

Reconocimiento de entidades en textos de redes sociales

Autoría: Manuel Mongelós: manuel.mongelos@correo.ucu.edu.uy | **Tutoría:** Bernardo Rychtenberg: brychtenberg@ucu.edu.uy
Departamento de Informática y Ciencias de la Computación - UCU

1- Introducción

A diario se realizan una gran cantidad de posteos en la red social Twitter, donde los usuarios comparten información en tiempo real de diversos temas entre ellos de las cosas que suceden a sus alrededores. Estos datos pueden ser valiosos para dar seguimiento a cualquier suceso que pueda estar ocurriendo en una ubicación específica, e incluso responder a este. Twitter ofrece la posibilidad al usuario de brindar su ubicación actual, adjuntándose esta en la metadata del tweet, pero para los casos donde no esté disponible es que se propuso en esta investigación probar modelos para lograr inferir esa información.

2- Trabajos anteriores

Como parte del grupo de investigación del Departamento de Informática, se viene desarrollando un framework que toma información de diversas fuentes de datos, las integra y eventualmente son insumo de algún modelo de Machine Learning. Este último caso es el motivo de la presente investigación, donde a partir de datos descargados con módulos desarrollados previamente (Twitter), se le aplica un modelo de Machine Learning con el objetivo de extraer información.

3- Descripción

Frente al planteo de cómo lograr discernir la ubicación del usuario a partir de lo dicho en un tweet, se probó un modelo de inteligencia artificial que trabaja en base al reconocimiento de entidades nombradas (NER) y sus resultados frente a distintos textos. El modelo empleado fue “*mrm8488/bert-spanish-cased-finetuned-ner*”¹, el cual parte del modelo *BETO*², un modelo entrenado sobre un aproximado de tres mil millones de líneas tomadas de documentos en español, que luego se ajustó para la tarea del reconocimiento de entidades en textos.

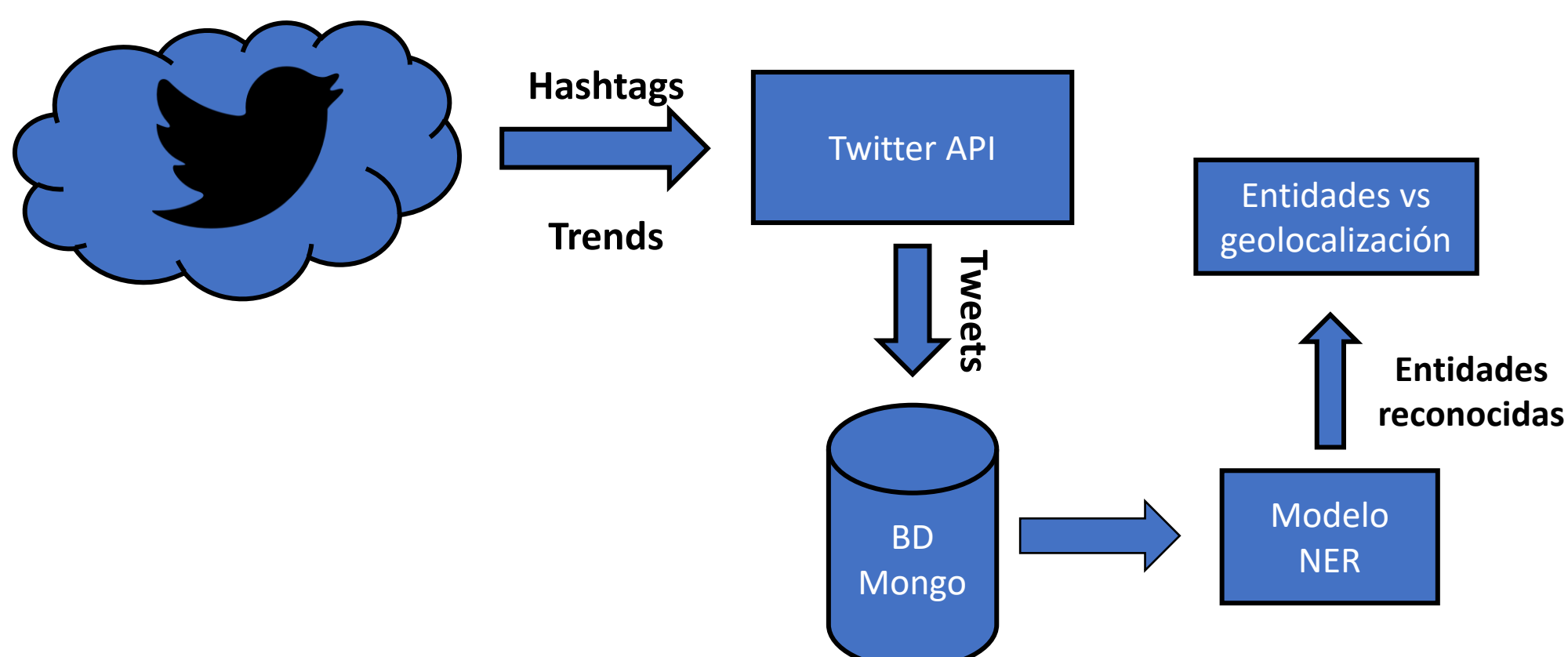


Figura 1: Diagrama del proceso de la aplicación

Bibliografía

1. Manuel Romero. Spanish BERT (BETO) + NER <https://huggingface.co/mrm8488/bert-spanish-cased-finetuned-ner>
2. Dccuchile. BETO: <https://github.com/dccuchile/beto>
3. Hugging Faces. Fine-Tune a pretrained model <https://huggingface.co/docs/transformers/training>
4. Xiaohua Liu, Shaodian Zhang, Furu Wei, Ming Zhou, Recognizing named entities in tweets. (2011)

El reconocimiento de entidades nombradas nos permite extraer cualquier localización que pueda haber descrito el usuario, sea un país, ciudad, etc., y el experimento consistió en tomar esas entidades y comprobar si efectivamente correspondían a la ubicación real. Para los tweets que contaban con los metadatos necesarios, se realizó un algoritmo que comparaba estos y las entidades que haya extraído el modelo. Los que no disponían de esta información, fueron comprobados de manera manual, buscando si el usuario había descrito de forma explícita que su ubicación fuese esa.

4- Resultados

Se analizaron 9.740 tweets, de los cuales 996 contaban con entidades reconocibles. De estos 996, 925 sí disponían de metadatos que brindan la ubicación del usuario, dejando solo 71 tweets para la verificación manual.

La comparación automatizada mostró que la eficacia del modelo elegido fue del 34%, acertando la geolocalización del usuario en 311 tweets.

Por otro lado, expuso una precisión del 25% para los tweets que no contaban con localización activa, considerándose este como un buen resultado pero que podría mejorar con el empleo de otras técnicas, detalladas a continuación en las conclusiones.

5- Conclusiones

Una de las principales limitantes fue el escaso texto de los tweets, con un máximo de solo 280 caracteres. Esto redujo de manera significativa la efectividad del modelo, siendo que este funciona mejor en textos de mayor longitud, lo que permite inferir de forma más acertada el contexto y por ende mejora la predicción. Además, la poca cantidad de palabras puede derivar en formas de comunicación propias del contexto de las redes, como el uso de abreviaciones, que el modelo puede no conocer si partimos de que sus datos de entrenamiento son de documentos de internet. Esto podría haberse visto contenido si se hubiese realizado un *fine-tuning*³ al modelo, dando como insumo tweets, y que así estuviese en mejores condiciones de analizar las publicaciones. Para pruebas futuras, se podría emplear una unión de tweets, respetando que todos respondan a un mismo contexto, como ya se aplicó en otra investigación⁴, generando así textos de mayor tamaño.

La amplia mayoría de fallos se presentaron como falsos positivos, donde se extraían ubicaciones, pero estas aparecían nombradas en la descripción de una situación ajena. Esto se contendría, al menos parcialmente, con una mejora de la lectura del contexto, existiendo para esto la opción de agregar un modelo de procesamiento de lenguajes naturales del tipo de análisis semántico.