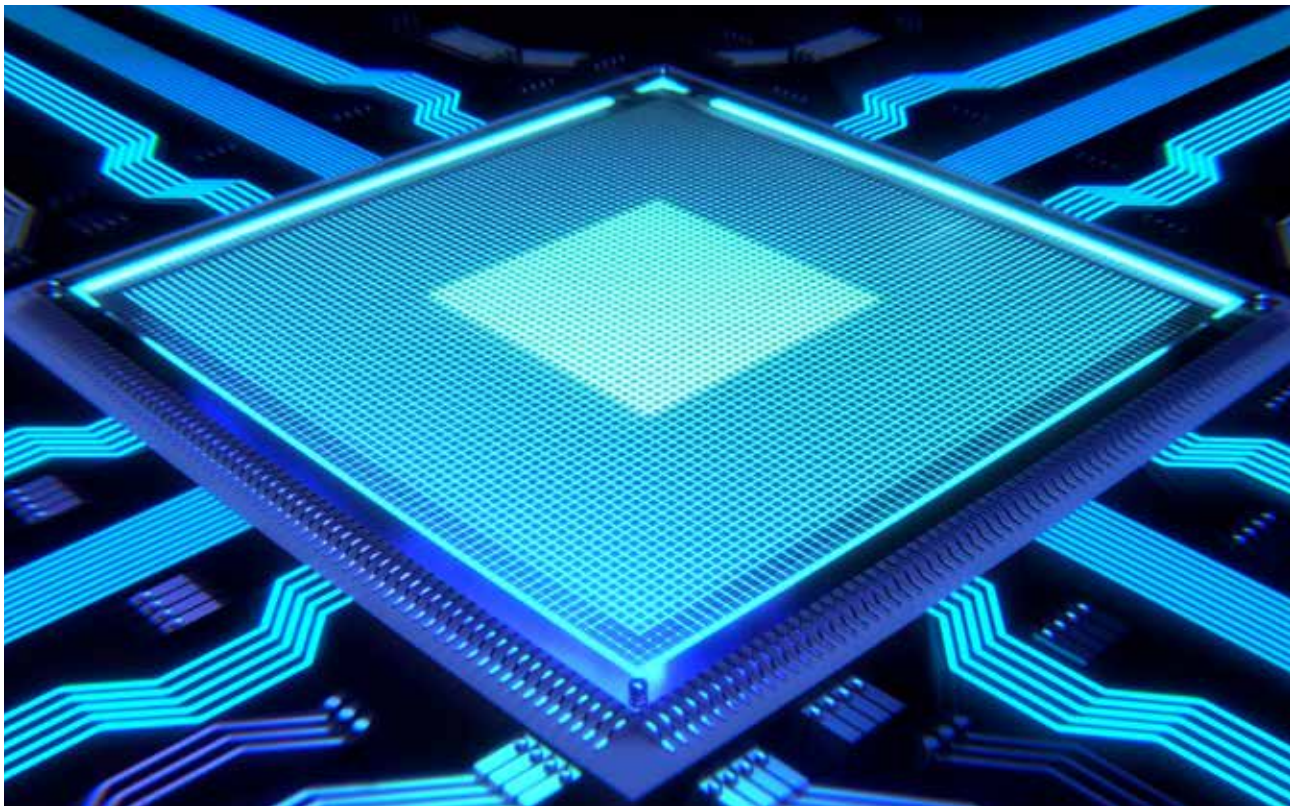
Para ello utilizan datos provenientes de tres categorías:

Información del enlace (que se indica mediante las consultas y las páginas correspondientes),;

a) Información del enlace (que se indica mediante las consultas y las páginas correspondientes),

b) Información temporal (tiempo) y

c) Contenido de la consulta (palabras clave de consulta), agrupan los datos en diferentes tópicos o temas.



A continuación, agrupan, en función de su similitud estos temas en clústeres hasta construir eventos.

Para realizar esta valoración o para aplicar estas a procedimientos de agrupación de resultados, utilizan una categoría interesante: el concepto *zona de explosión*. Es un periodo dentro de la secuencia de datos registrados donde el número de clics es mayor que el promedio del número de clics del resto de la secuencia de datos.

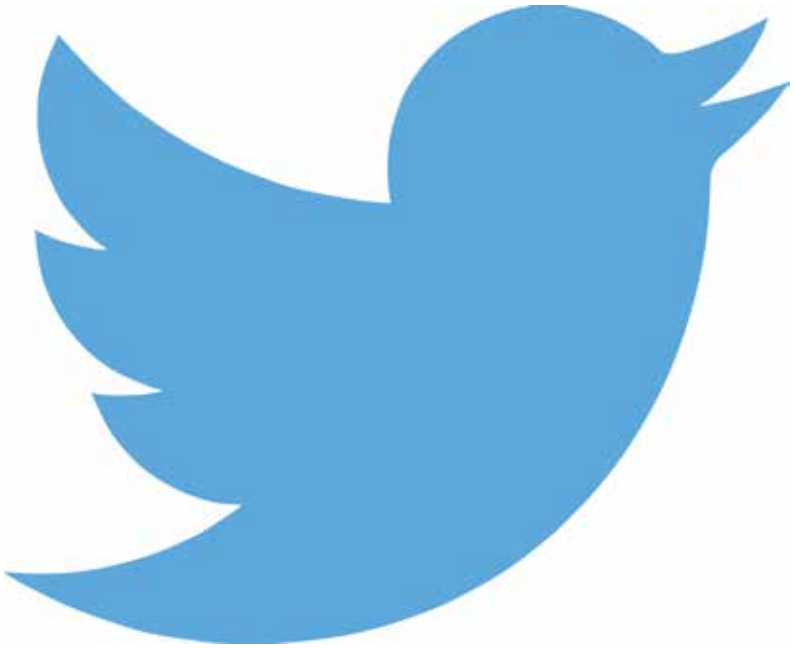
La práctica periodística también reconoce el valor de las zonas de explosión.

En términos generales, la práctica periodística también reconoce el valor de estas *zonas de explosión,* cuando utiliza criterios como las audiencias y su interés por determinados temas a la hora de seleccionar o publicar determinado tipo de noticias.

Desde nuestro punto de vista, el método propuesto por los autores, tiene validez y debe ser desarrollado y contrastado en otros casos y contextos. Pero, como ellos mismos proponen, seguramente debe de ser complementado con otras investigaciones que sirvan de contraste, tales como el análisis semántico del lenguaje natural de las páginas webs a las que se dirigen los *clicks* estudiados.

Análisis realizado sobre Twitter, eliminando los datos que aportan ruido

Aunque el análisis del registro de búsqueda web se ha revelado útil de cara a la detección de eventos, recientemente son las redes sociales las que han despertado más interés.



En general, se empieza a admitir que estas redes no solo se han convertido, por sí mismas, en fuentes de información que compiten con el periodismo, sino que, de alguna manera, resultan ser muy reveladoras sobre los intereses, motivaciones, gustos y tendencias del comportamiento de los usuarios. De aquí que pueden ser muy útiles cuando se utilizan para detectar hechos noticiosos.

En esta línea, Jagan *et al.* (2009) diseñan un sistema llamado **TwitterStand**[10] que se basa en una presuposición fundamental: "los usuarios geográficamente próximos a menudo twittean sobre las mismas noticias de última hora". Y, por eso se puede crear una especie de agregador de noticias –parecidos a los habituales- con el objetivo de que este pueda actuar como "nuestros ojos y nuestros oídos"; con la única diferencia que las fuentes utilizadas son siempre los datos proporcionados por Twitter.

Esta hipótesis afirma que los usuarios próximos twittean sobre a) las mismas noticias, y b) sobre las de última hora. Pese a su plausibilidad, deberíamos someterlas a examen y comprobación.

Lógicamente, los usuarios que tienen acceso a noticias de los medios convencionales –que se publican regularmente- tendrán tendencia a twittear en función de ese acceso reciente. Sin embargo, no hay que descartar casos en que una noticia o un tema de discusión reciente provenga de la reutilización de una noticia antigua o de la pervivencia de un tema que, aunque desfasado temporalmente, haya surgido por cualquier operación discursiva de algún grupo de usuarios. Por tanto, se abre aquí una línea de indagación y de experimentación que debe desarrollarse.

Por otra parte, los autores, utilizan el dato del tiempo o el criterio de temporalidad –a partir del concepto de última hora- para distinguir entre eventos pasados y actuales. Lo cual les acerca al concepto de actualidad en periodismo.

Lo esencial del sistema es distinguir lo que es noticia de lo que no lo es. Los autores manifiestan que el mejor sistema que han encontrado para hacerlo es identificar manualmente aquellos usuarios que suelen twittear noticias, creando así una especie de muestra inicial de las que pueden obtener un modelo para aplicar, con éxito, en gran cantidad de datos.

El principal objetivo del sistema es eliminar los ruidos –es decir, los contenidos que no son noticias- y determinar si el tweet es nuevo o no.

10 TwitterStand: News in tuits. *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*.

Para ello presentan el siguiente modelo que explica las diferentes funciones del sistema:

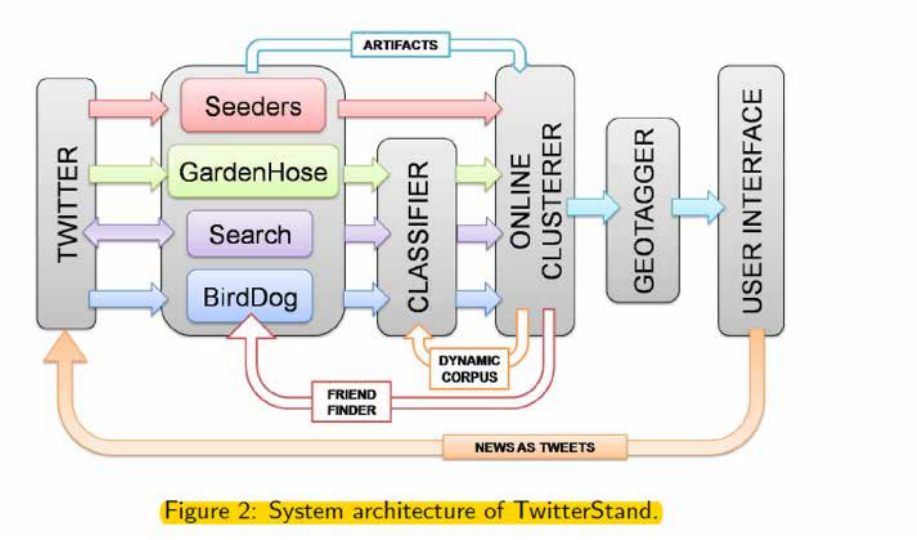


Figure 2: System architecture of TwitterStand.

A partir de los datos obtenidos directamente de Twitter, los autores establecen cuatro compartimentos en los que alojan datos según categorías.

Seeders representa una colección de usuarios de Twitter seleccionados específicamente y que están ligados a productores de noticias: diarios, revistas, blogs de noticias o información, emisoras de radio y TV, etc.

Garden House incluye una muestra de usuarios de Twitter sin ningún requisito específico.

BirdDog selecciona a usuarios de Twittter –has un número de dos mil- que se ha comprobado que habitualmente twittean noticias.

Search es el mecanismo que recupera activamente datos de Twitter.

El modelo incluye dos ámbitos que son utilizados en la categorización y clasificación de los datos. Por un lado, ***Friend Finder***, un buscador de *amigos* –con lo cual se introduce el factor social en el análisis-; y por otro, ***Artefacts*** que recoge todos los vídeos, imágenes o interactivos añadidos al texto del Tweet y que suelen ofrecer información complementaria interesante.

El modelo prevé un sistema de creación de clústeres para lograr una adecuada clasificación de los resultados, y un sistema de geolocalización capaz de filtrar y ordenar los datos.

Los autores reconocen que basándose en las experiencias de control de ruido, Lee *et al.* (2011) optimizan el rendimiento de la clasificación automática de temas y la evaluación de mensajes para identificar rápidamente el evento emergente que se define como una modalidad de localidad temporal, a saber, "un conjunto de mensajes que están altamente concentrados en un período de tiempo ".

También prestan atención al concepto de tema de energía "que abarca tres factores: a) **la popularidad; b) la explosión y c) la información**.

El enfoque de detección de noticias propuesto por los autores, como ellos mismos dicen, tiene la ventaja y la frescura de utilizar –a diferencia de lo que pasa con los agregadores de noticias convencionales- los textos producidos por personas usuarias, no por instituciones periodísticas o profesionales. Por otra parte, y en relación a otros sistemas que utilizan datos de Twitter Google Trends[11] y Twitter Search, Twitter Stand es automático y selecciona solo las noticias.

Sirve, sobre todo como una muestra del impacto de una determinada noticia en la opinión pública.

Desde nuestro punto de vista, el sistema y sus desarrollos posteriores, puede ser útil para detectar, y en su caso, reforzar la detección de noticias por parte de los periodistas, y sirve, sobre todo como una muestra del impacto de una determinada noticia en la opinión pública. Naturalmente, la posibilidad de analizar ese interés en relación con datos como la localización geográfica y algunas características de los usuarios, añade un interés mayor al método propuesto.

Dado el gran volumen de tuits sobre temas no relacionados, Guille et al. (2015) diseñan la detección de eventos basada en anomalías de mención (MABED) para mejorar la precisión y la solidez de la detección de eventos de noticias en presencia de contenido muy lleno de ruido de Twitter.

11 https://trends.google.com

Teniendo en cuenta no solo el contenido textual de los tuits sino también las prácticas mencionadas (por ejemplo, @nombre de usuario), **MABED detecta eventos** que se describen como (I) una combinación de una palabra principal y un conjunto de palabras relacionadas ponderadas, (II) tiempo y (III) la magnitud del impacto sobre la red mencionada.

Según los experimentos realizados, su enfoque produjo un mejor rendimiento al aprovechar la frecuencia de las menciones.

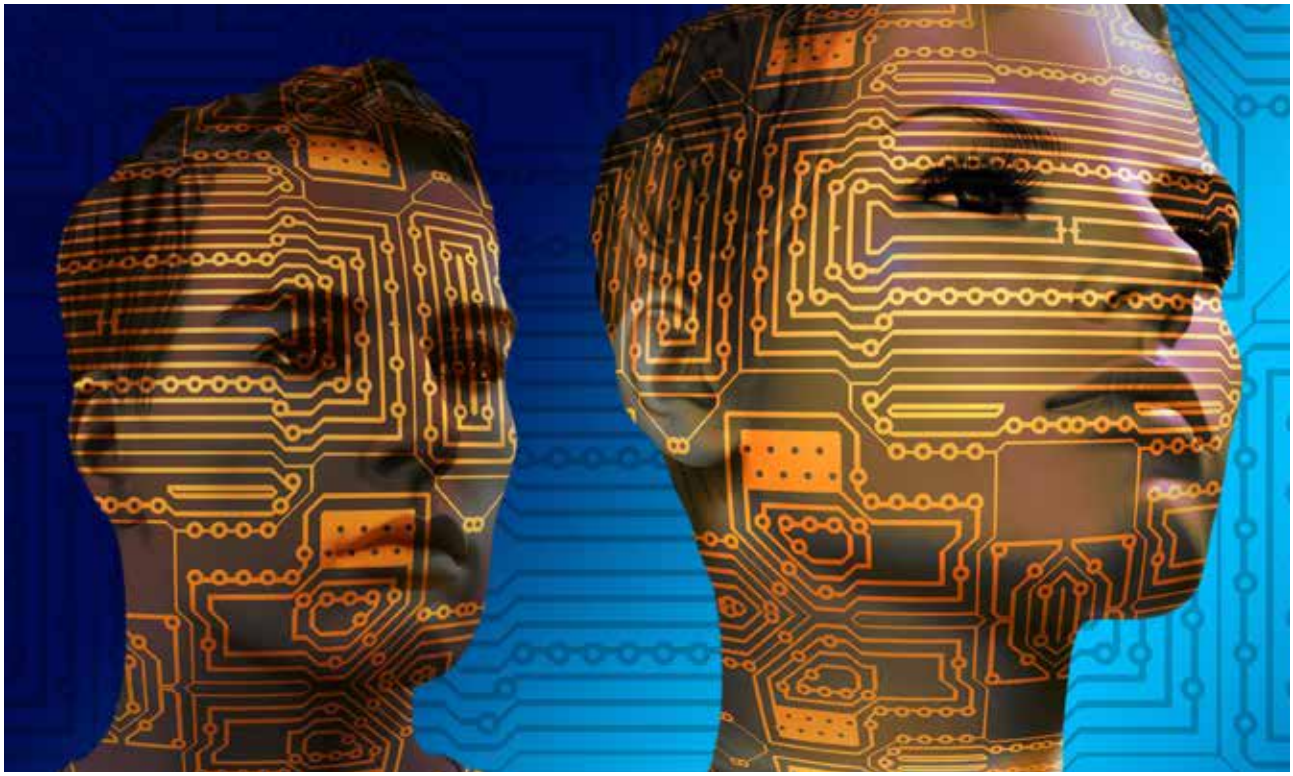
MABED está en la línea de los métodos propuestos con anterioridad, pero introduce algunas ventajas –acreditadas en investigaciones empíricas-. A saber:

a) Proporciona una mejor inteligibilidad de los textos producidos.

b) Refuerza la detección de noticias tomando en cuenta valores de tiempo flexible.

c) Utiliza datos sobre las redes de amistad en Twitter, con lo cual puede proporcionar información útil sobre el interés de los usuarios.

d) En el futuro, puede servir –combinado con otros datos- para medir el desarrollo de debates sociales en torno a noticias.



Análisis mediante aprendizaje supervisado

El artículo Hua *et al.* (2016) se centra en el contenido del "dominio dirigido" en Twitter; contenido que se relaciona con un tema / área específica dependiendo de lo que buscan los usuarios.

Por otro lado, los autores proponen utilizar –junto a la categoría de dominio- las de datos espacio-temporales. Para ello proponen una especie de escáner espacio-temporal, que "intentan construir un conjunto de datos y de etiquetas, de manera automática, apropiados y poder, así, utilizar estos datos generados automáticamente para detectar eventos espacio-temporales".

El método propuesto (Detección automatizada de eventos espacio-temporales en el dominio etiquetado, ATSED) se probó aplicándolo en 305 millones de tuits que muestran un mejor nivel de eficiencia en la detección de eventos en Twitter.

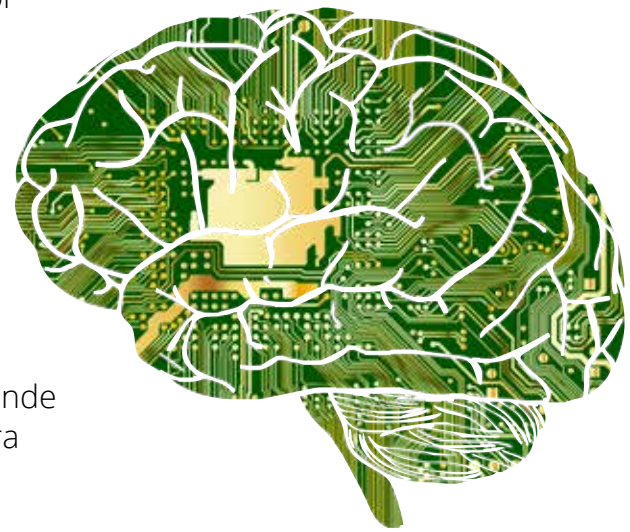
Los eventos espacio-temporales se refieren principalmente a eventos de la vida real que ocurren en un momento específico en un lugar determinado.

Según los autores, se realizan dos tareas principales:

a) Generación de etiquetas y

b) Detección espaciotemporal de eventos.

Hua *et al.* (2016) explican que "los datos históricos de Twitter y los artículos de noticias son utilizados por el componente de generación de etiquetas para producir pseudo etiquetas. Posteriormente, el módulo de detección de eventos espacio-temporales entrena al clasificador a través de estas etiquetas y detecta eventos de datos de Twitter en tiempo real".



Por lo tanto, la detección de eventos auténticos a partir de tuits depende de la generación automática de etiquetas con que la máquina aprende y, de este modo, se hallaría pronto capacitada para aplicar su aprendizaje a nuevos datos.

Después de agrupar los tuits relevantes para el tema específico buscado por el usuario (dominio de destino), se aplican "tecnologías de estimación de ubicación" como

siguiente paso para diferenciar los diferentes incidentes que ocurren al mismo tiempo.

Tres características principales contribuyen a la eficiencia del método utilizado (ATSED), según los autores, en la detección de eventos en Twitter:

a) La generación automática de etiquetas -que se considera más inclusiva que los procesos manuales-,

b) La capacidad del sistema para agrupar de manera eficiente los "tuits relacionados con eventos", y

c) La precisión mejorada que proporciona la consideración de la ubicación de los eventos.

Desde nuestro punto de vista, el interés de este enfoque radica en que incorpora un doble criterio para la detección de eventos. Uno el dominio o tema y el otro los datos temporales.

El uso de una orientación temática tiene la ventaja de que los periodistas –en la posibilidad de usar un sistema de este tipo- podrían dirigir el uso del programa en función de datos de agenda que consideran prioritarios. Pero la desventaja es que, por esa misma razón, el sistema no sería capaz de reconocer la emergencia de temas abruptos no previstos en la agenda. O, en todo caso, sólo podría recoger eventos previamente categorizados.

Respecto al uso de datos espacio-temporales, todos son ventajas, dado que –en general- los eventos noticiosos son de naturaleza fáctica y deben poder localizarse de algún modo. Además, estos datos pueden corresponderse, en todo caso, con los criterios de actualidad que el periodismo no puede dejar de tener en cuenta casi nunca.

Finalmente, el esfuerzo por diseñar sistemas de auto-aprendizaje semi-orientado parecen convenientes siempre, no sólo en relación con la eficiencia del sistema, sino, también, por el hecho de que tales métodos permiten la incorporación de los criterios y valores de los periodistas.

El artículo de Capdevila *et al.* (2017) presenta dos técnicas para la detección de eventos en Twitter: "una técnica de minería de datos llamada Tweet-SCAN y una técnica de aprendizaje automático llamada WARBLE" (Capdevila et al., 2017).

El citado artículo arroja luz sobre cómo identificar eventos a partir de los datos proporcionados por las redes sociales.

En primer lugar, los eventos son considerados hechos que ocurren en un momento y lugar específicos. Pero a esta característica los autores añaden un atributo, la importancia.

Teniendo esto en cuenta entienden que, a la hora de detectar eventos a través de las redes sociales, hay que considerar que el término evento debe referirse a "algo que causa un número anormal de acciones en el OSN" (Redes sociales en línea). En el caso concreto del cual se ocupa el artículo, para detectar cualquier incidente se trataría de tomar en cuenta el aumento o la disminución significativa de las acciones en Twitter.

Por otro lado, se parte de la constatación de que los "servicios de geoetiquetado" de la plataforma de redes sociales son más rápidos, a la hora de presentación de informes, que los utilizados por medios tradicionales. Y es así gracias a las referencias de ubicación con enfoque en Twitter. Por tanto, en su propuesta, el geotiquetado adquiere un papel fundamental.

Tweet-SCAN detecta principalmente la densidad de tuits tomando en cuenta tres características principales: tiempo, espacio y texto.

A continuación, el sistema agrupa "tuits estrechamente relacionados generados por un conjunto diverso de usuarios". Pero tomando siempre en cuenta que exista una considerable diversidad de usuarios que produzca tuits. Se evita así el error de considerar evento un hecho mencionado por un solo usuario que produzca un número elevado de tuits.

El sistema agrupa "tweets estrechamente relacionados generados por un conjunto diverso de usuarios".

Posteriormente el sistema se aplica al texto de los tuits identificando los términos clave.

Finalmente, el componente WARBLE detecta e identifica los tuits que están "relacionados con eventos", distinguiéndolos de la gran mayoría de tuits que "no está relacionada con eventos". Este procedimiento involucra mecanismos de aprendizaje automático (Capdevila et al., 2017).

El artículo de Dashdorj et al. (2016) tiene como objetivo principal evaluar la precisión de un enfoque introducido para detectar eventos en las redes sociales y en la Web.

Para evaluar el "Modelo de reconocimiento de eventos" EventMine, los autores utilizaron artículos de noticias de diferentes "fuentes web" y Facebook como plataforma de medios sociales.

El modelo introducido incluye cuatro componentes: Rastreador de artículos de noticias (artículos de filtro para "expresiones de eventos"; "fecha, lugar, nombre, tema", etc.); reconocedor de eventos, clasificación de eventos ("identificación de temas" y "agrupamiento de eventos") y visualización de eventos en mapas.

Sin embargo, los autores no explican claramente cómo funciona el método para detectar con éxito eventos en las redes sociales como se indica en el resumen.

Objetivos y métodos

A partir de nuestra lectura de los artículos académicos considerados, pueden obtenerse algunas conclusiones respecto a las perspectivas de detección de hechos noticiosos a través de las redes sociales y la Web:

- Existe una coincidencia generalizada en que los datos aportados por redes sociales pueden permitir la detección de eventos de la vida real. Y que el análisis de estos eventos puede facilitar dos tareas: a) La detección de hechos y su valoración; b) La comprensión de la dinámica de la vida social y la relación de las personas con respecto a determinados eventos.

- Los estudios realizados hasta ahora parten de datos que provienen de la web –y de los registros de búsquedas y clics-; del análisis de páginas webs de determinadas fuentes; datos de redes sociales como Facebook y otros; y, especialmente, datos provenientes de Twitter.

- En las aproximaciones consideradas se pueden señalar tres tipos de metodologías: Método basado en registro, Método basado en contenido y Método basado en estructura. El método basado en registro capitaliza los registros de búsqueda web (como el registro de búsqueda en Google). El método basado en contenido detecta eventos a través del análisis de información textual por medio del procesamiento de lenguaje natural. El método basado en la estructura, que también se denomina método basado en enlaces, utiliza estructuras de sitios web, estructuras de hipervínculos y/o estructuras de contacto social para detectar eventos.

- También se proponen métodos que combinan datos procedentes de diversas fuentes y registros, combinando y verificando resultados.

- Con respecto a los modelos y procedimientos propuestos, la mayoría de propuestas utiliza modelos de meta-datos, análisis de lenguaje natural, algoritmos, cálculos probabilísticos, creación de clústeres sucesivos, máquinas de aprendizaje supervisado. Y algunos de ellos incorporan tareas humanas en la selección de muestras o en el análisis de resultados. Finalmente, los muestreos varían entre los aleatorios y los determinados a partir de categorías previas.

- El problema fundamental que se presenta es cómo distinguir datos referidos a hechos o eventos de otros que no lo son. Cualquier sistema o programa que se plantee esta tarea se enfrenta al riesgo de confundir datos sobre hechos con otros que no lo son. Para evitar este riesgo son múltiples las estrategias que toman en cuenta factores como el análisis textual, datos de geolocalización, datos de intensidad de producción, etc.

- Se constata que, para la detección de eventos, la cuestión de la geolocalización de los datos es fundamental, porque, junto al análisis automático de texto, puede ayudar a la distinción de datos referidos a eventos de otros que no lo están.

- No obstante, el análisis de la geolocalización se convierte en un problema de difícil resolución en el momento en que existen muchos casos de ambigüedad y de confusión en los topónimos y en las denominaciones de localidades o territorios.

- La frecuencia e intensidad en la producción de datos referidos a un tema sirve, a la mayoría de los estudios, como un elemento que ayuda a identificar eventos.

- En algunos casos, los estudios proponen tomar en cuenta los enlaces de amistad y de relación que proporcionan las redes sociales. A partir de estos datos y en contraste con otros, se puede mejorar la precisión del sistema.

- En general, todas las propuestas analizadas consideran que es tan interesante detectar hechos como conocer las reacciones sociales que estos hechos suscitan en el público: interés, niveles de atención, reacciones emocionales, etc.

Los estudios proponen tomar en cuenta los enlaces de amistad y de relación que proporcionan las redes sociales.

Para facilitar la comprensión del estado de la metodología de detección de eventos y su evolución proponemos el siguiente gráfico.

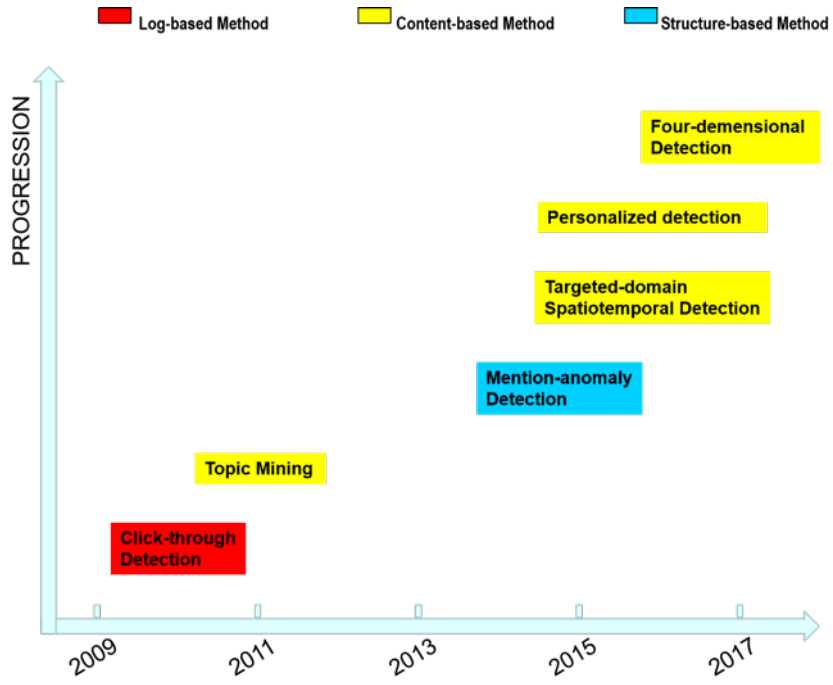


Figura. Detección de noticias

Como puede verse en él, la trayectoria académica de la detección de noticias evoluciona desde la detección de un solo factor -por ejemplo, palabras clave- a la detección de múltiples factores -por ejemplo, la composición de la localidad, la temporalidad, las palabras clave o el usuario personalizado-.

Específicamente, la detección de eventos de clics (Gu *et al*., 2010) es una técnica basada en registros –log- que incorpora información de enlaces, información temporal y de contenido de consultas.

Lee *et al.* (2011) utilizan la minería de temas, mediante la adopción y el desarrollo de varios algoritmos, como el algoritmo de modelo de ventana deslizante, para descubrir temas en tiempo real a partir de microblogs.

La detección basada en anomalías de mención (Guille et al., 2015) es un enfoque estadístico que se basa en los enlaces dinámicos que los usuarios insertan en los tuits para detectar eventos significativos y evaluar la magnitud del impacto en la comunidad.

La detección de eventos espacio-temporales de dominio dirigido (Hua et al, 2016) es una técnica semi-supervisada que

"aprende por primera vez las etiquetas de tuits a partir de datos históricos, y luego detecta eventos en curso de flujos de datos de Twitter en tiempo real" al tiempo que considera factores espaciotemporales.

Dashdorj *et al.* (2016) diseñan una detección de eventos personalizada en las redes sociales, que modela temas utilizando una asignación de Dirichlet latente (LDA) y personaliza los eventos por diferentes categorías, ubicaciones de usuarios y fechas.

La detección de eventos en cuatro dimensiones (Capdevila et al., 2017) emplea un algoritmo de aprendizaje automático y métodos probabilísticos para detectar eventos en redes sociales basados en la ubicación, con respecto a la dimensión de la ubicación, la dimensión del tiempo, la dimensión del texto y la dimensión del usuario.

Aplicaciones comerciales de detección de hechos noticiosos

A continuación, detallamos algunos de los sistemas de IA de alertas que se están utilizando por los medios y agencias de noticias en el panorama internacional. **En concreto, describimos los siguientes casos: *Dataminr, News Tracer, Newswhip Bertie, QuakeBot.***

Selección de los casos

Para determinar y categorizar la información de este apartado se creó una base de datos con los principales medios de comunicación que sirven de referencia globalmente. A partir de aquí, se identificaron diferentes casos sobre aplicaciones y herramientas que cada uno de ellos utiliza actualmente. Finalmente, se seleccionaron, de entre todos ellos, los que

más se acercaban a la funcionalidad de detección de hechos noticiosos. El resultado es el que sigue:

1. Dataminr
2. News Tracer, Reuters.
3. Newswhip, Associated Press
4. Bertie, Forbes
5. Quakebot, Los Ángeles Times

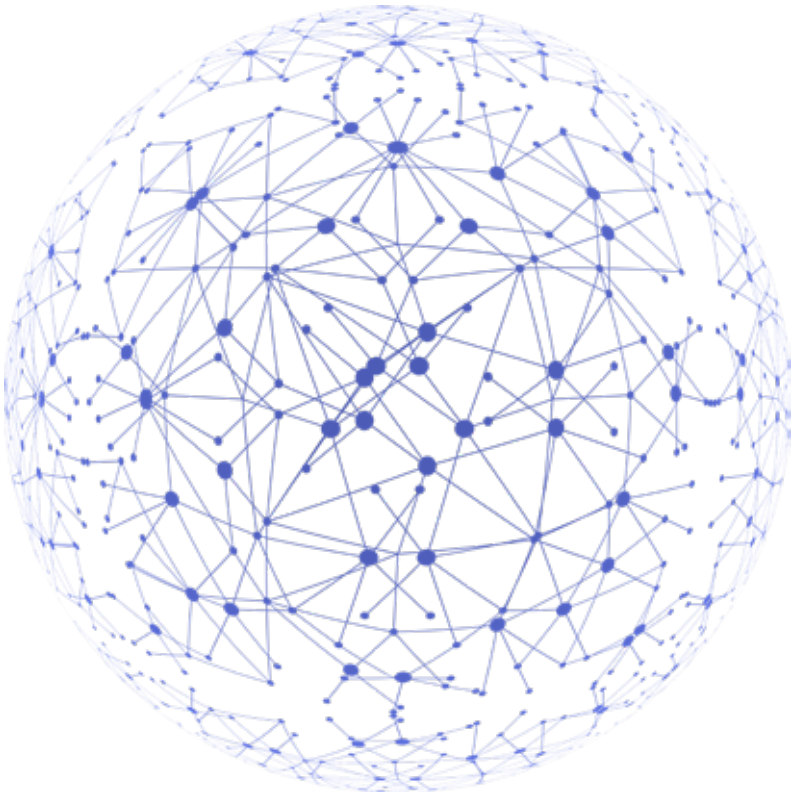
Las informaciones de cada caso seleccionado provienen de la fuente de la propia compañía.

Dataminr[12] es una compañía creada, en el año 2009, por Ted Bailey, como una empresa líder en análisis de datos y en la innovación del uso de inteligencia artificial para detectar y jerarquizar, en tiempo real, la información pública.

En realidad, los estudios de la compañía tiene a Twitter como fuente de datos básica.

Su metodología consiste en analizar los tuits que se van publicando hasta localizar un tema que es citado en gran cantidad y con mucha frecuencia –en sus términos, alcanza un determinado punto de calor-. Entonces, lo captura y lo ofrece a un equipo periodistas. En este sentido, funciona como una especie de alarma informativa que llama la atención sobre un hecho que, en principio, merecería la atención del mundo periodístico. Una vez la información es categorizada y verificada, se sirve al usuario como punto de partida para un trabajo posterior.

Una vez la información es categorizada y verificada, se sirve al usuario como punto de partida para un trabajo posterior.



12 https://www.dataminr.com/

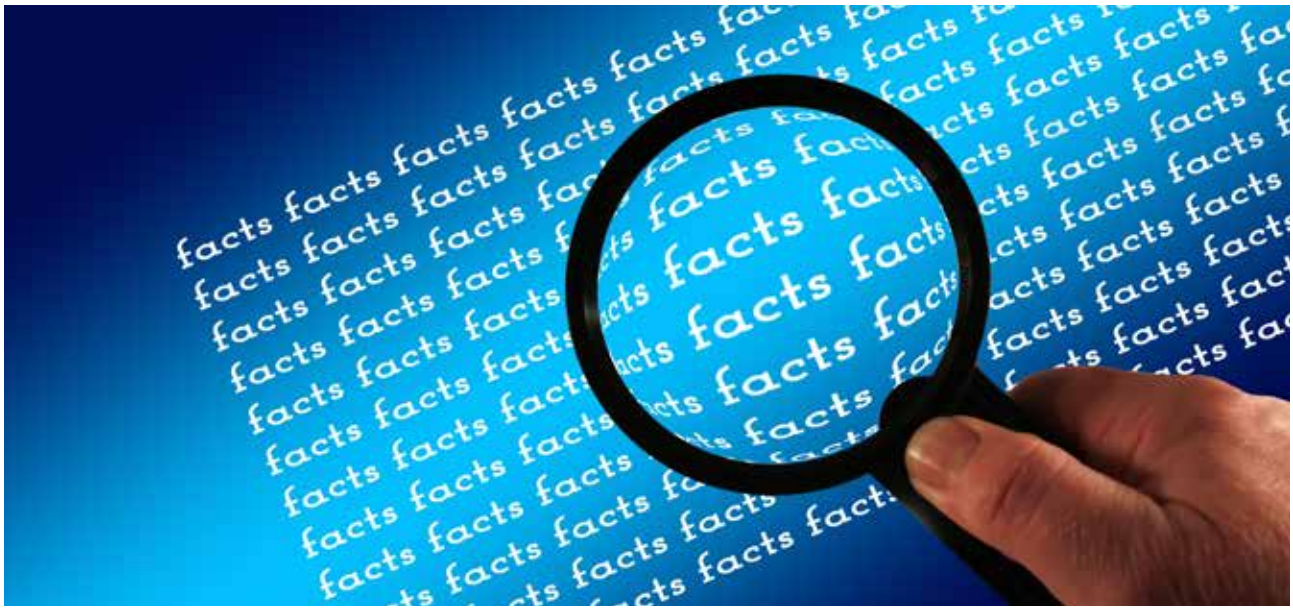
De este modo, los usuarios de la herramienta, pueden consultar la información aportada y analizar que está pasando; acceder a mapas; hacer seguimiento y contactar con la fuente directamente. Esto es importante porque incluso pueden interactuar las personas que están geocalizadas. Por otra parte, la información proporcionada puede contrastarse con fuentes oficiales y a personas que no dan información falsa.

La funcionalidad de la herramienta se organiza en diferentes áreas: seguridad corporativa, finanzas, sector público, relaciones públicas, comunicación y noticias. Según manifiesta la empresa, "periodistas situados en más de 600 salas de redacción de todo el mundo confían en *Dataminr* para ofrecer los primeros datos sobre noticias de última hora e historias pre-virales en su flujo de trabajo".

Laurent Frisch, directora digital de Radio France, le asegura a la empresa *Dataminr* que el software les ha permitido iniciar el proceso de recopilación de noticias antes que otros medios de comunicación. El ahorro de tiempo es uno de los principios fundamentales de esta herramienta tecnológica.

¿Cómo funciona?

En el ámbito periodístico, ***Dataminr for News*** toma como base de datos las redes sociales. Por su alianza con Twitter, puede explorar este servicio de microblogging y utilizar diferentes señales y sus metadatos para descubrir, clasificar y categorizar lo que merece atención y es realmente importante de lo que no, evadiendo así todo lo que termina siendo ruido o basura informativa.



La empresa lo explica en su sitio web (2019) de la siguiente manera:

> Millones de personas en todo el mundo transmiten lo que ven y escuchan en las plataformas de redes sociales públicas, creando la primera red de sensores presenciales del mundo. Dentro de este mar masivo de redes sociales, *Dataminr* detecta las primeras indicaciones de eventos de alto impacto e información crítica de última hora. Transforma estas primeras señales en alertas en tiempo real, alineadas con las principales prioridades de nuestros clientes e integradas directamente en su flujo de trabajo.

De tal forma, *Dataminr* recibe alertas de hechos e incidentes críticos y descubre historias con valor periodístico antes de que se vuelvan tendencia. Todo en tiempo real. Esto les permite a las agencias y medios de comunicación tomar decisiones editoriales de alto impacto y producir informes y textos de mayor calidad para atraer audiencia. CNN y The Telegraph son algunas de las empresas que usan y apoyan la inclusión de *Dataminr* en sus salas de redacción, así como RTVE.

La compañía informativa Thomson Reuters creó ***Reuters News Tracer***, una herramienta que permite a los periodistas detectar y validar noticias a través de algoritmos que extraen la información de las redes sociales.

Sameena Shah, directora de investigación para el desarrollo de tecnología, y Reginald Chua, editor ejecutivo de operaciones editoriales en la división de datos e innovación de Reuters, aseguran que *Reuters News Tracer* es capaz de detectar los eventos con mayor tendencia en Twitter e identificar su relevancia para alertar a los periodistas de manera confiable y en tiempo real. La herramienta analiza automáticamente el contenido de las redes sociales con el objeto de jerarquizar la información y descartar los mensajes de spam.

Tomando como base los datos históricos y las acciones de los periodistas que forman parte de Reuters, se generan las señales que los algoritmos usan para determinar la veracidad de un tuits. En todo caso, una vez recibida la alerta de un hecho noticiable cada periodista confirma nuevamente la información recibida.

Reuters News Tracer –declara la compañía- ha detectado antes que otros medios de comunicación más de 50 noticias importantes gracias a su codificación. Entre ellas destacan el bombardeo al aeropuerto de Bruselas y el atentado en el barrio de Chelsea en Nueva York en 2016. Lo que permitió a sus periodistas tener una ventaja de publicación de entre 8 y 60 minutos.

La agencia de noticias The Associated Press (AP) utiliza ***NewsWhip*** para rastrear en tiempo real el uso del contenido que se publica y el tipo de relación con el usuario (engagement) que se establece rodea.

En junio y julio de 2017, se analizaron más de 1.2 millones de artículos de AP para clasificar el alcance total en Facebook. Los resultados indicaron que el byline de AP clasifica regularmente entre los cinco primeros autores más comprometidos en esta plataforma.

Bertie13 es un sistema de gestión de contenido (CMS) creada por Forbes y su equipo de tecnología. Su función es proporcionar temas sobre los que escribir, recomendar titulares y sugerir imágenes.

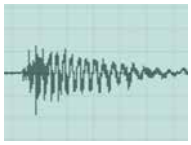
13 https://www.foliomag.com/forbes-cms-relaunch-artificial-intelligence/

El robot, en la primera fase de investigación, funciona basándose en el comportamiento de otras publicaciones, y en función de los resultados obtenidos sugiere, al medio o periodista, temas que están en laza y las tendencias en enfoques, así como en innovaciones de tratamiento.

El sistema puede ofrecer el primer borrador de un artículo para que el periodista pueda tomarlo como base y desarrollar en profundidad la temática planteada.

Los algoritmos y la inteligencia artificial que utiliza *Bertie* para construir sus recomendaciones se nutren del comportamiento de la audiencia y de los datos recolectados de la interacción con cada usuario.

En conclusión, a diferencia de otras herramientas, este servicio se acerca más a un sistema de apoyo a la escritura periodística que a un método de detección de hechos noticiosos.



Quakebot14 , creada por Ken Schwencke, periodista y programador de Los Angeles Times, tiene como objeto recibir alertas y actualizaciones de terremotos en tiempo real, y sin ningún tipo de intervención humana.

14 https://gizmodo.com/quakebot-an-algorithm-that-writes-the-news-about-earth-1547182732

Quakebot trabaja con los datos que proporciona el Servicio Geológico de Estados Unidos.

Detecta las alertas de terremotos que están por encima de un umbral determinado y extrae la información relevante para construir una primera plantilla de aviso. Esta plantilla es recibida por los periodistas y son ellos los se encargan de revisarla, editarla - si es necesario- y, posteriormente, publicarla.

El tiempo aproximado para llevar a cabo todo el proceso oscila entre los tres minutos.

Schwencke y el equipo de programación y datos del periódico, también diseñaron un algoritmo, con características similares a *Quakebot*, que genera informes automáticos sobre homicidios en las zonas que cubre la sección. Su publicación depende siempre, en todo caso, del criterio de los periodistas.

Como en algún caso anterior, la herramienta conjuga elementos de alerta temprana sobre hechos noticiosos –en este caso muy centrada en la observación empírica de los hechos (terremotos registrados)- con sistemas de escritura automática. A diferencia de otros sistemas, la fuente primaria no es aquí ni la Web ni las redes sociales, sino un servicio de seguimiento de terremotos.

La experiencia de RTVE en sistemas de detección de hechos noticiosos

RTVE ha desarrollado algunas experiencias de utilización de sistemas de detección de hechos noticiosos –o sistema de alarma de noticias- con los objetivos de a) obtener información sobre la conveniencia de su integración en sus diferentes redacciones informativas, y, b) diseñar herramientas propias que les permita conocer mejor los posibles sistemas de innovación tecnológica sobre la materia.

En este informe nos centramos en dos experiencias concretas: 1) El uso de la herramienta comercial Dataminr; 2) El desarrollo y experimentación de una nueva herramienta, Social Media Radar.

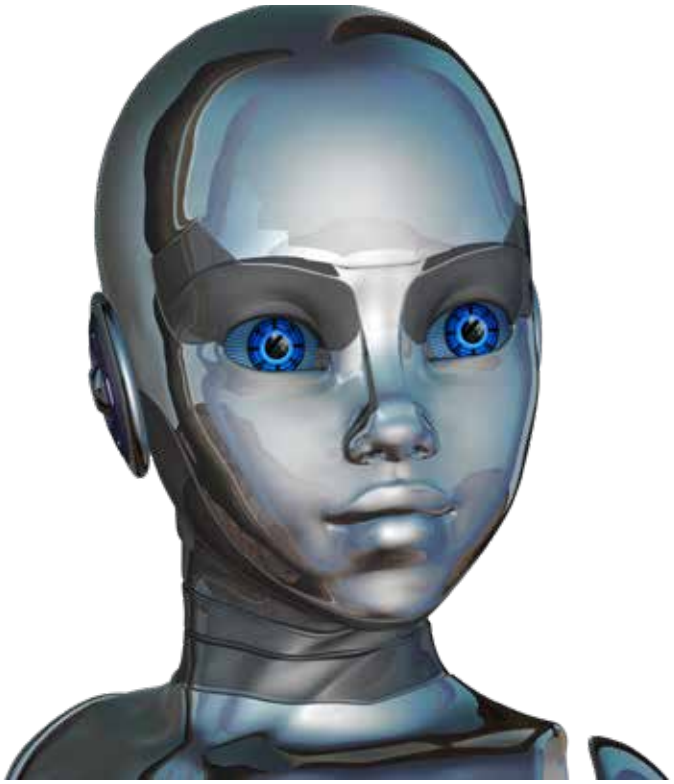
En ambos casos, se trataba de extraer las conclusiones pertinentes en cuanto su uso en RTVE –tanto a nivel de periodista como en el de la redacción en su conjunto-. En el caso de Social Media Radar debe considerarse un valor añadido: se trata de una herramienta desarrollada ad hoc para RTVE por la Universidad Carlos III, dentro de un convenio de colaboración con RTVE.

Observación de la experiencia

Los objetivos de este informe son, únicamente, describir someramente las experiencias realizadas con objeto de señalar las futuras líneas de investigación. Para ello se ha optado por combinar diversas aproximaciones, la observación participante; entrevistas en profundidad con los actores que han intervenido en la experiencia, bien como usuarios, bien como programadores y desarrolladores; y una encuesta dirigida a una muestra de participantes que ha sido analizada estadísticamente.

A continuación, se muestra la tabla de técnicas de investigación aplicadas correspondiente a cada una de las herramientas analizadas para este estudio de caso.

Técnica	Herramienta	Perfil	Fecha de realización
Entrevista	Social Media Radar	Responsable	20 marzo 2019
Entrevista	Social Media Radar	Usuaria	20 marzo 2019
Entrevista	Social Media Radar	Usuario	20 marzo 2019
Grupo focal	Social Media Radar	Desarrolladores	20 marzo 2019
Grupo focal	Social Media Radar	Testeadores	20 marzo 2019
Entrevista	Dataminr	Responsable	20 marzo 2019
Cuestionario	Dataminr	Usuarios	Mayo 2019



Dataminr

Ficha descriptiva del Dataminr- implementación

Web	https://www.dataminr.com
Inicio implementación	Junio/Julio 2018
Número de usuarios activos	100-150
Tipo de soporte	Móvil, ordenador
Principales funciones	Alertas en función de los tuits que se emiten sobre un tema en concreto, según los parámetros definidos.
Fuentes principales de datos gestionados	Twitter
Datos destacados	La función del verificador de la alerta previo al enviar la alerta a los medios es crucial, verifican periodistas repartidos en todo el mundo. Usuarios activos/ observadores, son fuentes fiables porque no publican fake news, es decir, las fuentes principales han sido seleccionadas porque no publican noticias falsas. Los medios que utilizan Dataminr ganan tiempo a las agencias de noticias

Dataminr se empezó a implementar el segundo trimestre del 2018, en julio del 2019 hará un año de su implementación. Inicialmente el uso de esta herramienta era para contar con un instrumento de detección de alertas para conseguir con mayor rapidez las noticias que están sucediendo. Los beneficios esperados son sobre todo obtener inmediatez. Además, Dataminr lo que ofrece es una verificación de noticias de lo que está ocurriendo sin ser una agencia. (R1)

El uso de Dataminr ha sido, para todas las personas de RTVE, de acceso universal y voluntario. Todos ellos tuvieron a su disposición tanto la herramienta como un manual de uso y acceso a algunas sesiones de formación.

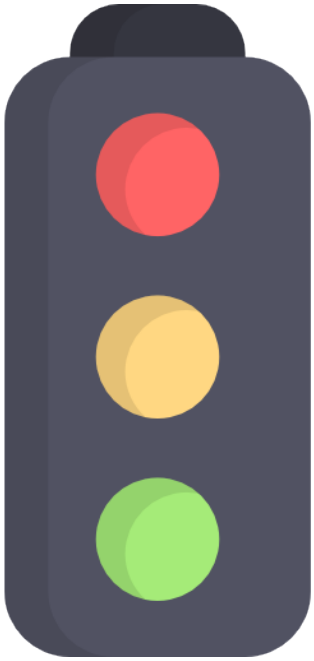
En el momento de la observación participante, lo estaban utilizando entre 100 y 150 personas usuarias, especialmente, periodistas de informativos, redes sociales, web, canal 24horas y radio nacional.

Resultado de los estudios cualitativos

Ventajas

"La principal ventaja es que llegas antes a la noticia, tenemos la noticia preparada antes que otros medios porque hemos accedido a ella antes"

Tras la observación participante y las entrevistas realizadas se pueden obtener las siguientes conclusiones, a tomando en cuenta la versión de los participantes:



- Las ventajas
- El uso de la herramienta permite ganar tiempo en la detección de noticias. En algunos casos, se consigue adquirir una información verificada antes de que las agencias de noticias las difundan, lo que puede permitir reaccionar con celeridad.
- Esta celeridad permite actuar rápidamente a través de las redes sociales y medios interactivos y preparar con más tiempo los servicios informativos que se difunden en horarios regulares.
- Permite acceder directamente a las personas que, por su participación en los hechos, están siendo fuente principal

de la noticia. Por ejemplo, en el caso de un incendio forestal, puedes acceder directamente a quien lo está viviendo en directo y lo está sufriendo.

- Otro beneficio que aporta la herramienta es la rápida detección de accidentes. En un caso concreto, llegó la alerta de un accidente de tren a la redacción, incluso antes de que los servicios de emergencias lo supieran –en concreto, la policía local-. Esto facilitó, además de la información periodística, una más ráida atención a los heridos.

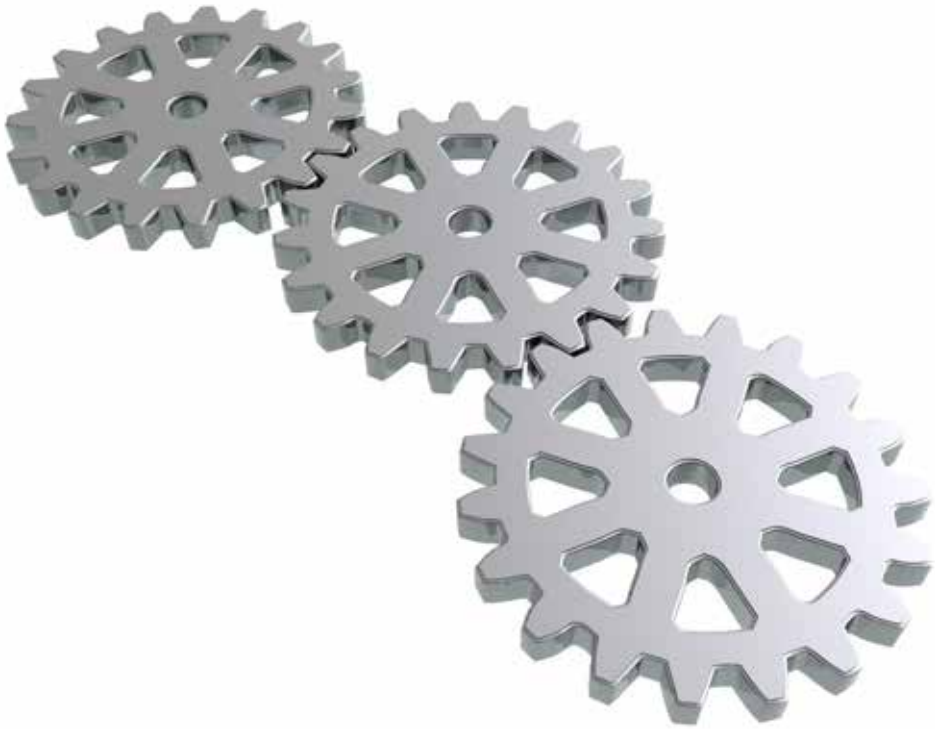
Mejoras recomendadas

- La principal mejora que se sugiere es **adaptar la herramienta al mercado español**. Es una herramienta pensada para un mercado anglosajón, principalmente Estados Unidos y otros países de habla inglesa, y por tanto hay una limitación en este sentido.

- La adaptación exigiría tomar en consideración los idiomas hablados en los territorios de interés del medio, así como depurar la clasificación basada en la toponimia, que, en determinados casos, puede resultar confusa.

Actitudes: Interesados en su utilización y no interesados apoyos y resistencias

Hay más periodistas que están a favor del uso de esta herramienta que en contra. Lo que no nos puede hacer olvidar, de ningún modo, algunas de los contras planteados.



En general, son más favorables al uso de la herramienta quienes trabajan en primera línea de la información en medios interactivos y en contacto con las redes sociales: web, canal 24horas, etc. Su actitud positiva se debe principalmente a que han vivido el beneficio directo que les comporta, en su trabajo diario, acceder antes a las noticias y a los participantes en ellas. De este modo, obtiene una ventaja competitiva en sus piezas y en su trabajo en general.

Hay que señalar, no obstante, los aspectos en contra;

- La más importante reside en el hecho de que la herramienta genera tal cantidad de información que se hace muy difícil procesarla adecuadamente. No obstante, tanto los periodistas como los responsables de la experiencia reconocen que este obstáculo se podría vencer, simplemente, configurando específicamente la herramienta y estableciendo determinados filtros. Pero hacerlo, es en sí misma, una dificultad.

- Otro de los aspectos complejos es el referido a la necesidad de cambiar de hábitos y rutinas de trabajo. Hay algunas personas que prefieren mirar cuenta por cuenta de Twitter que usar los resultados ya procesados por la herramienta. Aquí se percibe el peso de la rutina y de la costumbre.

Encuestas a usuarios

La encuesta se envió a los 100 usuarios potenciales de Dataminr de RTVE. Se envió un primer correo a finales de abril dando como plazo hasta la primera semana de mayo. Y posteriormente se amplió con un segundo correo hasta el 9 de mayo del 2019. De los 100 usuarios posibles, respondieron finalmente 15. A continuación se detallan las respuestas.

Sobre el perfil de las personas que han respondido al cuestionario, principalmente forman parte de la **franja de edad de 40-55 años** con **más de diez años de experiencia** en RTVE.

Tabla 1. Perfil Edad.

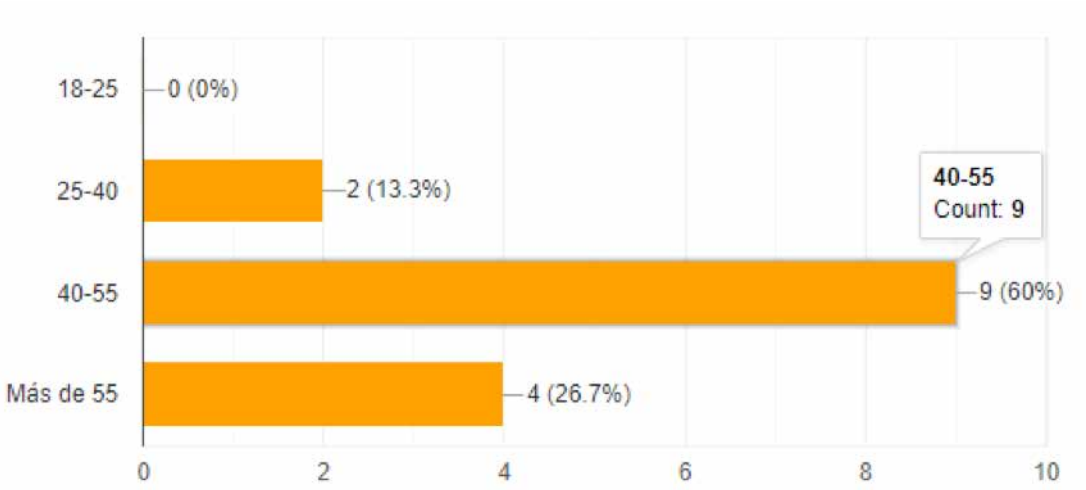
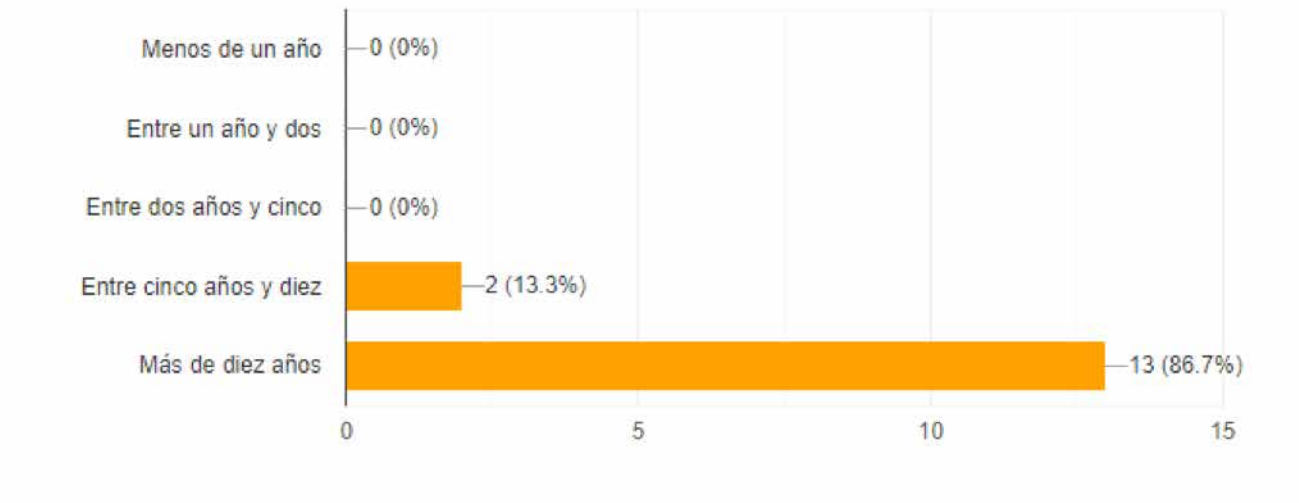


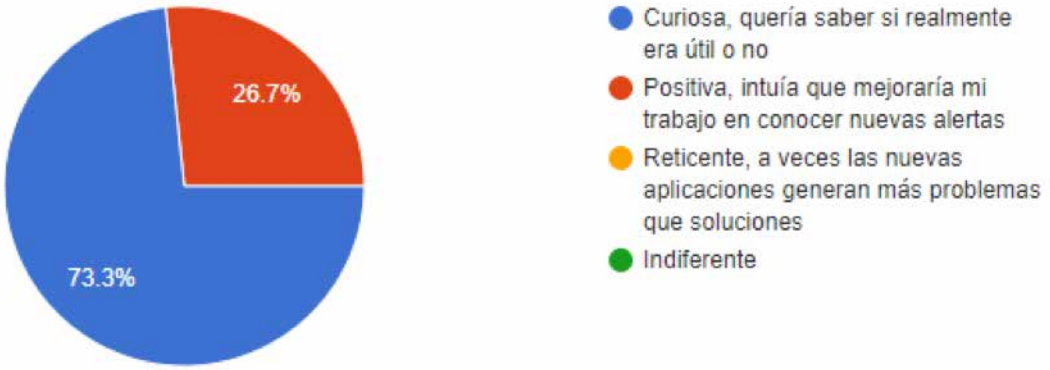
Tabla 2. Experiencia en RTVE



Respecto la cuestión si conocían la aplicación de Dataminr antes de usarla el 93.3% dijo que no. Solo una persona de las encuestadas conocía la herramienta.

En relación a la pregunta ¿Cuándo te plantearon utilizar el Dataminr como sistema de alertas cual fue tu primera reacción?, un 73,3% reconoce que era curiosa para ver si era útil o no y un 26,7% positiva.

Figura 1. Actitud previa ante la herramienta



Ante la cuestión si una vez testeada la noticia le ha generado satisfacción su uso, vemos como un 66.7% afirma que sí, aunque hay un 13.3% que sigue pensando que no le ve la ventaja a otros sistemas.

Figura 2. Satisfacción una vez utilizado Dataminr

Respecto **las funcionales que más se utilizan** según las respuestas cualitativas recibidas son:

- *Alertas, seguimiento de noticias, comprobar fuentes*
- *Las alertas y la verificación de la persona que escribe*
- *Búsqueda de material CGU*
- *Alerta noticias de última hora*

- *Localización de áreas de interés en el mapa, palabras clave, seguimiento (tracking) de temas*
- *Última hora en Twitter*
- *Alerta por correo electrónico*
- *Agencias, otros medios de comunicación, redes*

Respecto **las ventajas que le ha aportado**, son las siguientes:

- *Velocidad, anticipación, seguimiento de hechos que están sucediendo*
- *El servicio de alertas aporta inmediatez y te permite poner el foco sobre acontecimientos que no contemplabas*
- *Localizar material CGU para verificar*
- *Previsión*
- *Inmediatez*
- *En mi caso conocer nuevos sistemas de búsqueda en redes*
- *Rapidez, elementos de edición, fuentes*

- *Recibo alertas y noticias que originan en ocasiones ideas para coberturas y reportajes*
- *Apunta a una posible noticia*
- *Ninguna. No es aplicable a información Local*
- *Estar al día de temas antes que otros*
- *Verificar si algo que comentan en la redacción ha sucedido de verdad, ponerme en funcionamiento y gestionar al equipo humano cuando veo alguna alerta importante*

Y los **beneficios para la redacción**, se definen según los usuarios consultados

- *Mejor conocimiento de la actualidad*
- *Es una herramienta más de información sobre lo que se mueve en redes sociales pero no garantiza la fiabilidad del emisor y es demasiado genérica*
- *Localizar y conocer qué se está haciendo viral*
- *Alerta de noticias - capacidad de respuesta*
- *Inmediatez en las noticias*

- *Alertas tempranas en el caso de accidentes, atentados o noticias importantes*
- *El seguimiento de asuntos recientes*
- *Última hora, fuentes, elementos de edición*
- *Es otra fuente de información que alerta y da pie a coberturas con anticipación*
- *Otra fuente mas*
- *Que podamos empezar a preparar las historias con más antelación. Poder confirmar si algo es veraz porque viene de una fuente oficial o de varias fuentes concordantes*

Dificultades

Según las respuestas recibidas**, las principales dificultades** son:

- *Falta de personalización de la herramienta*
- *Discrimina poco, manda demasiada información que, en un alto porcentaje, no me resulta interesante*
- *La complejidad de la interfaz*
- *Imposibilidad de acceder a la Web de la herramienta, ya que el recibir solo la alerta a través de los mails resulta engorroso*
- *Imposibilidad de realizar búsquedas locales*

- *Las búsquedas locales*
- *La versión de escritorio, algo confusa. El mapa de la versión móvil, difícil de manipular. La información internacional es buena. La local, más limitada por la menor cantidad de fuentes contrastadas*
- *A veces se destacan como urgentes tuits que no tienen ninguna importancia*
- *El algoritmo está poco afinado aún. Hay percepciones informativas diferentes en cada medio*
- *DataminR se convierte en una fuente más, junto al gran número de fuentes ya existentes*
- *Restringir los campos de búsqueda para usos concretos muy locales*
- *Conseguir un nivel de alertas que sea relevante pero no sature. Ahora mismo hago que entren pocas alertas porque tengo un filtro alto*

Propuestas de mejora

Y en cuanto a **la mejora de la herramienta,** se propone lo siguiente:

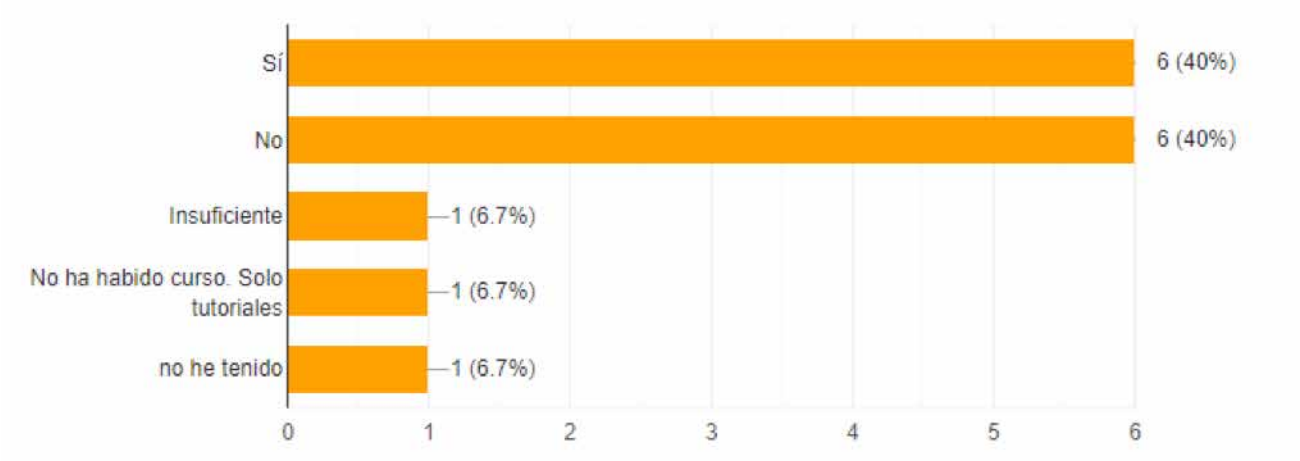
- *Más contenido específico en cuestiones locales*
- *Permitiendo afinar más los parámetros de búsqueda*
- *Mejorando la interfaz*
- *Sí en búsquedas locales*
- *Mejorar la información para España, agilizar la configuración del mapa*
- *Mejorando el filtrado de los tuits de alerta o urgentes*
- *Incluyendo la percepción del medio sobre lo que es y lo que no es noticia*
- *Con más parámetros que adapten las alertas a objetivos más concretos*

- *Afinando mucho más los filtros para que de verdad sólo lleguen avisos importantes de los campos seleccionados en áreas regionales*

- *Teniendo más alcance minucioso*

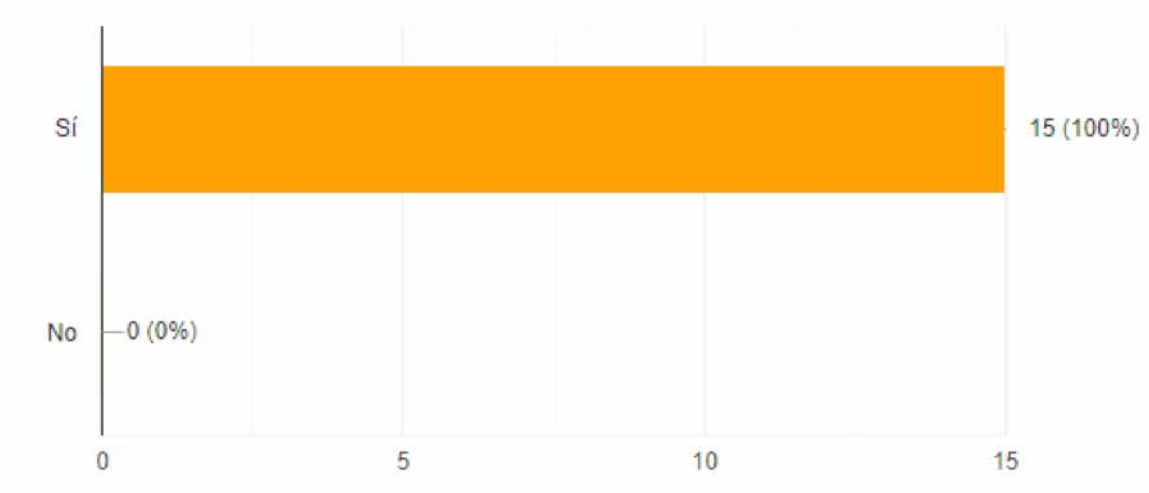
Respecto si la **formación recibida** para utilizar la herramienta ha sido adecuada, las personas consultadas responden lo mostrado a continuación en la tabla 3

Tabla 3. Adecuación formación



La experiencia ha sido positiva, aunque haya elementos de mejora como han señalado durante la encuesta.

En un dato están de acuerdo el 100% de los encuestados. Ante la pregunta: a partir del uso de esta aplicación, ¿te interesa seguir incluyendo aplicaciones de inteligencia artificial en el desarrollo de tu trabajo como periodista? Todas las personas consultadas responden afirmativamente. Este dato que nos confirma que sigue habiendo interés en la redacción para continuar conociendo otras herramientas de inteligencia artificial.



Social Media Radar

***Social Media Radar* (SMR)**15 es otra herramienta que aplica inteligencia artificial para detectar noticias. Específicamente, es un software diseñado y creado por la Universidad Carlos III de Madrid para RTVE, cuyo objetivo se centra, según su manual de instrucciones en:

> Combinar técnicas de monitorización de redes sociales, big data, procesamiento de lenguaje natural, data mining, búsqueda y análisis de sentimiento para dar soporte al diseño y desarrollo de un servicio de monitorización y alerta multi-nivel de contenidos en redes sociales para la detección temprana de noticias, clasificación y notificación a los profesionales de la información a través de la computación móvil. (p 1.)

A través de la aplicación Telegram, que existe para todos los sistemas operativos como IOS y Android, puede buscarse el bot de SMR y guardarlo como si fuera un usuario más.

15 http://www.kr.inf.uc3m.es/?s=SMR

Al interactuar con él, será posible darse de alta o subscribirse para recibir notificaciones del sistema de alerta de las diferentes comunidades autónomas.

Social Media Radar incluye también un sistema de feedback para responder a las alertas recibidas a través de la suscripción por correo electrónico y un sistema de validación y comparación con las alertas de noticias emitidas por la agencia EFE.

Adicional a esto, se desarrolló una app que monitorea las redes sociales para detectar accidentes de tráfico. Esta aplicación puede consultarse mediante Microsoft Azure, una nube de RTVE.

Otras aplicaciones

Así como RTVE o Radio France emplean e impulsan proyectos vinculados al uso de la inteligencia artificial para acelerar la detección de hechos noticiables, agencias comunicacionales de todo el mundo se unen a esta tendencia.

Es necesario destacar el carácter innovador del desarrollo de esta herramienta, es fruto de la colaboración de RTVE con la Universidad Carlos III, y en las diferentes fases de desarrollo han ido mejorando la aplicación. A continuación, ofrecemos las principales contribuciones extraídas fruto del trabajo de campo.

Web	http://www.kr.inf.uc3m.es/?s=social+media+radar
Inicio implementación	2018
Número de usuarios activos	Aprox. 90
Tipo de soporte	Móvil, canal de telegram
Principales funciones	Detectar alertas principalmente de accidentes de tráfico, pero no únicamente. Su principal fortaleza es la localización.
Fuentes principales de datos gestionados	Twitter

Responsable del proyecto

La entrevista se realizó a la responsable de implantación de Social Media Radar.

Destaca que la iniciativa se haya ideado a raíz de unas experiencias previas relacionadas con la Cátedra de Innovación. Concretamente, en este contexto, en el pasado se desarrolló un programa de "Sentiment analysis" en las redes para la elección de los presentadores que en la actualidad ya no se usa.

En la génesis del proyecto, se encargó a los desarrolladores generar una herramienta que permitiera a las redacciones locales ser las primeras, antes de las agencias, en detectar accidentes de tráfico. Dataminr lo hace a nivel global y querían algo de proximidad.

Se eligen los centros locales porque tienen límites de tiempo (horarios laborales muy concretos) y presupuestos ajustados

Las redacciones locales usan tanto las agencias como fuentes periodísticas propias, y no se quejan. Se pensó el proyecto para innovar y sistematizar la abundante y desorganizada información que aparece en las redes.

Social Media Radar se implementa en cuatro centros locales: Madrid, Castilla la Mancha, Extremadura y Cataluña.

En este momento, más allá de los accidentes, se están intentando incorporar en Cataluña y en Castilla alertas para accidentes y catástrofes naturales.

El proyecto tiene recursos (la responsable, una documentalista y 2 testeadores) durante 8 meses y va a seguir desarrollándose hasta agosto. El futuro es incierto y la evaluación del proyecto no tiene un protocolo establecido.

Ventaja

- Efectivamente, comparándolo con EFE, las redacciones que usan SMR llegan antes.

- Es de ayuda para las redacciones locales que querían algo de proximidad que permitiese ayudarles en su situación de escasez de recursos y problemas de horario laborales.

- Hay estadísticas de acierto.

Desventaja

Hay un problema de coincidencia toponomástica en Extremadura (muchas ciudades extremeñas tienen ciudades, e inclusive clases, con el mismo nombre en México y en otros lugares de Latino América). Su desarrollo e implantación precisa de más recursos tanto en la fase de diseño como en la fase de desarrollo e implantación posterior

> "Se incorpora lo existente y exitoso, lo nuevo da miedo y hay que incorporarlo con mucho cuidado". (Responsable SMR)

Desarrolladores

La entrevista se realizó al responsable y a cuatro de los desarrolladores de Social Media Radar.

Su objetivo inicial era de generar una pre-alerta de accidentes no comprobada, que necesita confirmación por parte del ser humano, a través de la lingüística computacional.

El punto de partida es que la información en las redes es abierta. Se elige Twitter porque es donde aparecen las noticias y las apps existentes son principalmente de marketing.

¿Cómo funciona social media radar?

El API de Twitter es infinito. SMR selecciona solo noticias en idioma español, donde aparezcan palabras del dominio (accidente, catástrofe natural, etc.) y, en la medida de lo posible, geolocalizadas (aunque muy pocos tuits estén geolocalizados). Antes Twitter proporcionaba el offset y franja horaria, pero ahora ya no se puede.

Para seleccionar las palabras del dominio se han indexado y analizado por frecuencia las palabras relacionadas con accidentes del diario *El País*. Esta operación ha permitido aislar 400 palabras claves relacionadas con el universo lexical de los accidentes.

FASE 1: Filtros

El interés se establece antes de todo a nivel cuantitativo: el número de tuits sobtre un mismo hecho establece un nivel de alerta (numérico).

FASE 2: Veracidad

SMR tiene listas blancas y negras.

La lista blanca está formada por cuentas de Twitter fiables (periódicos locales, protección civil, bomberos, guarda forestal, 112, etc.).

Para construir la lista blanca se partió de "cuentas semilla" (bomberos, etc.) que se expandieron manualmente siguiendo a las cuentas que las "semillas" siguen.

Los desarrolladores admiten que en esta fase hubiera sido de mucha ayuda la participación de un periodista.

La lista negra, también construida manualmente, no perjudica pero descarta tuits y está formada por cuentas, principalmente de Latino américa (por el citado tema de la toponomástica parecida).

FASE 3: DESAMBIGUACIÓN POR LOCALIZACIÓN

Se priorizan/descartan los Tuits si indican, por ejemplo, ciudad y calle y si tienen foto tienen más valor.

Sin embargo, tal y como afirma un desarrollador "si las descripciones del usuario son muy largas es un desastre".

Esta información llega por e-mail y Telegram y los testadores, son los encargados de dar el visto bueno final.

FEEDBACK

De momento, sólo el 70% de las alertas resultan verídicas en Extremadura, por el citado tema de la toponomástica parecida con otros países, mientras en Cataluña el 85%, debido a la toponomástica única en catalán.

Ventaja

En dos casos los equipos de los centros locales de RTVE llegaron antes que el 112 (el caso del tren y el caso de las elefantas).

Desventajas

El desarrollo de una herramienta de este tipo exige esfuerzo y recursos, que no siempre están disponibles.

A nivel estructural los desarrolladores asumen que SMR no funciona bien en casos donde hayan aparecido nombres de personas que se llaman como lugares.

Además, consideran que SMR falla a la hora de proporcionar información completa: Twitter solo da una parte de los datos que el periodista necesita.

Otro tema que genera "ruido" es el hilo (la gente sigue hablando de un tema aunque ya no se trate de un accidente y SMR sigue mandando alertas).

Mejoras

Según los desarrolladores el futuro será "Taylor made", es decir, la ambición es crear un sistema de alerta construido a medida de los intereses de los usuarios.

Testeadores

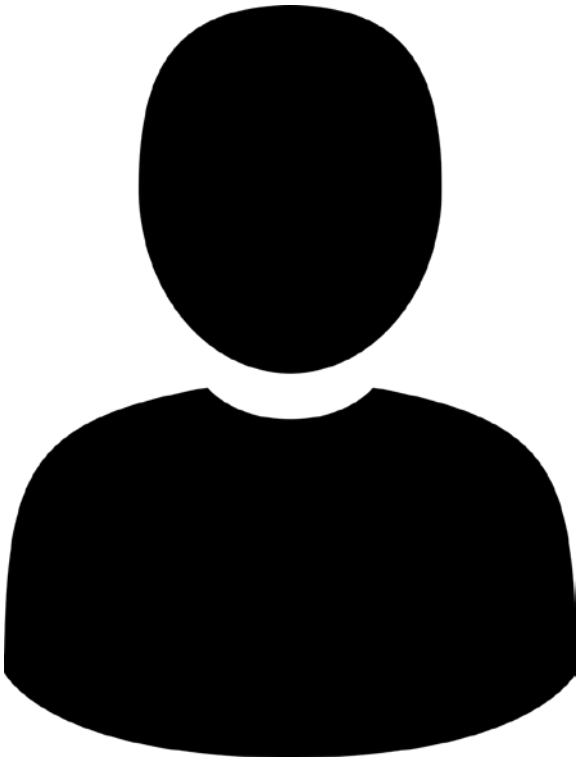
Se realizó un grupo de discusión con dos personas que en este momento están llevando a cabo el rol de testeadores. Se han incorporado recientemente, y su función se basa principalmente analizar los resultados que envía la herramienta para verificar su validez. En este sentido durante el diálogo destacaron que, si bien la herramienta está siendo bastante útil, si se detectan algunos fallos que deben mejorarse en cuanto la captación de la información.

Usuarios

Las entrevistas se realizaron a dos periodistas usuarios de Social Media Radar (unidad Castilla La Mancha y Extremadura).

Castilla La Mancha

En el caso de Castilla La Mancha iniciaron el uso desde la fase de testeo, hace dos años. **Las expectativas** de las personas eran positivas, según la entrevistada. Había un desconocimiento de lo que podría ofrecer, pero todo lo que fuera añadir una fuente más de acceso a noticias se consideraba interesante para explorar. La red se comenzó a usar hace un par de años.



Ventaja

La rapidez de acceder a un hecho noticiable, las fuentes oficiales suelen tardar más en dar la información.

Inconveniente

En un primer momento el principal inconveniente era el filtrado de la información, pero en la segunda fase ha mejorado mucho.

Respecto la formación se realizó un seguimiento mediante e-mails, se proporcionó una guía de usuarios y también la posibilidad de establecer contacto directo con los desarrolladores, lo cual fue positivo para ir resolviendo las dudas a medida que iban surgiendo.

Se valora positivamente su uso, ya que constituye una fuente de información más, siempre es útil acceder a una información nueva.

Sobre la pregunta sí se consideraría positivo utilizar otras herramientas de IA en las diferentes fases de elaboración de noticias, la respuesta es afirmativa.

"Sí, la IA no es el futuro, ya está aquí, y todo lo que nos ayude a ejercer nuestra profesión mejor, bienvenido sea". (Usuaria SMR)

Extremadura

Social Media Radar se está utilizando desde septiembre del 2017 en esta unidad, desde los principios, desde la fase de testeo. En este momento hay 3 personas utilizando la herramienta. La expectativa previa a su implementación era positiva, un sistema de este tipo que salta una alerta resultaba interesante. Como equipo han ido participando en los reportes de fallos y viendo cómo se podría ir mejorando.

Mejoras

En el caso de Extremadura, la aplicación de SMR tiene tres principales limitaciones; 1) La densidad de la población es baja y hay pocos usuarios activos en Twitter, 2) Al ser una tierra de conquistadores, la mayoría de poblaciones extremeñas tienen su homólogo en Sudamérica, llegan muchas alertas de Trujillo, Medellín,etc.. pero de Sudamérica, 3) y por suerte no existen casi accidentes de tráfico.

Desde la nueva ley de protección de datos, también se ha visto limitada la alerta por la deslocalización de los tuits. También se constata en el territorio que hay más usuarios activos en Instagram (jóvenes) y Facebook (población mayor).

Las propuestas de mejora en este caso después de la entrevista serían:

- Intentar adaptar la herramienta a la realidad de cada población, en el caso de Extremadura bajar el índice de tuits para considerarlo incidente puesto hay menos densidad de población, si fuera posible.

- Ampliar, en el caso que se pudiera, a otros temas. Les llega que los usuarios están activos en el tema del tren (reivindicación del AVE) o en el caso de una central nuclear.

En relación a la pregunta si están interesados en seguir aplicando otras herramientas de IA en las sucesivas fases de elaboración de noticias, la respuesta es afirmativa.

> "Bajar el índice de actividad para generar la alerta en nuestro caso es importante por la realidad poblacional de Extremadura" (Usuario SMR)

Conclusiones generales

UN CAMPO EN PLENO DESARROLLO

La primera conclusión que se deriva tanto del análisis de la investigación científica como industrial y de los estudios de casos, es que la aplicación de la inteligencia artificial al campo de la detección de noticias es un campo en pleno desarrollo. El número de investigaciones científicas detectadas es ya significativo y parece que aumentará en el futuro inmediato. Y aunque en su mayoría proviene aún del campo de la informática y de la ingeniería, los estudios aplicados al periodismo o campos relacionados, tienden a aumentar.

Del mismo modo, la comercialización y uso de herramientas de IA aplicadas al sector, no solo aumenta, sino que está aumentando su capacidad gracias al uso intensivo de big data y al aprendizaje automático que las máquinas desarrolladas están experimentando. En este sentido, la actividad en el campo va a aumentar, sin duda en los próximos años.

Por otro lado, los estudios de caso analizados, pese a su novedad y singularidad revelan que -tras las primeras y lógicas resistencias- los profesionales van a ir incorporando a sus prácticas diarias las herramientas de inteligencia artificial, lo que, sin duda, proporcionará vialidad y nuevas experiencias al sector.

Del mismo modo, la observación del creciente número de actividades de debate, experimentación y formación sobre la materia hace presagiar que, poco a poco, la IA se v a ir incorporando, de un modo natural, a las actividades periodísticas, tanto en el campo de la formación como en el de la actividad profesional. Lo cual proporcionará cambios sustanciales en la organización del trabajo. E impactará, decisivamente, en el sistema de búsqueda de noticias y de producción audiovisual en relación con ellas.

LA INGENIERÍA COMO PRINCIPAL VECTOR DEL CAMBIO

De los estudios realizados, se desprende que es la ingeniería y la ingeniería informática en particular la que actúa de vanguardia en el sector. Mientras el ámbito periodístico suele actuar reactivamente ante ella. Lo cual viene a significar que: a) tanto las organizaciones periodísticas como sus profesionales no han llegado a formular aún demandas meditadas y conscientes como para convertirse en el tractor esencial de la innovación. Entre tanto, es la industria de la información y de la tecnología la que lidera los cambios.

De ahí las dubitaciones e incertidumbres que se ciernen, en algunos casos, sobre la adecuación de las estrategias de innovación. Y de aquí, también, las desconfianzas e inercias profesionales. En general, los periodistas ven su labor más amenazada por la IA que complementada por ella.

Se puede concluir, por tanto, que urge una toma de conciencia periodística sobre las posibilidades y sobre las demandas a las que puede responder la IA. Si, durante los próximos años, esta toma de conciencia se produce, la innovación tecnológica estará mejor fundada. Si no, se corre el riesgo de ir siempre a remolque de una innovación dominada por el determinismo tecnológico.

UN PROCESO DE MADURACIÓN

La mayoría de experiencias y casos de uso de la IA en materia de detección de noticias se halla, aún, en fase incipiente de desarrollo. Cabe esperar en que en los próximos años, los mecanismos empleados se afinen y se desarrollen mejores y más sofisticadas aplicaciones.

Por ahora, es la extrapolación de las experiencias vividas en otro campos distintos al periodismo, los que están marcando la agenda experimental. Y hay que esperar que esta sea una etapa de pronta superación, que desemboque en una maduración de la tecnología que se acerque más a las necesidades específicas de las organizaciones periodísticas.

LAS GRANDES ORGANIZACIONES PERIODÍSTICAS, MÁS PROCLIVES AL USO DE APLICACIONES DE IA EN LA DETECCIÓN DE NOTICIAS

La aplicación de herramientas de IA en los equipos de redacción de los medios y agencias internacionales más relevantes son un hecho. Agencias como Reuters, Associated Press tienen su propio sistema de detección de noticias, siendo las redes sociales una de las fuentes principales de detección de" breaking news".

Este hecho se debe, muy probablemente, a que son las grandes entidades periodísticas -que abarcan un amplio territorio y producen ingentes cantidades de información las que mejor se benefician de la automatización de los procesos de detección de noticias. Son, al mismo tiempo, estas entidades las que se benefician mejor de un pronto acceso a la información, en la medida en que este produce un incremento del valor de la información.

Por todo ello, cabe prever que en pequeñas organizaciones más centradas en el análisis y en la aplicación de criterios cualitativos a la hora de difundir la información, la aplicación de herramientas de IA será más lenta y más indirecta.

LA CONSTANTE NECESIDAD DEL FACTOR HUMANO

Dada la naturaleza misma del valor de la noticia, cabe pensar que la participación humana en su detección y su valoración seguirá siendo decisiva. Los sistemas de detección de noticias estudiados solo sirven para proporcionar más y mejores fuentes de información a los periodistas. Y les permitirán, responder mejor y antes a las demandas del entorno, pero siempre exigirán la valoración con criterios cualitativos y en función de factores contextuales -entre otros- para que el proceso sea eficaz y tenga éxito. Lo cual asegura que la IA, en este campo, no sustituye nunca la labor de los profesionales, sino que la complementa y la mejora.

CAMPOS PRIVILEGIADOS DE USO

Los casos analizados, así como las experiencias estudiadas, ponen de relieve que existen algunos campos en los que la detección de hechos noticiosos puede ser más productiva. En general son los que reúnen algunas de las condiciones siguientes:

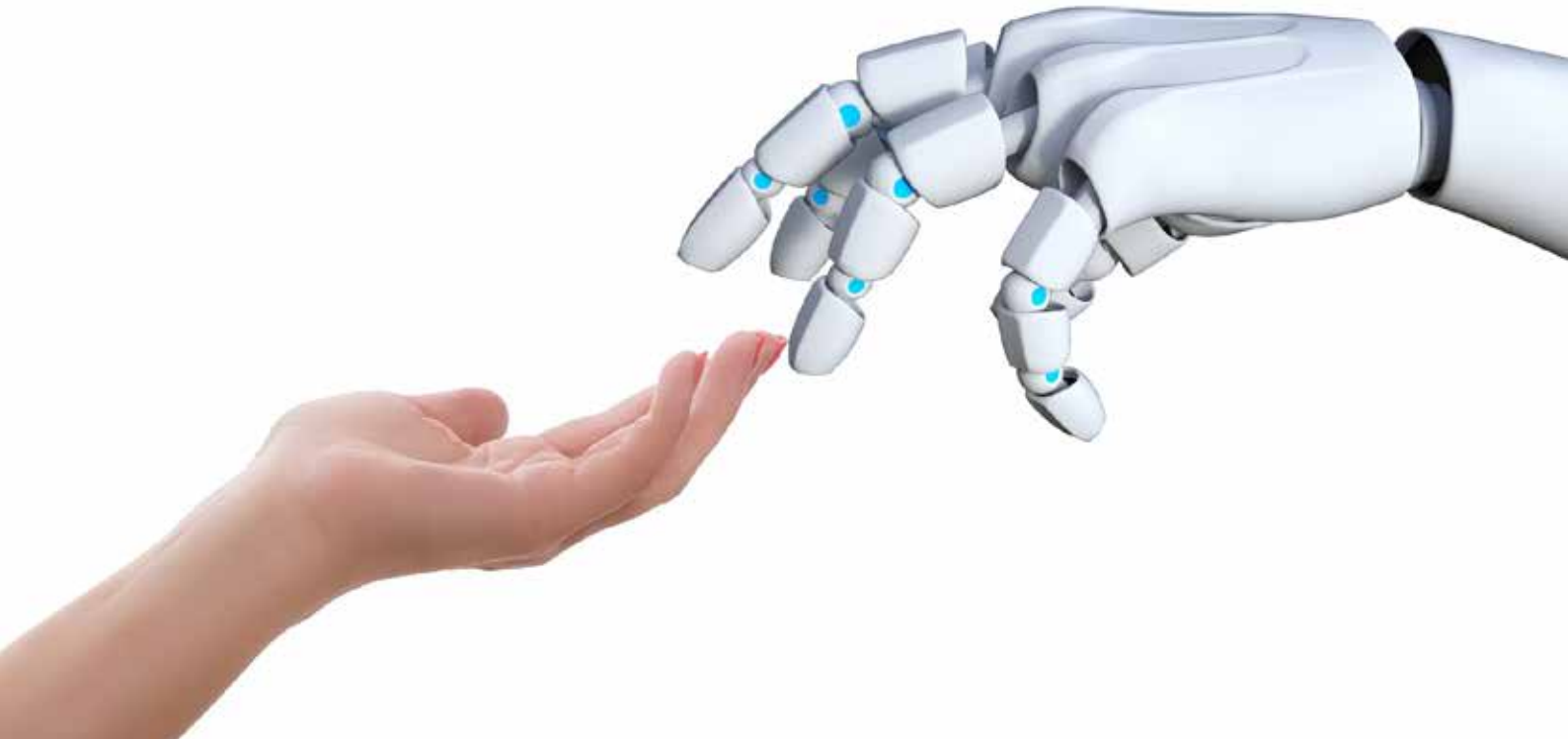
1. Se derivan de la existencia previa de grandes datos: fuentes de datos metereológicos, redes sociales, información disponible en internet o indicadores generados automáticamente. A partir de su estudio y de su análisis pueden reconocerse determinados hitos que inducen a proporcionar pistas sobre hechos noticiosos.

2. Se relacionan con áreas periodísticas en las que la inmediatez y la celeridad informativa son un valor: catástrofes naturales, accidentes, alertas meteorológicas, epidemias, etc.

3. Responden mejor cuando el hecho noticioso es capaz de provocar por sí mismo una alteración de la conversación pública: de aquí la importancia que tiene las redes sociales como indicador de hechos noticiosos.

4. Proporcionan datos precisos sobre el hecho noticioso que no recogen otras fuentes ya existentes.

5. Pueden permitir la participación del público en la producción de datos: otra vez, aquí, las redes sociales juegan un papel fundamental.

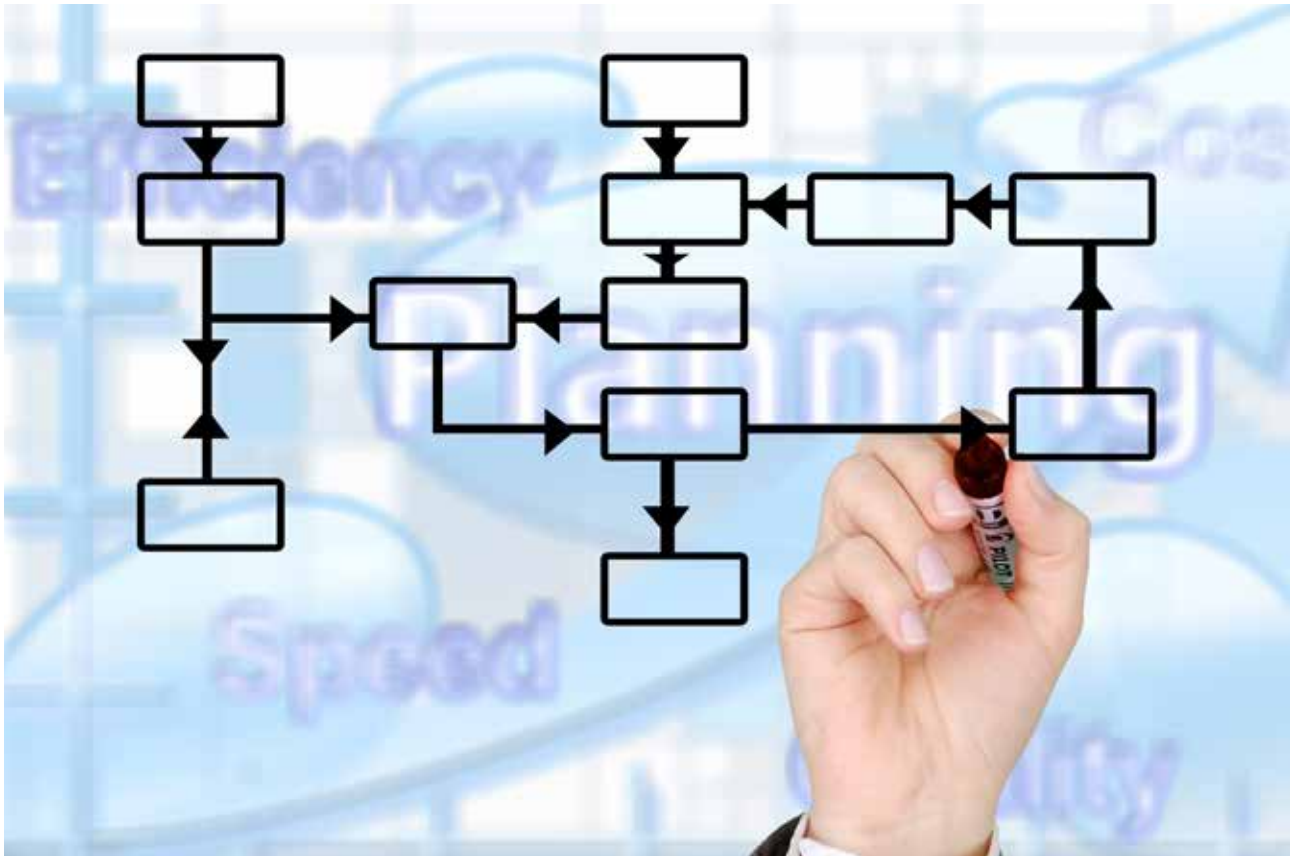


UN PROCESO DE EXPERIMENTACIÓN E INNOVACIÓN NECESARIO

Todos los datos analizados apuntan a la necesidad de continuar con un proceso de investigación iniciado y la necesidad, también, de involucrarse directamente en proceso de experimentación. Sólo así se podrá obtener un beneficio claro de las herramientas existentes.

Lo que se ha podido comprobar es que las aplicaciones de IA exigen tanto un aprendizaje de las mismas máquinas, aprendizaje automática, como de las organizaciones en que estas se implementan. Sin ese doble aprendizaje, las herramientas resultan inútiles. Lo cual exige: a) experimentación; b) aprendizaje; y c) cambios estructurales que permitan la innovación.

El éxito de est proceso depende de la capacidad de las organizaciones para poner en marcha estrategias de innovación a corto, medio y largo plazo. De su capacidad de invertir recursos, tanto humanos como financieros. Y del grado en que la cultura de las propias organizaciones sea capaz de incorporar el principio de innovación crítica a sus prácticas habituales.



Referencias de la literatura científica

Bollier, D. Artificial Intelligence: The Promise and Challenge of Integrating AI Into Cars , Healthcare and Journalism. Maryland: The Aspen Institute. (2017).

Capdevila, J., Cerquides, J. & Torres, J. Event Detection in Location-Based Social Networks. *Book - Data Science and Big Data: An Environment of Computational Intelligence* (2017).

Dashdorj, Z., Tsogtbaatar, B., Tumurchudur, A. & Altangerel, E. High Level Event Identification in Social Media. *Proceedings - 12th International Conference on Semantics, Knowledge and Grids, SKG* (2016).

Gu, Y. *et al.* Detecting Hot Events from Web Search Logs. *Conference Proceedings - Web-Age Information Management* (2010).

Guille, A. & Favre, C. Event detection, tracking, and visualization in Twitter: a mention-anomaly-based approach. *Social Network Analysis and Mining.* 5, 1–18 (2015).

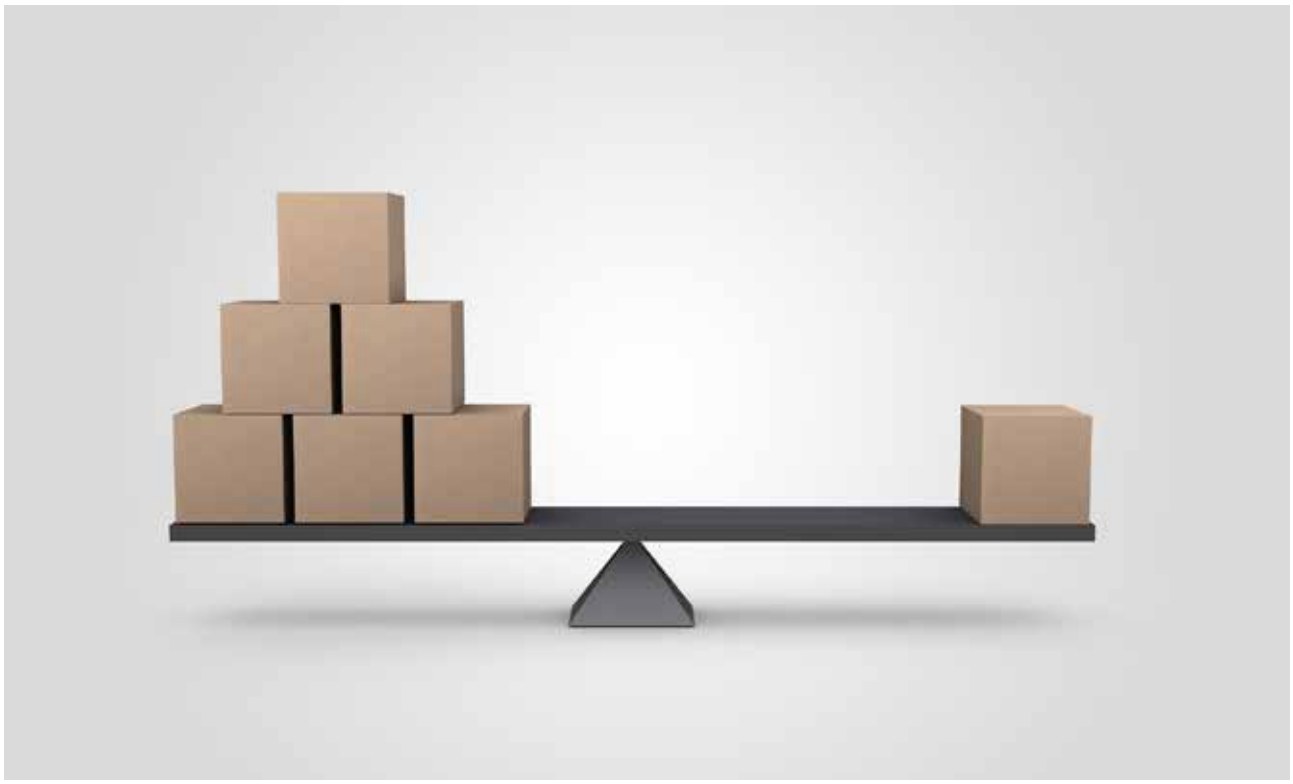
Hua, T., Chen, F., Zhao, L., Lu, C. T. & Ramakrishnan, N. Automatic targeted-domain spatiotemporal event detection in twitter. *Geoinformatica.* 20, 765–795 (2016).

Lee, C. H., Chien, T. F. & Yang, H. C. An automatic topic ranking approach for event detection on microblogging messages. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics.* 1358–1363 (2011).

Marconi Alex, Journalist, Machine, F. S. (2017). A guide for newsrooms in the age of smart machines. The Future of Augmented Jounralism. Retrieved from https://insights.ap.org/uploads/images/the-future-of-augmented-journalism_ap-report.pdf

Newman, N. (2019). Journalism, Media and Technology Trends and Predictions 2019. Disponible en: http://www.digitalnewsreport.org/publications/2019/journalism-media-technology-trends-predictions-2019/

Webb, A. (2018). 2018 Industry Trends: Journalism, Media, Technology. Disponible en https://www.amic.media/media/files/file_352_1341.pdf



Referencias del Benchmarking

Álvarez, J. Rozalén, M Rodríguez & F. Jiménez, A. (18 de febrero de 2019). *Proyecto Social Media Radar 3. Entregable M6.* Madrid.

BBC News. (2014). *Robot writes LA Times earthquake breaking news article.* Recuperado de: https://www.bbc.com/news/technology-26614051

Fundación Tena. (2018). *Inteligencia Artificial: para qué puede usarse en periodismo y qué están haciendo los medios.* Recuperado de: https://www.laboratoriodeperiodismo.org/inteligencia-artificial-para-que-puede-usarse- en-periodismo-y-que-estan-haciendo-los-medios/

Micó, J. (2017) ¿Cómo puede transformar la inteligencia artificial las noticias? La Vanguardia. Recuperado de: https://www.lavanguardia.com/tecnologia/20170623/423605313675/inteligencia- artificial-noticias-periodismo-prensa-medios-de-comunicacion.html

Narrative Science. (2019) Products: Quill. Recuperado de: https://narrativescience.com/products/quill/

Zalatimo, S. (2018). *Entering The Next Century With A New Forbes Experience.* Forbes. Recuperado de: https://www.forbes.com/sites/forbesproductgroup/2018/07/11/entering-the-next-century-with-a-new-forbes-experience/#3fdf74a93bf4

Websites

https://blogs.thomsonreuters.com/answerson/making-reuters-news-tracer/

https://archive.annual-report.thomsonreuters.com/2016/is-it-news-or-noise.html