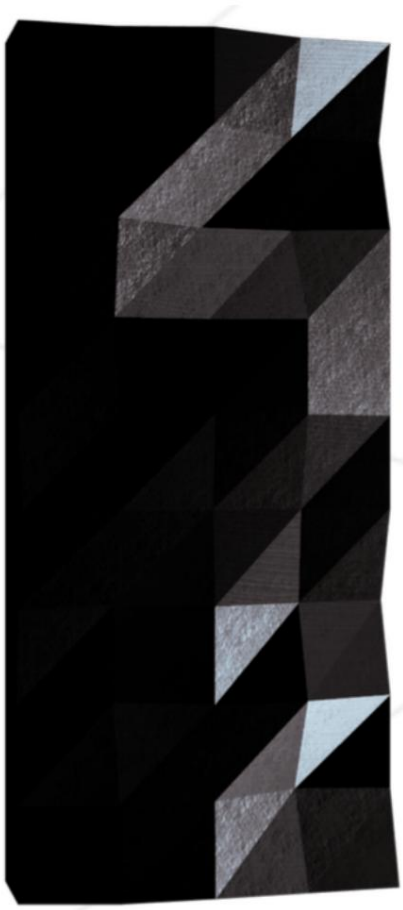
Aprendizaje por transferencia y perfeccionamiento de los LLM

Una primera aproximación: Análisis de Sentimientos

42 Proyecto de IA de Urduliz

Septiembre de 2024



Prefacio

Tras ver la última película de Nolan, me encontré inmerso en un mundo donde la coherencia narrativa no solo está ausente, sino deliberadamente abandonada, donde el significado mismo parece un espejismo que se desvanece a medida que uno intenta comprenderlo. El director, en lo que parece un elaborado ejercicio de simulacro de inteligibilidad, orquesta una secuencia de eventos que, al examinarlos con más atención, se desploman en un vacío de indiferencia semiótica. Los personajes flotan en un espacio liminal de significantes.

Es precisamente aquí, en esta intersección liminal de lo real y lo simbólico, donde el intento de la película de crear una narrativa cohesiva se desintegra en un puro juego de significantes, una especie de goce lacaniano de parálisis interpretativa. Incluso se podría argumentar que el acto mismo de buscar significado se convierte en el tema central de la película, un bucle recursivo de aplazamiento infinito, similar a la estructura de una cinta de Möbius, donde el espectador, al igual que el sujeto de Lacan, se encuentra suspendido entre el deseo y la ausencia de deseo.

De esta manera, Nolan construye un edificio simbólico tan denso que la única interpretación viable es renunciar por completo a cualquier esperanza de interpretación, entregándose a la pura opacidad de la experiencia cinematográfica. Y, sin embargo, en esta rendición, emerge cierta claridad —no del mensaje pretendido por la película, pues dicho mensaje se pierde irremediablemente en el laberinto de significantes—, sino de la propia compulsión del espectador por descifrar un rompecabezas que nunca se pretendió resolver.

Creo que he dejado claro mi punto.

1 Introducción

El análisis de sentimientos es una potente técnica de procesamiento del lenguaje natural que se utiliza para determinar si un texto transmite un sentimiento positivo o negativo. Desempeña un papel fundamental para comprender las opiniones y emociones de las personas a partir de diversos tipos de datos textuales.

Se puede aplicar a reseñas de películas o libros. Una reseña de una película popular como Origen podría decir: «Nolan se ha superado con esta obra maestra, una experiencia alucinante de principio a fin», expresando claramente una perspectiva positiva. Por otro lado, una reseña negativa de la misma película podría decir: «Origen fue demasiado confusa, no pude entenderla».

Otro caso de uso común son las reseñas de restaurantes. Por ejemplo, un comensal podría dejar comentarios como "¡La comida estuvo deliciosa y el servicio fue excelente!" (positivo) o "La comida estaba fría y el servicio fue pésimo" (negativo).

Este proyecto es una introducción al aprendizaje por transferencia y al ajuste fino, un potente enfoque de aprendizaje automático que aprovecha modelos preentrenados para resolver nuevos problemas. En este caso, utilizarás técnicas de ajuste fino en modelos preentrenados para desarrollar un sistema de análisis de sentimientos binario. Investigarás diversos modelos preentrenados, técnicas de tokenización y configuraciones de hiperparámetros para encontrar el equilibrio óptimo entre rendimiento y eficiencia computacional.

Sin embargo, debido a la limitación de recursos, podría ser necesario elegir modelos más pequeños y reducir el tamaño del conjunto de datos para garantizar que el proceso de entrenamiento sea viable dentro de sus limitaciones informáticas. De esta manera, adquirirá experiencia práctica con uno de los avances más impactantes del PLN moderno.

2 Objetivos

Aclaración importante

El objetivo de este proyecto no es profundizar en los detalles altamente técnicos o matemáticos de arquitecturas complejas como transformadores, codificadores, decodificadores o mecanismos de atención.
En lugar de eso, debería centrarse en aplicar estos modelos al análisis de sentimientos mediante técnicas de aprendizaje por transferencia y ajuste fino. Comprender los conceptos generales es suficiente para este proyecto.

El objetivo principal de este proyecto es brindarle experiencia práctica en aprendizaje por transferencia, perfeccionamiento
Ajuste y procesamiento del lenguaje natural. Aprenderás a:

- Utilice computación en la nube de nivel gratuito o recursos locales (como sgoinfre) para capacitar y ajustar modelos de aprendizaje automático.
- Investigar y seleccionar modelos previamente entrenados (por ejemplo, BERT, RoBERTa, GPT-2, DistilBERT, Electra, XL-Net) y métodos de tokenización.
- Ajuste los modelos entrenados previamente ajustando los hiperparámetros, incluido el número de épocas y el aprendizaje. tasa de procesamiento y tamaño del lote para optimizar el rendimiento del modelo.
- Comparar las técnicas de tokenización y cómo se relacionan con el modelo preentrenado.
- Evaluar la precisión del modelo y su capacidad de generalización.

3 Instrucciones generales

Nota

Veremos algunos pasos básicos del análisis de sentimientos. Por supuesto, estas no son las únicas técnicas disponibles ni los únicos pasos a seguir. Cada conjunto de datos y problema debe abordarse de forma única. Seguramente encontrará otras maneras de analizar sus datos en el futuro.

Puede usar cualquier lenguaje de programación, bibliotecas, modelos, métodos de tokenización y recursos informáticos (en la nube o locales) que prefiera. Sin embargo, debe centrarse en perfeccionar los modelos preentrenados, asegurándose de adaptarlos eficazmente a su tarea de análisis de sentimientos. Se aplican las siguientes reglas:

- Debe decidir los hiperparámetros clave: número de épocas, tasa de aprendizaje y tamaño del lote.
- Se espera que realice su propia investigación para seleccionar modelos, tokenizadores y conjuntos de datos adecuados. y plataformas de computación en la nube de nivel gratuito.
- Si bien se pueden utilizar bibliotecas de alto nivel como Transformers de HuggingFace, es importante comprender los conceptos y las decisiones subyacentes.
- Los proyectos se evaluarán en función de su claridad, estructura y comprensión de los conceptos involucrados.

4 Parte obligatoria

Advertencia

Se recomienda encarecidamente realizar los siguientes pasos en el orden prescrito, especialmente Cuando se trata de recursos limitados.

Su proyecto debe incluir las siguientes tareas:

Investigación y selección de modelos y plataformas: Investigue varios modelos preentrenados y compare plataformas en la nube para ajustar el modelo. Tenga en cuenta el tamaño del modelo; dadas las posibles limitaciones de recursos, podría ser necesario elegir modelos más pequeños y computacionalmente eficientes.

- Investigación y selección de tokenización: explore diferentes técnicas de tokenización y seleccione una adecuada según el conjunto de datos y el modelo.

Investigación y selección de conjuntos de datos: Se le proporcionará un conjunto de datos de muestra llamado sam-ple reviews.parquet para que pueda empezar a analizar sus opiniones. Sin embargo, deberá elegir su propio conjunto de datos para el modelo final (p. ej., reseñas de IMDB, Yelp o Amazon) y dividirlo en subconjuntos de entrenamiento y prueba. Tenga en cuenta que, debido a posibles limitaciones de recursos, podría ser necesario reducir el tamaño del conjunto de datos.

- Entrenamiento del modelo: ajuste el modelo previamente entrenado seleccionado en el conjunto de datos elegido, ajustando hiperparámetros como la tasa de aprendizaje, el tamaño del lote y la cantidad de épocas.
- Evaluación del modelo: evalúe el rendimiento de su modelo en el conjunto de pruebas y apunte a una precisión de al menos 0,75, aunque este es un punto de referencia recomendado más que un requisito estricto.

5 Parte adicional

Si has completado con éxito las tareas obligatorias, puedes intentar los siguientes desafíos de bonificación:

Pruebe su modelo con conjuntos de datos adicionales para evaluar su capacidad de generalización a diferentes tipos de datos de sentimiento. Debe seleccionar los otros conjuntos de datos y evaluar su modelo con estos nuevos conjuntos de datos.

- Explorar múltiples técnicas de tokenización y comparar sus resultados, brindando información sobre sus diferencias.
- Implemente su modelo entrenado en la web, lo que permite la predicción de sentimientos en tiempo real a través de una interfaz de usuario. Puedes ver un ejemplo en esta página web.

6 Presentación y evaluación por pares

Esta es una prueba de concepto, por lo que no se enviará ningún repositorio de Git. Debe tener una carpeta específica con su código. No incluya el modelo entrenado ni el conjunto de datos en la carpeta.

- Si ha utilizado recursos de computación en la nube para entrenar su modelo, debe mostrar su implementación. ción en línea durante la evaluación.
- Si ha entrenado su modelo localmente, debe ejecutarlo localmente durante la evaluación.
- El proceso de capacitación no debe durar más de 15 minutos, dejando tiempo para la discusión y Explicación de su implementación.

Asegúrese de que su presentación sea clara, bien organizada y funcional, ya que esto afectará su evaluación.