MIRA

Monitor de temas de Inmunización en Redes y medios de Argentina

METODOLOGÍA

para el análisis del discurso sobre vacunas en Twitter

VERSIÓN 1.0 (JUNIO 2021)

1. Fuente de datos (Twitter)

MIRA accede en tiempo real al *"stream"*[1] de Twitter, es decir la secuencia de mensajes originados en todo el mundo a medida que ocurren. MIRA utiliza el acceso oficial a Twitter, siguiendo los parámetros y reglas de uso que la plataforma exige.

En cada segundo, la comunidad global de usuarios genera miles de tuits. El acceso público de Twitter, utilizado por las entidades que recopilan esta información, limita la cantidad de tuits que se pueden recuperar en tiempo real. En la práctica, quienes monitorean la fuente pueden recuperar una fracción pequeña del total de mensajes, que se estima en 1%. Esta limitación se ve paliada al utilizar filtros de búsqueda, que permiten obtener sólo los tuits que cumplen determinadas condiciones. MIRA realiza una búsqueda por palabras clave, asegurando que todos los tuits de la muestra recibida segundo a segundo contengan términos de interés para el monitoreo de discurso sobre vacunas. En la práctica, dado que el discurso sobre cualquier tema en particular representa una pequeña fracción de todo lo que se dice en la plataforma, es posible asumir que todos o una gran parte de los tuits de interés son capturados con esta modalidad.

MIRA consulta en forma continua los sistemas de Twitter, recopilando cada tuit encontrado con alguna de las palabras clave en su contenido.

La lista de palabras clave es la siguiente:

'vacuna', 'vacunas', 'vacunar', 'vacunación', 'vacunacion', 'vacunados', 'vacunadas', 'vacunado', 'vacunada', 'vacuné', 'vacunen', 'vacunemos', 'vacunando', 'vacunaron'

Para determinar la procedencia del usuario utilizamos el campo user_location, que responde al lugar de residencia que cada usuario de la plataforma elige declarar. Como los términos de búsqueda son palabras del lenguaje español, los tuits acumulados provienen de hispanoparlantes en todo el mundo. Analizando el campo user_location se ha logrado determinar como Argentina la procedencia de algo más de un 12% del total de los tuits. Dado que muchos usuarios optan por especificar su provincia de residencia, es posible analizar los datos a ese nivel, e incluso a escala ciudad para ciertas urbes suficientemente representadas: Ciudad de Buenos Aires, Bahía Blanca, La Plata entre otras.

Los tuits son almacenados en una base de datos para su posterior análisis.

1. Puede consultarse la documentación de la API de Twitter y particularmente del stream en el siguiente link: https://developer.twitter.com/en/docs/tweets/filter-realtime/guides/connecting.html

2. Clasificación e interpretación de temas

Como puede inferirse, un primer problema a resolver es el volumen de información que se genera. En efecto, al 31 de Diciembre del 2020 contábamos con alrededor de 3.650.000 de tuits. Esta magnitud de texto parece imposible de ser abordada mediante técnicas manuales de lectura. Es por ello que se procedió a implementar algunos procesos y estrategias que permitieran automatizarla.

2.1 Primera estrategia: clasificación supervisada

En una primera etapa (año 2019), diseñamos una estrategia supervisada. Utilizando un método llamado fastText[2], parte de la familia de algoritmos de aprendizaje automático de tipo "supervisado". Este método requiere de un conjunto de ejemplos previamente clasificados -mediante intervención humana- de los cuales "aprender" una serie de atributos que diferencian una categoría de otra. Una vez realizado el aprendizaje, el modelo puede aplicarse a ejemplos de texto nuevos (nunca "vistos" por el algoritmo de clasificación) sobre los cuales estima la probabilidad de pertenecer a alguna de las categorías con las que fue entrenado; por ejemplo "antivacunismo", "faltante de vacunas", o "tema no relacionado".

2. Joulin, Armand, et al. "Bag of tricks for efficient text classification." *arXiv preprint arXiv:1607.01759* (2016).

Pero la irrupción de la pandemia de COVID-19 y, particularmente, del inicio de las investigaciones sobre el desarrollo de una vacuna, hizo que las discusiones en Twitter cambiaran profundamente. El clasificador automático entrenado desde un enfoque supervisado dejó de captar de forma fiable los temas de la discusión pública en Twitter. Esto implicaba que sería necesario reentrenar un modelo de clasificación, cosa que parecía poco viable en función de los tiempos y la escala. Es por esto que adoptamos una nueva estrategia.

2.2 Segunda estrategia: análisis no supervisado y modelado de tópicos

Se avanzó en el diseño de una metodología no supervisada que pudiera detectar de forma semi-automática los temas del corpus. Este tipo de tareas forma parte del dominio del modelado de tópicos, un conjunto amplio de métodos de procesamiento de lenguaje natural que, dado un corpus (sin etiquetar) de textos, busca detectar estructuras semánticas ocultas dentro del mismo. Dentro de las múltiples técnicas para la detección de tópicos, se optó por la denominada Latent Dirichlet Allocation (LDA), una de las más utilizadas.

Sabemos que un mismo tuit puede estar hablando de varios temas simultáneamente. Puede tocar temas vinculados a la gestión política de la vacunación, a los impactos que tiene en la sociedad y a la efectividad de las vacunas. LDA[3] trata de dar cuenta de esta complejidad asumiendo que cada documento (en nuestro caso, cada tuit) del corpus presenta una cierta composición de tópicos. Se asume que el "sentido" del documento puede modelarse como una distribución de probabilidades o proporciones. De esta forma, nuestro objetivo es predecir o detectar esa distribución en cada tuit: queremos saber en qué porcentaje un tuit habla de vacunas covid, en qué porcentaje lo hace sobre educación, etc.

3. Blei, D. (2012) "Probabilistic topic models". *Communications of the ACM* Vol. 55, N°4, p. 77-84.

Surge una pregunta obvia: ¿cómo se pueden definir e identificar los tópicos? Una vez más, nuestra intuición nos dice que es esperable que ciertas palabras se relacionen más con ciertos temas que otras. Así, en los tuits que hablan de la gestión política de la vacunación esperaríamos que palabras como "campaña", "aduana", "distribución" aparezcan con mayor frecuencia. En cambio, en los tuits que tocan el tema de las clases en la pandemia sería razonable asumir que aparezcan términos como "presencialidad," "clases", "docentes", etc.

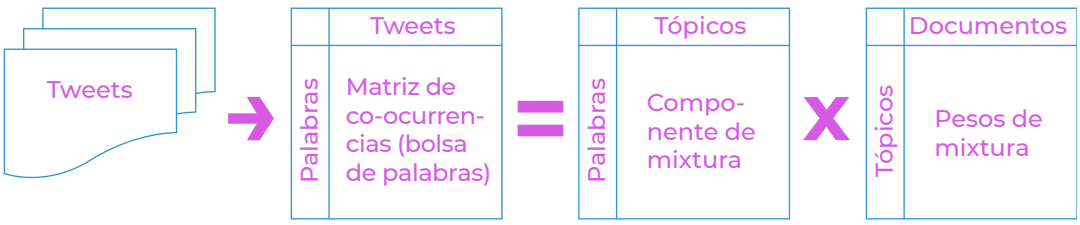
Una vez más, LDA intenta formalizar esta intuición en un modelo: cada tópico aparece definido como una distribución de probabilidades sobre la lista de todas las palabras únicas del corpus (llamado formalmente vocabulario). Así, cada palabra tendrá una cierta probabilidad de pertenecer a cada tópico detectado y esas distribuciones de probabilidad es otro de los parámetros que queremos estimar con LDA.

Ahora bien, hasta aquí, lo que esperamos obtener de LDA (distribución de palabras por tópicos y de tópicos por tuits), pero ¿qué datos es necesario proveerle a LDA para que haga su trabajo? Para ello, es necesario tomar el texto crudo de cada tuit y "vectorizarlo", es decir, representar esa información en un formato de filas y columnas. Hay muchas formas y modelos de vectorización de texto, la que utiliza LDA es la llamada "bag of words" (bolsa de palabras). La idea es construir una matriz en la que cada columna sea un documento (en este caso un tuit) y cada columna sea una palabra del vocabulario. A su vez, la intersección entre las filas y columnas contiene la cantidad de veces que esa palabra ocurre en el tuit[4].

Así, a partir de esta tabla (una matriz de co-ocurrencias de las palabras en los tuits) LDA busca producir dos output básicos:

- una tabla que nos dé la distribución de probabilidades de palabras para cada tópico, es decir, cuáles son las palabras más probables de cada tópico; esta información es fundamental para interpretar los tópicos (volveremos sobre esto enseguida)
- una tabla que nos permita cuantificar el peso de cada tópico en cada tuit; la idea es poder analizar en qué proporción habla cada tuit de cada tema; sobre esta tabla construiremos varios de los indicadores acerca de la evolución temporal de los temas

Es decir, podríamos resumir el proceso mediante el siguiente esquema:



Veamos un ejemplo simple. Supongamos que encontramos un tuit como el siguiente. Nuestro modelo de tópicos, estimado sobre la totalidad de los tuits, detecta algunas palabras que corresponden a dos de los tópicos: "controversias sobre las vacunas" y "vacunas covid". Estima la frecuencia de cada una de las palabras y calcula la prevalencia total de cada tema en el tuit.

4. En este trabajo, se utiliza una variante de ese modelo: en cada celda habrá un conteo de la cantidad de veces que aparece cada palabra en el texto pero ponderado por una métrica que se llama TF-IDF (por sus siglas en inglés Term frequency – Inverse document frequency). Se trata de una medida numérica que expresa cuán relevante es una palabra para un documento en una colección. El valor tf-idf está asociado al número de veces que una palabra aparece en el documento y también (con signo inverso) a la frecuencia de la palabra en el corpus. Esto permite balancear informatividad e importancia del término.

Se observa, entonces, que este tuit habla predominantemente de dos temas: las vacunas contra el covid y las discusiones a su alrededor.

Figura 1

Un problema a resolver tiene que ver con interpretación misma. El algoritmo devuelve algo similar a una probabilidad: la proporción en que cada tuit habla de cada tópico. Pero a su vez, cada tópico es una distribución de probabilidad sobre el vocabulario. Es decir una “bolsa de palabras”. De esta forma, para caracterizar e interpretar cada tópico se requiere intervención humana. El método no realiza una “interpretación” de estas palabras. Si encontráramos un tópico que concentra su probabilidad en las siguientes palabras: “efectividad”, “testeos”, “rebaño”, “inmunización”, etc. sería tarea del analista interpretar este conjunto de palabras como un tópico referido a la *efectividad* de las vacunas.

Esta operación (al igual que cualquier clasificación manual) tiene una carga potencial de ambigüedad. Es por ello que se avanzó en una etapa de etiquetado manual de los tópicos.

3. Flujo de trabajo

A continuación se detallan el flujo de trabajo realizado para la base de datos de tuits:

1. en primer lugar, se realizó un pre procesamiento estándar del texto:
    - se normalizó a minúsculas,
    - se eliminaron dígitos, sitios web, caracteres extraños y puntuación,
    - se reemplazaron los emojis por texto que lo describe ("pulgares_arriba", "cara_feliz", etc.),
    - se eliminaron stopwords tanto por lista (preposiciones, artículos, etc.) como por frecuencia (aquellas palabras de alta ocurrencia en el texto),
    - se construyó una matriz de co-ocurrencias de términos de unigramas para documento, ponderada por TF-IDF,
2. se dividió el dataset en bimestres,
3. luego de algunas pruebas se decidió entrenar para cada bimestre de forma independiente un modelo con 20 tópicos,
4. cada uno de los 10 modelos (uno por cada bimestre, desde el bimestre mayo-junio de 2019 al bimestre noviembre-diciembre de 2020) fueron etiquetados de la siguiente manera:
    - se analizaron las palabras de mayor probabilidad para cada tópico,
    - se leyó una muestra de los 20 tuits con mayor valor de cada tópico,
    - luego se agruparon de forma manual los 20 tópicos definidos en cada bimestre en 9 temas agregados,
    - En este análisis intervinieron tres personas. Se identificaron 9 temas principales relacionados con vacunas, luego de eliminar los tópicos asignados como "no relacionado" para los casos en los que los tuits no estaban hablando de vacunas humanas (ej. carne vacuna, vacunas de mascotas, etc.).
    - A cada tópico se le asignó un tema en base a las palabras de mayor probabilidad. Estos temas nos ayudan a entender de qué trata el tópico desde una perspectiva general. Entre los temas identificados encontramos a: *vacunas general, vacunas covid, vacunas no covid, controversias sobre vacunas, gestión, gestión otros países, política, educación, y sociedad*. Los temas se pueden repetir, en mayor o menor medida, a lo largo de todos los bimestres. Cada una de estos temas fueron definidos por los investigadores[5].

5. Esta metodología de trabajo entra dentro de lo que se ha propuesto llamar "teoría fundamentada computacional". Nelson, Laura K. 2020. "Computational Grounded Theory: A Methodological Framework", *Sociological Methods and Research*, 49 (1): 3-42.

Tabla 1. Listado de temas (tópicos agrupados), su interpretación y ejemplos de palabras

Tema	Descripción	Ej. de palabras frecuentes
Vacunas general	Valor de las vacunas, tanto vacunas generales (ej. calendario) como vacunas de covid. Vacunas que no son de covid (ej las del calendario): valor, efectividad, etc.	efecto rebaño, gripe, anticuerpos, sarampión, rubeola, calendario.
Vacunas covid	Vacunas para COVID-19, investigación, efectividad, funcionamiento.	fase, coronavirus, efecto rebaño, oxford
Controversia sobre vacunas	Posturas antivacunas o de reticencia a la vacunación. Tuits tanto a favor como en contra de estas posturas.	dióxido de cloro, conejillos de indias, bill gates, terraplanistas, chip
Gestión	Gestión política relacionada con las vacunas en Argentina	campaña, aduana, cadena de frio,
Gestión otros países	Gestión política relacionada con las vacunas en otros países	rusia, chile, eeuu, trump
Política	Mención a gobernantes, partidos, referentes de Argentina. Fuerte contenido político tanto a favor como en contra	alberto fernandez, kicillof, rubinstein, intendente, comunista, macrismo
Educación	Presencialidad en escuelas, vacunas a docentes	presencialidad, trotta, clases, docentes
Sociedad	Decisiones y actitudes de la sociedad. Opinión de los ciudadanos no en relación a la política sino más al cuidado	barbijo, alcohol, cuidandonos, medidas

4. Construcción de indicadores

Para la construcción de los indicadores a analizar tomamos como input principal las tablas de mixtura agregada (la distribución de la prevalencia se cada tema agregado para cada tuit) de cada bimestre y se apilaron en una sola matriz. La estructura de datos con la que se trabaja es la siguiente:

Tabla 2. Pesos de mixtura de documentos por temas agregados

id	bimestre	gestión	no relacionado	política	controversia sobre vacunas	...	educación
1	2019_3	0.6	0.01	0.01	0.2	...	0.1
2	2019_3	0.1	0.8	0.01	0.01	...	0.01
...	...	...	...	...	...	...	...
n	2020_6	0.01	0.01	0.1	0.7	...	0.1

- cada fila es un tuit
- cada columna es un tema agregado
- cada celda es el valor que el tuit tiene en el tema correspondiente.

Así, en el ejemplo hipotético anterior, el primer tuit tiene una alta prevalencia (0.6) en el tema gestión. Mientras que el último tiene una alta prevalencia en el tema controversias sobre vacunas.

Cada fila suma 1, por lo cual la estructura de temas puede pensarse como una probabilidad.

A partir de esta estructura se calculan la siguientes métricas:

1. **Promedio de importancia de temas por bimestre**: se calcula para cada bimestre el promedio de cada columna de tópicos.

Este indicador refleja el “peso promedio” que cada tema tiene en cada bimestre, y nos indica, de algún modo, cuánto se habla de ese tema en ese período, en relación a los demás temas.

2. **Conteo de tema mayoritario**: para cada tuit se identifica cuál es la columna que mayor prevalencia tiene (ese es el tema mayoritario). Se consideran todos los tópicos mayoritarios, independientemente de su prevalencia. Es decir que se considera mayoritario un tópico con prevalencia de 0.1 y uno de 0.5. A su vez, para cada bimestre se realiza el conteo de los tópicos mayoritarios de cada tuit.

Este indicador muestra qué proporción de los tuits tienen cada tema como mayoritario, en cada bimestre.