Maestría en Inteligencia Artificial Aplicada / MNA / Tecnológico de Monterrey

Actividad en Equipos - Semanas 10:

Transformer GPT & Web-Scraping

ITESM Campus Guadalajara

Escuela de Ingeniería y Ciencias

Dr. Luis Eduardo Falcón Morales

Nombre(s): _____

Matrícula(s): _____

En la actividad de esta semana trabajarás en equipos mediante el uso del Transformer GPT y comentarios de usuarios obtenidos mediante WebScraping en Tripadvisor.

A través de Tripadvisor (https://www.tripadvisor.es/) en español obtendremos los comentarios de usuarios sobre su experiencia en un hotel en particular. En esta plataforma se evalúa la experiencia mediante puntos de 1 a 5, que corresponden de manera ascendente a “Pésimo”, “Malo”, “Normal”, “Muy bueno” y “Excelente”.

Una de las formas en que se han venido utilizando los generadores de texto como el GPT, es para generar datos sintéticos de texto que no tengan problemas de privacidad. En particular en esta actividad los usaremos para generar datos sintéticos sobre comentarios positivos y negativos de usuarios y posteriormente los usaremos para ver si ayudan a mejor el desempeño de un modelo de clasificación.

De la siguiente liga de Tripadvisor selecciona un hotel en la zona de la ciudad de Monterrey, México, que consideres adecuado para extraer los comentarios de español y evaluaciones de usuarios:

https://www.tripadvisor.es/Hotels-g150782-Monterrey_Northern_Mexico-Hotels.html

En particular trata de seleccionar algún hotel cuyas calificaciones consideres desbalanceadas, al verlo como un problema binario: Positivo (4, 5) y Negativo (1, 2 3).

1. Mediante web scraping realiza una extracción de los comentarios y calificaciones a dicho hotel. Recuerda que Tripadvisor representa las calificaciones mediante círculos verdes. Los comentarios deben ser de usuarios que los redactaron en idioma en español. Indica la cantidad de comentarios y sus calificaciones que extrajiste.

2. Para considerarlo como un problema binario, convierte las calificaciones a valores de 0 (comentario negativo) y 1 (comentario positivo).

    a. En particular indica cómo considerarás las calificaciones entre 3 y 4, como 3.5, por ejemplo. Justifica la decisión tomada.
    b. Indica la proporción de comentarios positivos y negativos que tienes ahora.

3. Selecciona y aplica un modelo Transformer generador de texto en español de la plataforma de HuggingFace para “balancear” los porcentajes de comentarios positivos y negativos.

Indica la cantidad de datos sintéticos que vas a generar, en particular contesta las siguientes preguntas:

   a. ¿Vas a generar solo datos de la clase negativa minoritaria o de ambas clases? ¿Por qué?
   b. ¿Qué porcentaje de nuevo balanceo de clases vas a generar? ¿Por qué? NOTA: Recuerda que no siempre es lo mejor balancearlos al 50% cada clase, porque si el conjunto original está demasiado desbalanceado y los datos sintéticos generados son de muy buena calidad, los resultados obtenidos al final podrían ser bastante desastrosos.

4. Realiza una partición de los datos en entrenamiento, validación y prueba con los porcentajes que consideres adecuados.

5. Mediante un Transformer que consideres adecuado para idioma español de la plataforma HuggingFace, aplica un modelo de clasificación binaria (análisis de sentimiento) para los datos originales, es decir, sin incluir los comentarios sintéticos. ¿Qué desempeño obtienes? NOTA: Justifica el procedimiento utilizado. Es decir, podrías utilizar un modelo Transformer de manera directa (out-of-box), en cuyo caso la evaluación la haces de manera directa sobre el conjunto entrenamiento+validación. O bien, usar un modelo Transformer con fine-tuning en cuyo caso entrenas con el conjunto de entrenamiento y validas con el de prueba.
   En cuanto al Transformer a utilizar, está “PlanTL-GOB-ES/gpt2-large-bne” de la versión de MarIA, pero puedes utilizar algún otro si lo consideras adecuado.

6. Ahora realiza el entrenamiento y evaluación de los conjuntos aumentados (reales más sintéticos). En este caso sí aplica un modelo con fine-tuning. ¿Por qué en este caso no se justificaría utilizar la técnica out-of-box para entrenar y evaluar directamente los datos con los comentarios sintéticos?

7. Compara los resultados e incluye tus comentarios. En particular, indica si consideras que los comentarios sintéticos fueron de ayuda en este caso.