Análisis de tendencias sociales.

Enrique Vilanova Vidal[*] Brais Suarez Souto [**]

5 de noviembre de 2020

[*] Todos los que han colaborado en el proyecto

[**] Las comunidades online que nos han permitido realizar este trabajo

1. Contexto

En una sociedad en constante y rápido cambio es importante entender las *tendencias sociales* de un momento en particular y poder pronosticar su posible evolución. Entendiendo por *tendencia social* algo que es importante para un segmento de la población o mercado en un momento dado.

¿Qué hace a algo relevante para la sociedad? esta pregunta clásica está en el corazón de muchas disciplinas sociales. Disciplinas que van desde la *sociología* puramente académica hasta disciplinas aplicadas cómo *marketing publicitario*, para todas ellas es fundamental encontrar algún tipo de respuesta a la pregunta acerca de lo que interesa a la sociedad.

Esta pregunta ha sido abordada hasta la fecha, desde un punto de vista matemático, por modelos *theory-driven* [1, 2, 3]. Los recientes avances en inteligencia artificial y la abundancia de información disponible en Internet, hace posible intentar responder a la misma pregunta clásica, pero esta vez con modelos *data-driven*[4].

Contestar de forma general, ¿Qué hace a algo relevante para la sociedad? es un primer paso muy ambicioso por lo que es razonable elegir un contexto acotado, por lo que, en concreto nos preguntamos por:

- España
- Noticias
- Actualidad

En este sentido, consideramos relevante obtener información de agregadores de noticias y foros o redes sociales, donde sea razonable pensar que son los usuarios los que eligen qué tema es relevante.

Meneame es el mayor agregador de noticias de España. Obteniendo datos de su sección *portada* podemos tener información de las noticias de actualidad que han sido relevantes. Además, Meneame ofrece una sección *nuevas* donde podemos ver la mayoría de las noticias subidas por los usuarios. Comparando *portada* y *nuevas* podemos intentar entender las diferencias entre las noticias que llegaron a portada y las que no. *A priori*, entendemos que las noticias que llegan a *portada* han sido más relevantes para la mayoría de los usuarios.

Utilizaremos otras webs que funcionan como agregadores de noticias, con el fin de minimizar el sesgo proveniente del hecho de que los usuarios de Meneame pueden estar concentrados en un segmento específico de la población. Además también se estudiará el impacto de la noticia en redes sociales.

Para el propósito de esta práctica presentamos los datos provenientes de:

- Meneame: *portada*.
- Reddit: *subreddits de noticias de actualidad sobre España*.
- Twitter: las *tags* extraídas de las noticias se usan para estudiar el impacto de la noticia.

La información de Meneame se ha extraído con *web scrapping* (pese a contar con API) como ejercicio académico (practicar). Además, puede ser interesante usar *web scrapping* como forma de evitar las limitaciones impuestas por las API.

2. Títulos

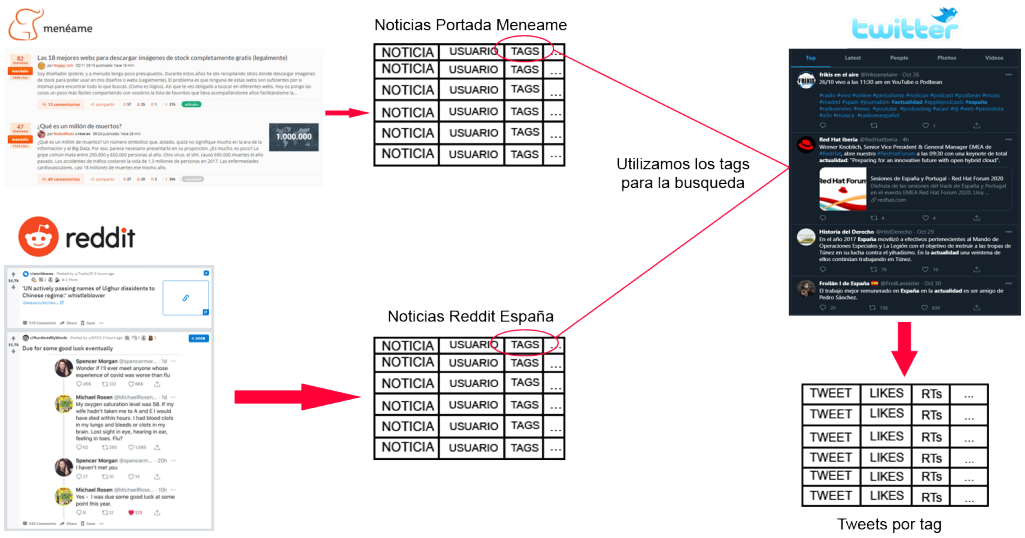
- Noticias *portada* Meneame.
- Noticias Reddit España.
- Tweets por *tag*.

3. Datasets

Se trata de tres datasets, provenientes de tres fuentes diferentes, Meneame, Twitter, y Reddit. Para los tres datasets se ha capturado el máximo de información relevante con el fin de realizar un estudio de la relevancia de la noticia. Entre otras cosas se pretende estudiar:

- El tiempo que tarde una noticia en convertirse en relevante.
- Los factores comunes en noticias relevantes.
- Número de usuarios relacionados con la noticia (comentan, twittean...)
- Sentimiento en la noticia y hacia la noticia.
- Reconocimiento *nombre-entidad* en la noticia.
- Estratificación de usuarios de acuerdo a una noticia
- Estudio comparativo usuarios con distribuciones población extraídos de otras fuentes (INE).

4. Representación gráfica.



5. Contenido

5.1. Meneame

El dataset para Meneame se ha hecho utilizando *Python* y la librería *Beautifulsoup*. Meneame dispone de una API, pero por motivos académicos y con la intención de presentar un trabajo lo más completo posible, para el caso de Meneame no se ha utilizado la API.

Cabe remarcar, que tal como se puede ver en el código del scrapper, para cada noticia se genera un segundo dataset con los comentarios de cada usuario para una noticia en concreto, así como otra información relevante de la noticia. Se elige esta aproximación para mantener el dataset principal legible, siendo el dataset principal el obtenido de Meneame *portada*. Toda la información puede accederse a través de una base de datos relacional. Aunque los datasets específcos de cada noticia no se incluyen en el trabajo.

- Código: clave principal de la noticia.
- Titular: titular de la noticia.
- Fecha de envío: hora a la que noticia es enviada a Meneame.
- Publicada: hora a la que la noticia es publicada
- Web origen: origen de la noticia

- Código usuario: clave numérica foránea que permite identificar a usuarios
- Nick: nombre del usuario
- Tema principal: primera etiqueta descriptiva de la noticia
- Columnas Sub Tema: todas las otras posibles clasificaciones que el usuario aporta para una noticia
- Clicks: número de veces que se ha clicado en la noticia
- Positivos": número de votos positivos ofrecidos por usuarios registrados
- Anónimos: número de votos positivos ofrecidos por usuarios no registrados
- Negativos: número de votos negativos
- Karma: factor de relevancia numérico asignado por Meneame
- Número de comentarios: número de comentarios que tiene una noticia

5.2. Reddit

El objetivo de este dataset es hacer una comprobación cruzada con las noticias extraídas de Meneme y ver si también han sido relevantes en Reddit y que grado de relevancia han tenido.También se quiere buscar posibles noticias relevantes que no hayan llegado a *portada* de Meneame pero si hayan sido relevantes. Además de comparar el sentimiento de los comentarios hacia las noticias (entre otras cosas).

Desafortunadamente, Reddit tiene una repercusión (en número de usuarios activos) menor que Meneame. Para evaluar los datos, esto añade la dificultad de definir un factor de peso para las noticias dependiendo del medio.

En el código del scrapper también se ha tenido en cuenta que Reddit no tiene sección *portada* y *nuevas* (como Meneame) por lo que se ha incluido una primer filtro donde únicamnete se seleccionan noticias que han tenido una cierta cantidad de votos positivos (sería nuestro equivalente a *portada*). Esto se ha hecho por motivos de consistencia con el dataset que se entrega de Meneame, ya que este solo incluye las noticias de *portada*. Otra cosa a tener en cuenta es que las noticias de Reddit no tienen *tags* proporcionadas por los usuarios, sin embargo los subreddits escogidos, podrían ser todos categorizados por las *tags*; noticias, actualidad, España.

Al igual que en el caso anterior, en el código del scrapper se puede ver como se genera un segundo dataset con todos los comentarios de cada noticia y otra información relevante, que no se presenta pero se incluye en nuestra base de datos relacional.

Para este caso y por sencillez hemos escogido realizar el scrapping usando la API de Reddit.

Descripción de los atributos:

- Id del post: clave principal para identificar el post.
- Título: título del post.
- ups: votos positivos o ups.
- downs: vostos negativos o downs.
- Número de comentarios: cantidad de comentarios para una noticia.
- Autor: autor del post.
- Fecha: fecha en la que se creó el post.
- link: enlace a la web de origen de la noticia.
- subreddit: subreddit donde se ha extraído el post.

5.3. Twitter

El scrapper de Twitter se utiliza para seguir la relevancia de las noticias en un medio generalista. Se usan las *tags* extraídas de las noticias para comprobar su relevancia. En primer paso, las tags se seleccionan manualmente y se compruebará de forma manual que los twits correspondan a cierto tema. El propósito final es automatizar todo el proceso, lo que presenta el desafío de comprobar que ciertos tweets corresponden a un tema en concreto. Sin entrar en detalles, para afrontar este desafío se usará redes neuronales para la identificación de *nombre-entidad* en el cuerpo de las noticias, con el fin de generar vectores de proximidad, así como *contextual topic recognition*.

Para el propósito de esta práctica se entrega el dataset correspondiente a dos noticias, donde las *tags* se han seleccionado (manualmente), a partir de las *tags* que los usuarios de Meneame proporcionan para describir una noticia.

Descripción del dataset:

- query: La query usada para generar el dataset.

- Tweet ID: ID del Tweet.
- Text: Texto del Tweet.
- RT Count: Número de veces que el Tweet ha sido retweeteado.
- Username: Nombre de usuario del autor del Tweet.
- Likes: Número de veces que el Tweet ha recibido un like.
- Created at: fecha en la que se creó el Tweet.

5.4. Otros

A medida que avance el proyecto otros foros y plataformas se irán incluyendo, cabe mencionar especialmente *Facebook* por su cantidad de usuarios.

Es importante mencionar, que con el fin de estimar apropiadamente, las distribuciones de población (entre otros), se consultarán bases de datos del *INE*.

6. Agradecimientos

Agradecimientos a las plataformas de Meneame, Reddit y Twitter por permitirnos utilizar sus plataformas y recursos para obtención de datos generados en ellas.

Nos gustaría agradecer especialmente a los miembros de su comunidad por su participación en las distintas plataformas ya que hacen posible este tipo de trabajos que una década atrás hubiese requerido de una gran cantidad de capital para entrevistar a una parte significativa de la población.

6.1. Reddit y Twitter

El scrapping de estos sitios webs se ha realizado a través de sus API para python. En ambos casos se ha utilizado una cuenta de desarrollador y nos hemos mantenido dentro de las limitaciones especificadas (número de peticiones por minuto y número de twits por hora)

6.2. Meneame

Para el caso Meneame nos hemos mantenido dentro de los límites especificados en el archivo robots.txt. En este archivo, se permite todos los *user agent*. No ha sido necesario acceder a ninguna de las carpetas que figuran como *Disallow*.

7. Inspiración

Cómo ya se ha mencionado la pregunta que se intenta responder es ¿Qué hace a algo relevante para la sociedad?, pero también se quiere averiguar ¿Qué temas son relevantes para la sociedad, en este momento?, ¿Por cuanto tiempo se mantiene relevante? Este tipo de preguntas son fundamentales en muchas áreas, tanto académicas como relacionadas con los negocios. Poder contestar satisfactoriamente estas preguntas, podría permitir, entre otras cosas, planificar adecuadamente campañas publicitarias dirigidas a un segmento de la población y sabiendo cuanto tiempo han de extenderse en el tiempo. Así cómo reaccionar en tiempo real al cambio de opiniones y gustos.

8. Licencia

Todo el material entregado en esta práctica está bajo licencia:

- Released Under CC0: Public Domain License

Debido a que nosotros estamos formándonos, y estamos utilizando contenido generado por comunidades de usuarios abiertas y todavía no contiene ningún análisis, vamos a lanzar nuestro dataset con licencia CC0, por lo tanto rechazamos cualquier tipo de derechos sobre la autoría del mismo, para que cualquier persona interesada pueda trabajar con él y hacerlo suyo, sin necesidad de darnos ningún tipo de atribución. Más información se puede encontrar en la página oficial de CC. https://creativecommons.org/share-your-work/public-domain/cc0/

9. Código

Todo el código utilizado para el desarrollo de los scrappers puede encontrarse en la carpeta Scrapper del repositorio de gitHub.

10. Dataset

Enlaces:

- Siguiendo el Link se pueden encontrar los datasets de Reddit y Meneame con DOI 10.5281/zenodo.4243130 subidos a Zenodo. El dataset de Twitter no se ha subido a Zenodo ya que tiene un cáracter de análisis mas específico, pero se ha incluido dentro del repositorio de github.

- Usando este link se puede encontrar el repositorio de GitHub dónde se han subido los últimos datasets generados.

11. Participación

Actividad	Integrante 1	Integrante 2
Investigación previa	Enrique Vilanova	Brais Suárez
Redacción de las respuestas	Enrique Vilanova	Brais Suárez
Desarrollo código	Enrique Vilanova	Brais Suárez

Referencias

[1] Leonard Berzkowitz, *Advances in Experimental Social Psychology*, 1967, Academic press

[2] Richard McElreath, Robert Boyd, *Mathematical Models of Social Evolution: A Guide for the Perplexed*, 2007, The University of Chicago Press

[3] Thomas L. Saaty and Joyce M. Alexander,*Thinking with models: Mathematical Models in the Physical, Biological, and Social Sciences*, 2015, RWS Publications

[4] Jason Radford and Kenneth Joseph,*Theory In, Theory Out: The Uses of Social Theory in Machine Learning for Social Science*, 2020, doi.org/10.3389/fdata.2020.00018