Leyendo todas las noticias del mundo con R (and friends)

Anónimo

Palabras clave: machine learning - natural language processing - news - anti-vax - media analysis

Introducción

A pesar del éxito de la vacunación, que en el último siglo ha prevenido la muerte y enfermedad de millones de niños, crece la desconfianza hacia las vacunas alrededor del mundo. Encuestas globales indican que en los países desarrollados está declinando la confianza respecto a la seguridad y valor de las vacunas, a la vez que reaparecen brotes de enfermedades antes controladas, como el sarampión en Europa.

La desinformación sobre los peligros de las vacunas circula en medios masivos y en redes sociales, y la reticencia a vacunar a los menores se transforma en un movimiento, llamado "antivacunas" o *antivax*. Ante la gravedad de la situación, en 2019 la Organización Mundial de la Salud declaró al movimiento antivacunas como una de las 10 mayores amenazas a la salud pública mundial.

La repercusión del mensaje antivacunas depende de la producción y la circulación por medios de comunicación y redes sociales de discurso, a favor o en contra, acerca de la inmunización. Para poder atender el problema de la desinformación, necesitamos entender la dinámica de la circulación de noticias y artículos de opinión, contestando preguntas como

```
Qué mensajes circulan sobre las vacunas en los medios locales?
Qué peso tiene el discurso antivacunas en los medios de comunicación?
Cuáles son los principales formadores de opinión sobre de las vacunas?
```

A leer todos los diarios con R… y sus amigues

Una forma encontrar respuestas sería accediendo a un archivo diario, que cubra los últimos años, con el contenido de todas las noticias publicadas en un país por sus periódicos y portales de noticias. Tal cosa existe: el proyecto GDELT *(Global Database of Events, Language, and Tone)*[1]. GDELT es una inmensa base de datos de acceso libre sobre la sociedad humana, considerada como la "de mayor tamaño, más completa, y de mayor resolución jamás creada". Crece cada día gracias a una iniciativa que monitorea las noticias *online* de una gran cantidad de países en más de 100 idiomas, acumulando las noticias publicadas e identificando las poblaciones, ubicaciones, organizaciones, temas y emociones presentes en cada artículo. Actualizada cada 15 minutos.

Y si: podemos usar **R** para acceder a la base GDELT y procesar sus registros, utilizando paquetes clave como **`dbplyr`** y **`bigrquery`**. Y en combinación con un plantel de invitados *open source* especializados (como la librería de **Python `newspaper3K`** para extracción de contenidos periodísticos, o la librería de **C `fastText`** para clasificación automática de texto) podemos crear un flujo de trabajo automatizado para monitoreo continuo.

El sistema de clasificación de texto *fastText* emplea una variante de la estrategia conocida como "bag of words". Convierte cada palabra una serie de atributos numéricos (un vector). Al analizar un texto, sólo extrae el identificador numérico de cada palabra, sin que sea necesario considerar el orden en que aparecen los términos ni su semántica -de allí el nombre, una simple "bolsa de palabras". La particularidad del método fastText es que incorpora una mejora respecto al *bag of words* tradicional, al considerar n-gramas (secuencias de palabras) y no sólo vocablos sueltos al realizar la conversión en vectores.

Con estas herramientas podemos poner en marcha un sistema que, día a día, identifica los artículos del país de interés. Y luego extrae los que mencionan a las vacunas (o el tema que se investigue), recupera su contenido completo y detecta los tópicos presentes utilizando un modelo entrenado por *machine learning* supervisado.

[1] https://www.gdeltproject.org/

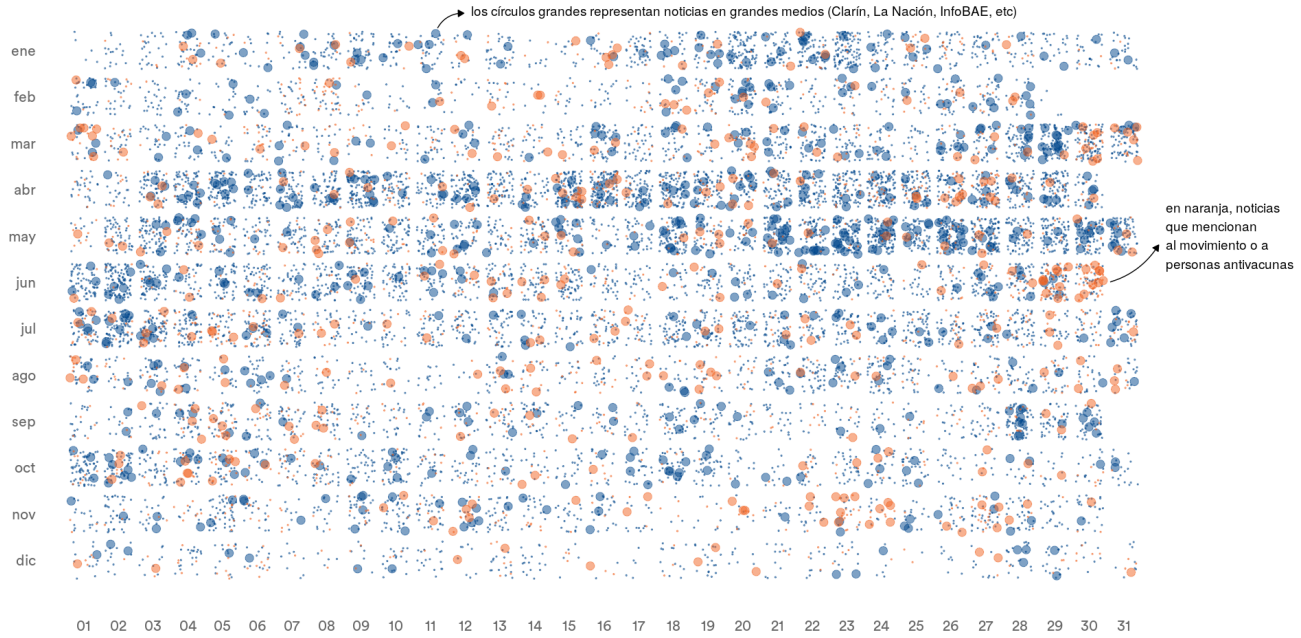

Figura 1: Noticias sobre vacunas en medios online argentinos durante 2019

Resultados

La clasificación automatizada de noticias vía fastText logró un muy alto nivel de precisión, basándose en el etiquetado por revisión humana del tópico central en algunos centenares de artículos.

En cuanto al contenido de las noticias, encontramos unos 42 artículos por mes que hablan sobre el fenómeno del antivacunismo; 8 de ellos en alguno de los grandes medios de alcance nacional. El tema es siempre abordado sin atacar la efectividad o valor de la vacunación pública desde la línea editorial. En general se reportan percances asociados al antivacunismo -como el rebrote de enfermedades prevenibles que se creían controladas, o el arresto de militantes antivacunas en algún país- sin insinuar que la postura antivacunas merezca ser considerada. Con una excepción: el aspecto de la farándula, cuando se informa que una u otra persona famosa declara estar en contra de las vacunas, o que en algún show de TV se han invitado a exponentes pro y contra vacunas a exponer en igualdad de condiciones. Esto sugiere que un importante vector para la erosión de la confianza pública en la vacunación podría ser la atención dedicada a personalidades mediáticas que toman la causa del antivacunismo.

Contenido de la presentación

La presentación incluirá una breve introducción a la la problemática del antivacunismo, y un repaso al proyecto GDELT. También describirá los métodos utilizados para conectar R a GDELT para extraer noticias de medios locales sobre vacunas, y luego la técnicas de clasificación de contenido de noticias con *machine learning*. Los minutos finales serán dedicados a presentar resultados preliminares.