

Recuperación de Información y Recomendaciones en la Web 2022

Análisis y Clasificación de Tweets

Informe final

Grupo 13

Integrantes

Nicolás Ribero - CI: 4.590.761-1

Cristhian D'Alessandro - CI: 5.290.469-2

Franco Serratto - CI: 5.266.237-3

Docente Libertad Tansini

Problema a abordar	3
Enfoque de la solución	3
Diseño	4
Funcionalidades y uso	6
Evaluación y resultados	8
Conclusiones	8
Trabajo Futuro	9
Referencias	9

Problema a abordar

Las redes sociales son una fuente interminable y variada de contenidos. En particular nos enfocaremos en Twitter, esta red social que tiene millones de usuarios generando información de forma constante nos da infinitas posibilidades de analizar y extraer datos.

Hoy en día, Twitter es un termómetro, o una muestra parcial, de las cosas más relevantes que están pasando en la actualidad a nivel local o global. Sobre un tema en particular, se pueden generar miles de opiniones al respecto, opiniones que pueden ser tanto negativas como positivas.

En estas opiniones nos vamos a centrar en nuestro proyecto. Nos interesa analizar cuál es el sentimiento del público dado un tema de interés. Por ejemplo, qué opina la gente sobre el lanzamiento de un nuevo producto al mercado o medir qué aceptación tiene cierta decisión política. Además, nos interesa saber de dónde llega esa opinión, saber si es del interior, de la capital, o diferenciarla por departamento.

Enfoque de la solución

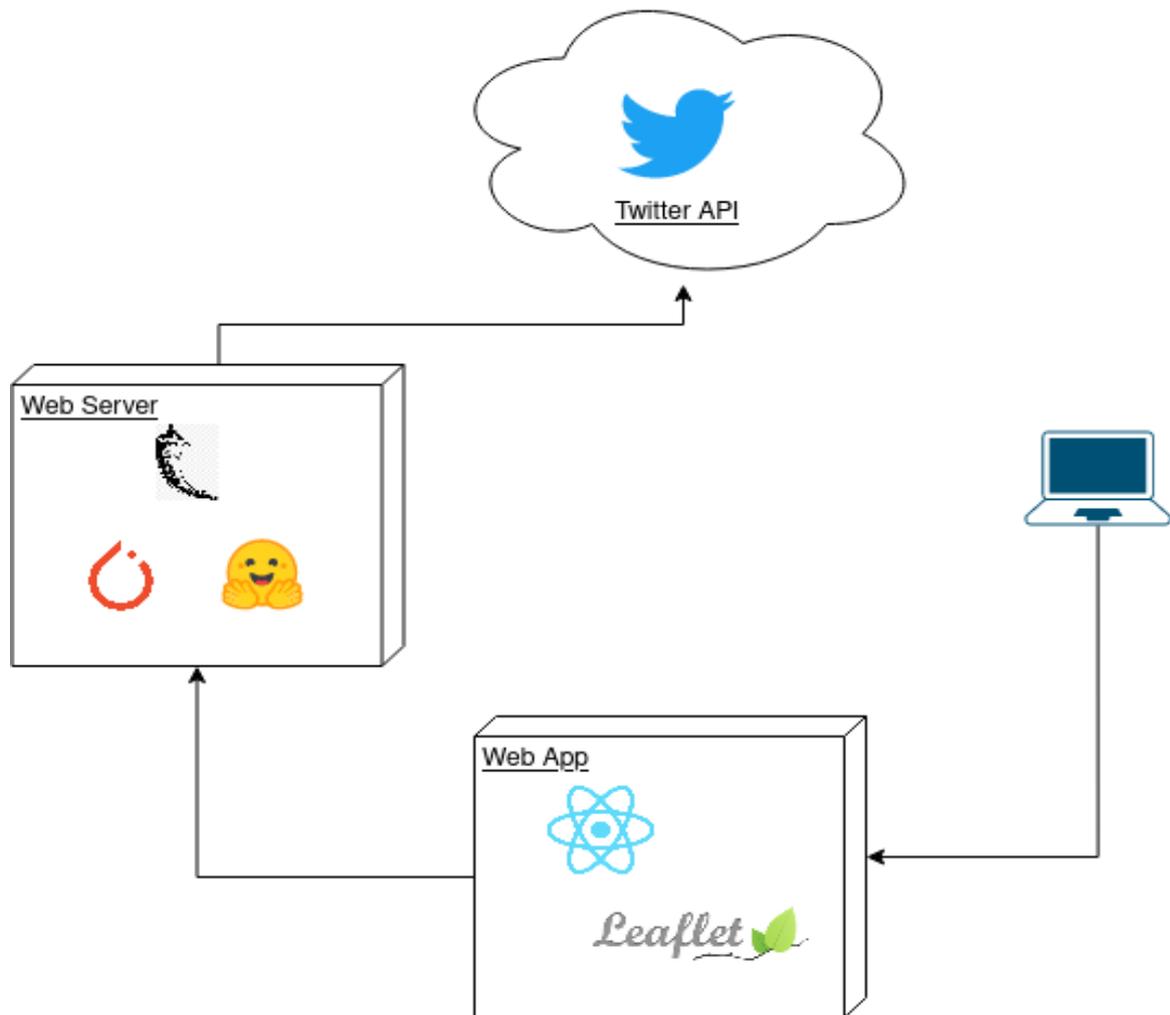
La plataforma Twitter ha implementado una API muy completa que nos permite no solo acceder a la información, sino que nos brinda enormes posibilidades de recuperación de la misma.

Proponemos crear una aplicación web que permita ingresar una palabra o varias (sin espacios) y a partir de ella recuperar los tweets que contengan *#<palabraingresada>*.

En base a esos tweets recuperados, se realizará el análisis categorizando los tweets en positivos, negativos y neutros. También la aplicación permitirá visualizar en un mapa, la ubicación desde donde se realizaron las publicaciones.

Diseño

En esta sección detallaremos las tecnologías y componentes utilizados durante el desarrollo de la aplicación, así como el diagrama de la misma.



Flask Web Framework

Un módulo de python que permite desarrollar aplicaciones web de manera sencilla.

El mismo será utilizado para desplegar una API REST que se encargará de traer la información desde la API de Twitter para posteriormente ser procesada por un modelo de procesamiento de lenguaje natural.

PyTorch

Un paquete de Python que provee 2 funcionalidades importantes de alto nivel

- Cálculo de tensores (como NumPy) con aceleración por GPU
- Redes neuronales profundas construidas en un sistema de auto graduación basado en cinta

Modelo XLM-roBERTa-base

En particular utilizaremos un modelo provisto por la comunidad de Hugging Face pre entrenado específicamente para la clasificación de tweets.

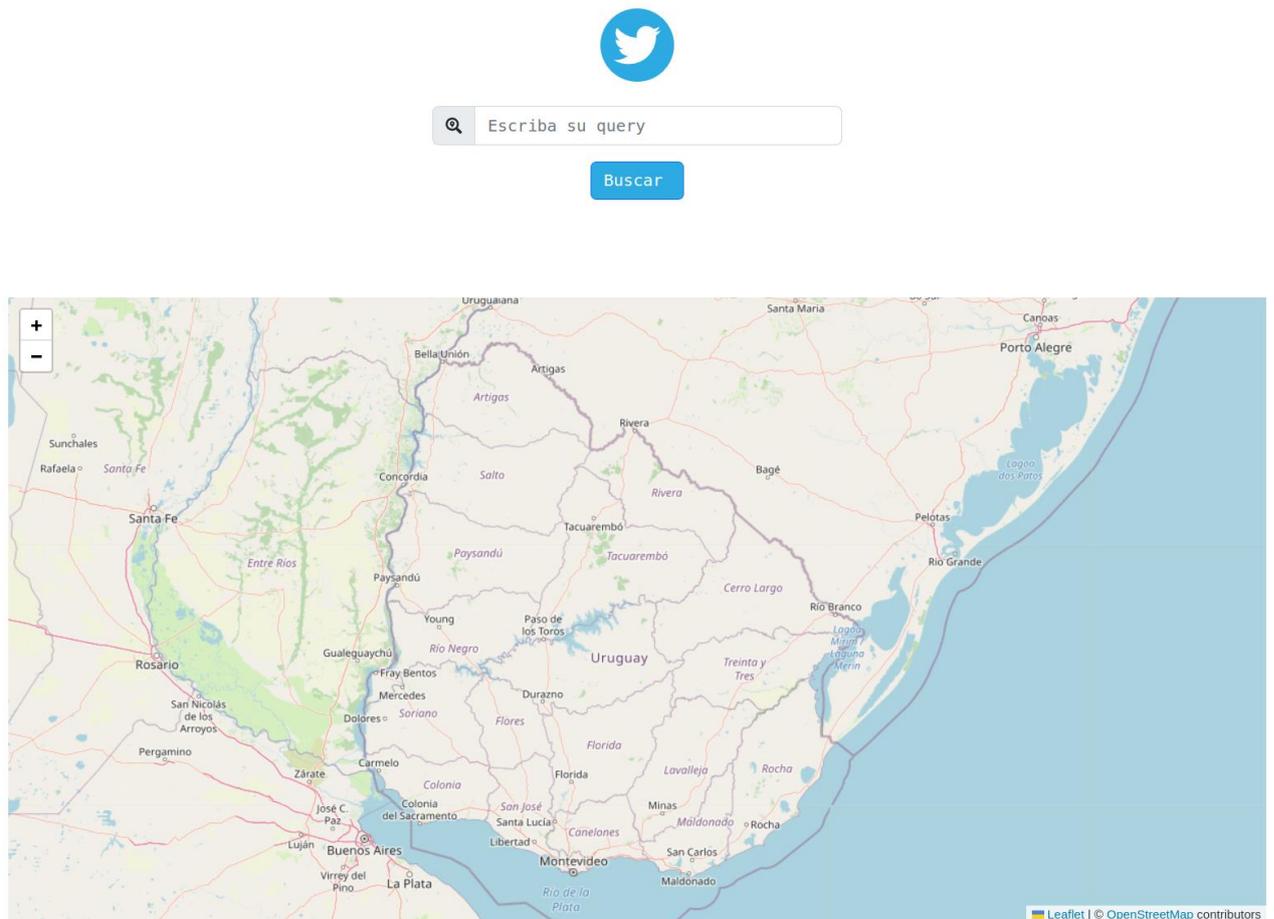
El modelo que utilizaremos específicamente es: `pysentimiento/robertuito-sentiment-analysis`

React

Un framework frontend para Javascript que utilizaremos para desplegar información relevante sobre una búsqueda particular. Esta aplicación consumirá directamente desde el backend desarrollado en Flask para obtener dicha información.

Funcionalidades y uso

Al ingresar a la aplicación veremos una barra de búsqueda en la que podremos ingresar la palabra o hashtag a consultar.



Luego de realizar la búsqueda, se puede ver la cantidad de resultados positivos, negativos y neutros recuperados. Esta funcionalidad utiliza por detrás un endpoint particular provisto por la API de Twitter que limita los tweets obtenidos a un marco de tiempo en los últimos 7 días.



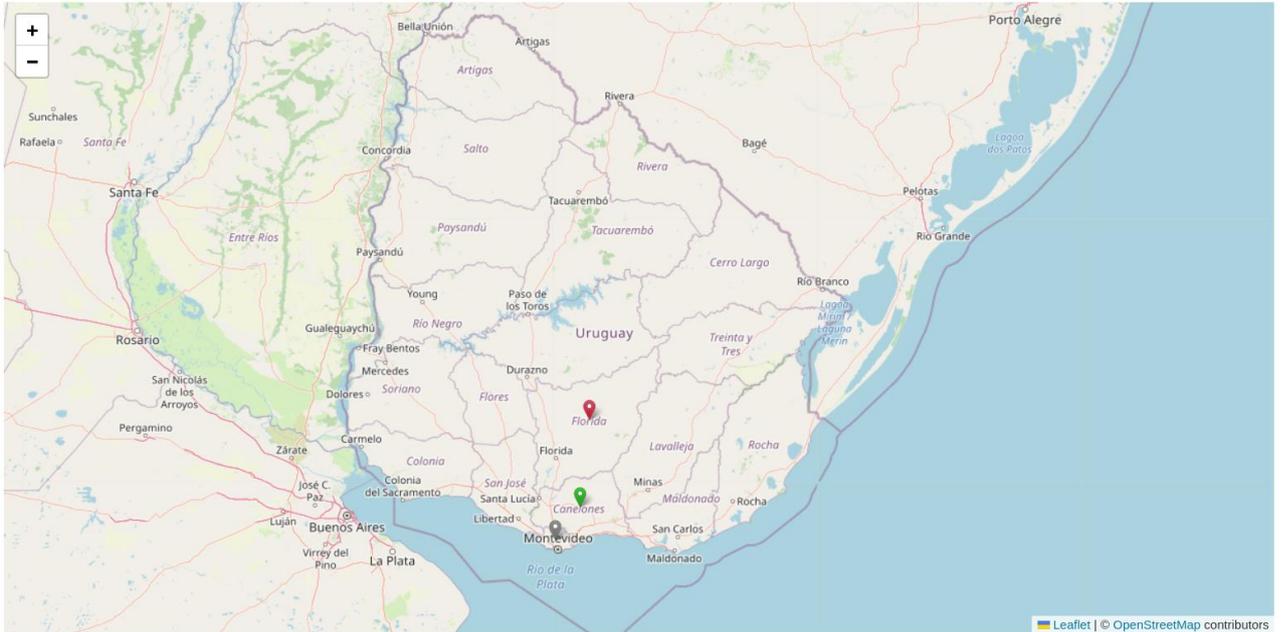
🔍 #Uruguay

Buscar

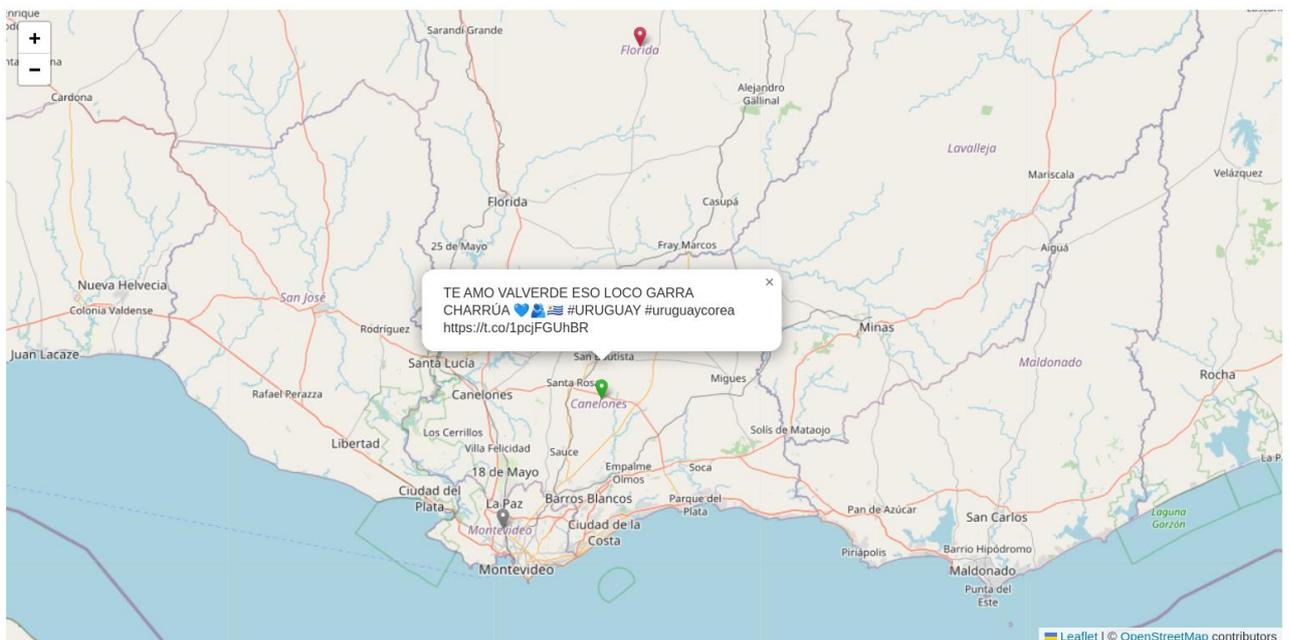
✓
1

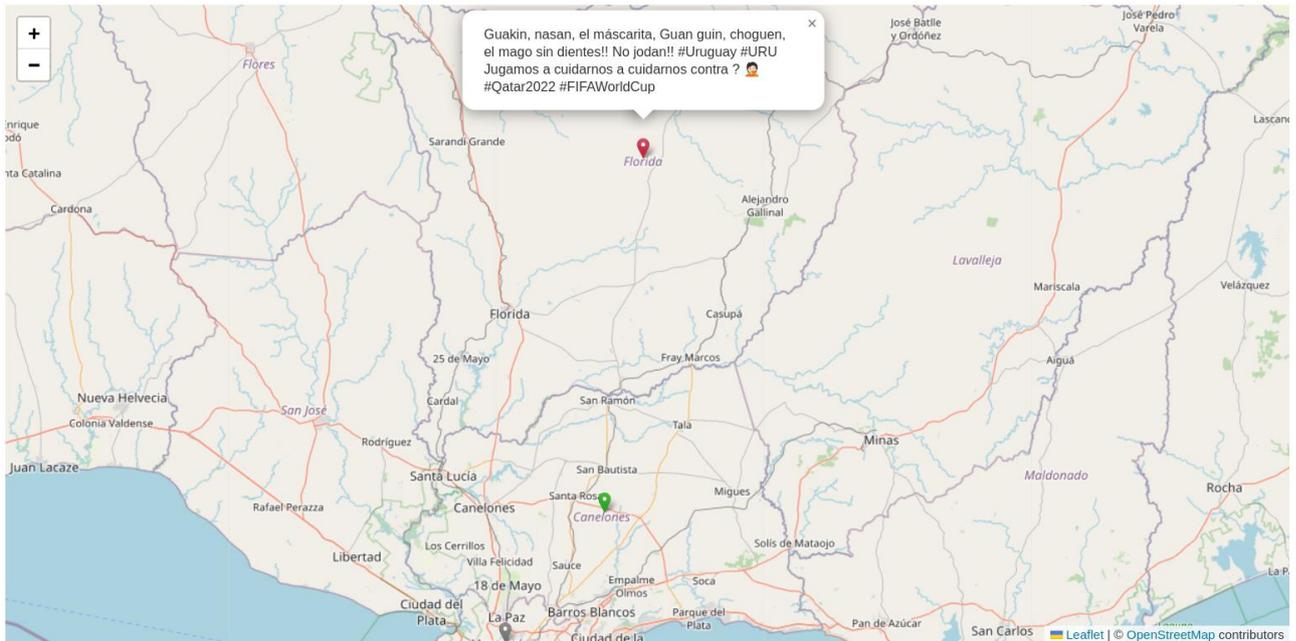
🗨️
1

✕
1



Además podemos recorrer el mapa viendo la ubicación de cada tweet identificados con un color según el resultado y ver el post completo al hacer click en dicha ubicación.





Evaluación y resultados

Debido a los ajustes que se debieron realizar dadas las limitaciones que la API gratuita de Twitter supone y el enfoque que le dimos a este proyecto se obtienen resultados variables dependiendo de la búsqueda a realizar.

Aquellos términos que son más utilizados dentro del territorio nacional que fuera de él acarrearán resultados más interesantes y diversos, como por ejemplo búsquedas con los hashtag: *#Peñarol*, *#CNdeF*, *#Montevideo*, *#UruguayNatural*.

Mientras que otro tipo de consultas que involucren temas más generales arrojan muy pocos, o incluso en algunos casos, ningún resultado.

Conclusiones

Llegamos a una aplicación interesante que nos permite conocer a muy grandes rasgos el sentimiento general sobre tópicos de interés en el Uruguay.

Aunque esto se da con un sesgo importante debido a las limitaciones que nosotros mismos impusimos a la hora de considerar únicamente aquellos tweets que cuentan con información geográfica (dado que en la mayoría de los casos, el autor no provee estos datos).

Aún así, de ser mejor refinada, la aplicación tiene el potencial de convertirse en una herramienta atractiva y bastante útil (particularmente si se consiguiera hacer uso de la API paga de Twitter para poder mejorar la búsqueda de manera significativa).

Trabajo Futuro

Sería interesante no restringir la búsqueda sólo a la red social twitter. Por un lado, hacer la comparación de los resultados de distintas redes sociales podría brindarnos información valiosa, ya que los perfiles de las personas que usan una u otra red pueden ser muy distintos. Ejemplo de ello puede ser las distintas edades que prefieren utilizar una plataforma u otra, también la cantidad de perfiles que ocultan su verdadera identidad.

Por otro lado sería atractivo tener acceso a la API paga de Twitter para aprovechar otro tipo de funcionalidades que la versión gratuita no provee, así como también la posibilidad de trabajar con otras API.

Otro aspecto a tener en cuenta es la performance de la aplicación, dado que se podría paralelizar la tarea de análisis de sentimiento en un cluster por ejemplo, a modo de reducir los tiempos de respuesta de la misma.

Por último, se podría almacenar esta información en una base de datos y realizar consultas sobre algún tema en particular a lo largo del tiempo para tener un registro de cómo fue evolucionando un tópico de interés durante dicho período.

Referencias

- Modelo ya entrenado de análisis de sentimiento orientado a tweets
<https://huggingface.co/pysentimiento/robertuito-sentiment-analysis>
- Documentación de la API de Twitter
<https://developer.twitter.com/en/docs/twitter-api>
- Documentación de la librería de Tweepy para Python
<https://docs.tweepy.org/en/stable/>
- Documentación de la librería Leaflet para la representación de mapas con javascript
<https://leafletjs.com/index.html>
- API para el componente de mapas con Leaflet en aplicación React
<https://react-leaflet.js.org/docs/api-map/>