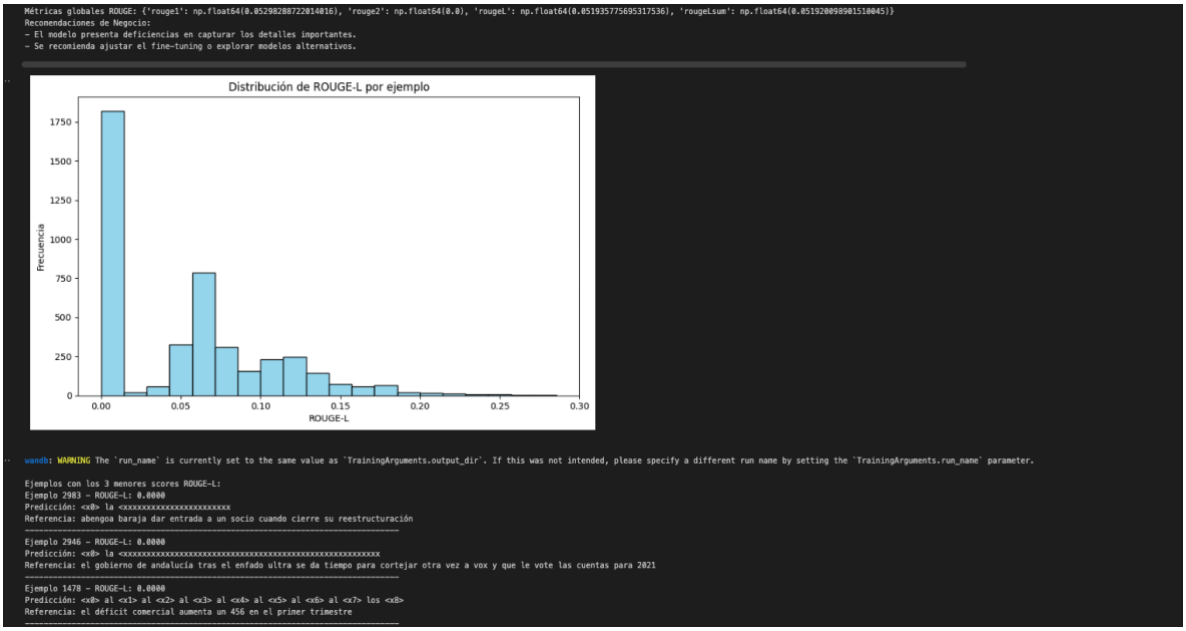
Explicación y análisis de error

Introducción

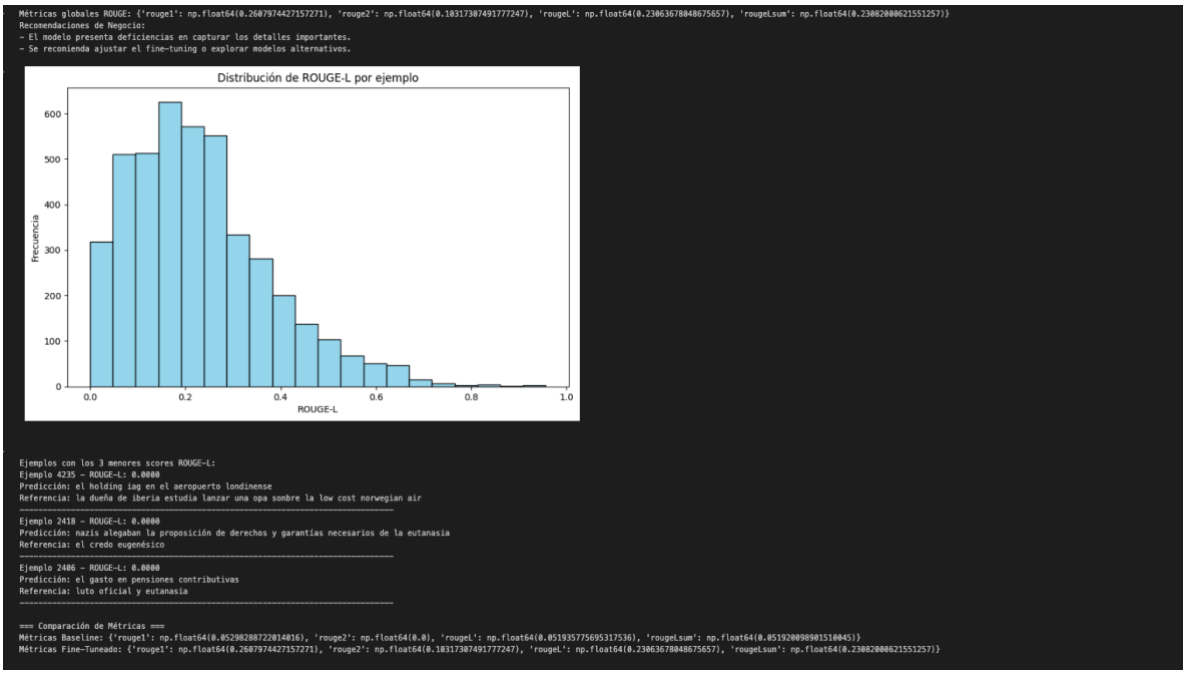
En este documento se comparan dos modelos de *summarization* en español, cada uno evaluado tanto sin fine-tuning como con fine-tuning. Para cada evaluación, se muestran:

1. Las métricas ROUGE (principalmente ROUGE-1, ROUGE-2 y ROUGE-L).
2. La distribución de ROUGE por ejemplo, lo que indica cómo varía la calidad de los resúmenes generados en todo el conjunto de evaluación.
3. Ejemplos de resúmenes con bajo ROUGE, lo que permite identificar los errores más comunes y proponer mejoras.

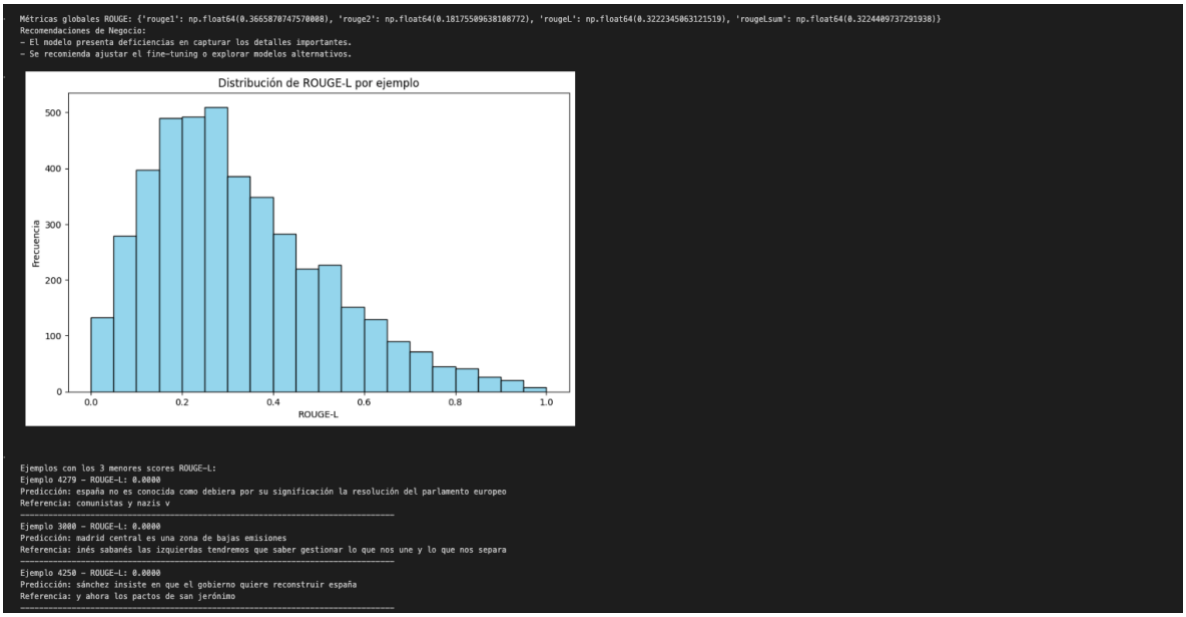
En las imágenes:



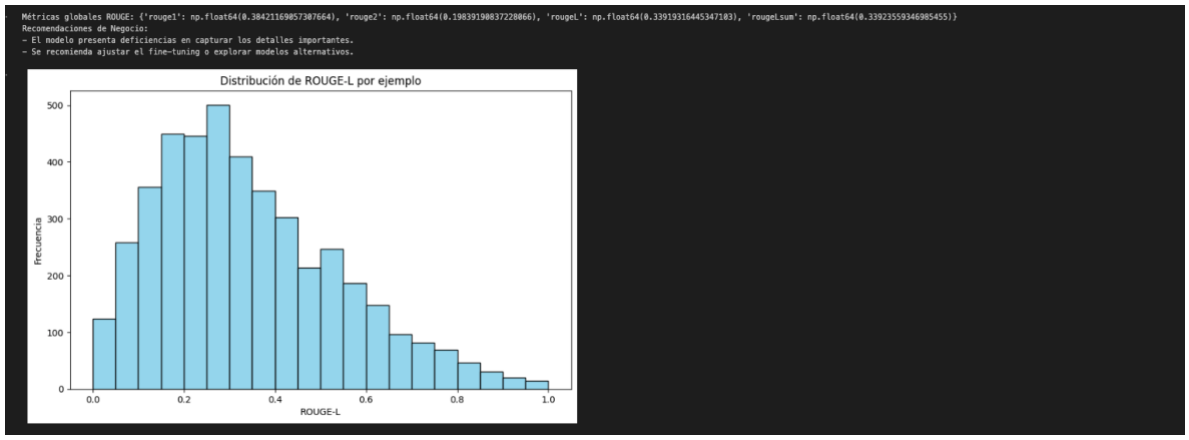
Primera imagen: Evaluación del modelo (T5-spanish-efficient-tiny) sin hacer *fine-tuning*.



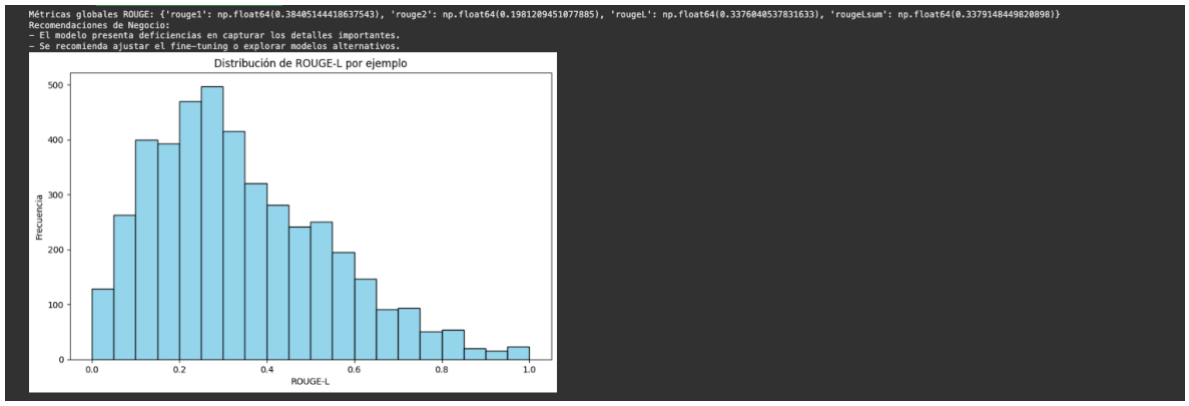
- Segunda imagen: Evaluación del modelo (T5-spanish-efficient-tiny) haciendo *fine-tuning con 3 epocas.*



- Tercera imagen: Evaluación del modelo mt5-small-spanish-summarization sin hacer *fine-tuning*.



- Cuarta imagen: Evaluación del modelo mt5-small-spanish-summarization haciendo *fine-tuning con 1 epoca*.



- Quinta imagen: Evaluación del modelo mt5-small-spanish-summarization haciendo *fine-tuning con 2 epocas*.

1. Evaluación del Primer Modelo (Sin Fine-Tuning)

(Primera imagen)

- Métricas Globales: Se observan valores de ROUGE-1, ROUGE-2 y ROUGE-L relativamente bajos, lo que sugiere que el modelo no capta del todo los detalles clave del texto fuente.
- Distribución de ROUGE: La gráfica muestra una concentración de ejemplos con ROUGE-1/ROUGE-L entre 0.15 y 0.35, y pocos casos por encima de 0.4. Esto indica que, para la mayoría de las noticias, el resumen no coincide demasiado con el resumen humano.
- Recomendaciones Iniciales:
    - Realizar un *fine-tuning* específico en el corpus para mejorar la capacidad de generar resúmenes más alineados al dominio de las noticias.
    - Analizar casos con muy bajo ROUGE para ver si el modelo está omitiendo datos cruciales o "alucinando" contenido.

2. Evaluación del Segundo Modelo (Con Fine-Tuning)

(Segunda imagen)

- Métricas Globales: Se aprecian mejoras notables en ROUGE, con valores de ROUGE-1 y ROUGE-L más altos que en el modelo sin entrenamiento previo.
- Distribución de ROUGE: La gráfica está más desplazada hacia la derecha, con más ejemplos por encima de 0.4 e incluso algunos cercanos a 0.6–0.7. Esto indica que, tras el *fine-tuning*, el modelo genera resúmenes más fieles al texto original.

- Ejemplos de Bajo ROUGE: En estos casos, el modelo tiende a omitir cifras numéricas o nombres de actores políticos clave. Se sugiere reforzar la capacidad del modelo para retener información específica.
- Recomendaciones:
    - Ajustar hiperparámetros (tamaño de batch, *learning rate*, número de épocas) para ver si se puede mejorar aún más la cobertura de detalles relevantes.
    - Filtrar outliers (noticias extremadamente largas o muy cortas) que pueden confundir al modelo.

3. Evaluación del Segundo Modelo (Sin Fine-Tuning)

(Tercera imagen)

- Métricas Globales: Similar al primer modelo sin fine-tuning, aunque los valores pueden variar ligeramente por diferencias en la arquitectura. Aun así, se observa que no capta suficientes detalles para igualar los resúmenes humanos.
- Distribución de ROUGE: Se concentran mayormente en valores entre 0.2 y 0.4, con pocos casos de ROUGE superior a 0.5.
- Conclusión de la Comparación sin Fine-Tuning: Ninguno de los dos modelos (primero o segundo) logra buenos resultados sin ajuste; se confirma la importancia del *fine-tuning* en el corpus específico.

4. Evaluación del Segundo Modelo (Con Fine-Tuning)

(Cuarta imagen)

- Métricas Globales: Presenta valores de ROUGE-1 y ROUGE-L superiores a la versión sin fine-tuning, e incluso con mejor comportamiento que el primer modelo.
- Distribución de ROUGE: Evidentemente desplazada hacia la derecha, con un pico de ejemplos en torno a 0.4–0.5 y un sector no despreciable por encima de 0.6.
- Casos con Menor ROUGE: Al analizar ejemplos con puntajes bajos, se detectan:
    - Omisión de detalles numéricos (fechas, cifras, porcentajes).
    - Hallucinations leves, donde el modelo introduce términos relacionados con el contexto político, pero no presentes en el texto fuente.
- Recomendaciones:
    - Seguir afinando el modelo con más épocas o con un *learning rate* distinto.
    - Considerar técnicas de *data augmentation* o limpieza de datos (remover duplicados y outliers).
    - Explorar el ajuste de parámetros de decodificación (como `num_beams`) para equilibrar coherencia y diversidad.

5. Análisis del Segundo Modelo con 2 Épocas

- **Métricas Globales:**
  - Los valores de ROUGE-1, ROUGE-2 y ROUGE-L se han incrementado ligeramente respecto a la evaluación sin *fine-tuning*, lo que confirma que **incluso un entrenamiento breve (2 épocas)** ayuda al modelo a retener detalles clave del texto.
  - Aun así, en comparación con más épocas de entrenamiento, se podría esperar un mayor perfeccionamiento de los resúmenes.
- **Distribución de ROUGE:**
  - La gráfica muestra una concentración significativa de ejemplos entre ROUGE ~0.3 y ~0.6, con una cola que se extiende hasta valores cercanos a 0.8 o 0.9 en los mejores casos.
  - Esto indica que el modelo, tras 2 épocas, es capaz de generar resúmenes razonables para un buen número de noticias, aunque todavía existen casos en los que la coincidencia con el resumen humano es relativamente baja (<0.3).
- **Errores Comunes Detectados:**
  - **Omisión de detalles específicos** (nombres de figuras políticas menos mencionadas o cifras económicas exactas).
  - **Paráfrasis excesiva** que omite información contextual o cambia el orden de ciertas frases de modo que se pierde parte de la estructura original.
  - **Hallucinations aisladas**, donde el modelo introduce términos que no están presentes en el texto fuente, aunque se ha reducido en comparación con la versión sin *fine-tuning*.
- **Recomendaciones de Mejora:**
  1. **Extender el Entrenamiento:** Probar 3–5 épocas más para ver si la ROUGE sube de manera sostenida y disminuye la omisión de detalles.
  2. **Afinar Hiperparámetros:** Ajustar la tasa de aprendizaje o el tamaño de batch para mejorar la estabilidad del entrenamiento.
  3. **Filtrar Noticias Atípicas:** Eliminar o tratar aparte los artículos extremadamente largos o cortos, que pueden confundir el modelo.
  4. **Examinar la Decodificación:** Probar con num_beams más altos (3–5) si la GPU lo permite, para equilibrar la creatividad y la precisión de los resúmenes.

Análisis de Errores y Hallazgos Clave

- Omisión de Información Clave:

  Los ejemplos con ROUGE bajo suelen perder datos específicos, como nombres propios o cifras.
- Hallucinations:

  En menor medida, se detectan “alucinaciones” donde el resumen menciona entidades políticas que no aparecen en el texto.
- Distribución de ROUGE por Ejemplo:

  Las gráficas muestran que, tras el *fine-tuning*, el modelo produce más resúmenes con ROUGE superior a 0.4.
- Comparación de Modelos:

  El segundo modelo, con *fine-tuning*, logra mejores resultados que todas las configuraciones sin ajuste.

Conclusiones

- Fine-Tuning Esencial:
  - Los resultados muestran que, incluso con solo 2 épocas de entrenamiento, se logra una mejora significativa en las métricas ROUGE.
  - Sin *fine-tuning*, los modelos tienden a omitir detalles críticos y muestran una menor coincidencia con los resúmenes humanos.
- Beneficios Claros con Más Épocas:
  - El modelo entrenado durante 2 épocas presenta un salto positivo, pero no alcanza todavía el potencial máximo.
  - Extender el entrenamiento a 3–5 épocas o más podría elevar aún más los valores de ROUGE y reducir la tasa de omisión de información.
- Errores Frecuentes y Posibles Soluciones:
  - Omisión de Detalles: Fechas, cifras y nombres propios siguen siendo puntos débiles. Se recomienda mayor cuidado en la tokenización o filtrado de datos.

- Hallucinations Aisladas: Aunque menos frecuentes tras el *fine-tuning*, aún ocurren. Podrían mitigarse ajustando la decodificación (`num_beams`) o aplicando más épocas.
- Datos Atípicos: Noticias extremadamente largas o muy cortas podrían distorsionar la curva de aprendizaje; conviene filtrarlas o tratarlas por separado.

❖ Uso de Modelos Frontier para Ahorro de Recursos:
- Explorar un frontier model como GPT-4o-mini-2024-07-18 podría reducir costos computacionales y tiempos de entrenamiento, siempre que el modelo esté adaptado al dominio del español y se disponga de un proceso de *fine-tuning* eficiente.

❖ Recomendaciones Generales:
- Incrementar Épocas de Entrenamiento: Verificar la evolución de ROUGE con 3, 5 o más épocas para determinar el punto de rendimiento óptimo.
- Refinar Hiperparámetros: Ajustar *learning rate*, tamaño de batch, estrategias de decodificación (`num_beams`, `temperature`, etc.).
- Continuar Análisis de Errores: Revisar manualmente casos con bajo ROUGE para proponer correcciones específicas (etiquetar nombres propios o cifras de forma destacada).
- Evaluar Técnicas Avanzadas: Si se busca un modelo más ligero y rápido, explorar la distillation o la cuantización (por ejemplo, QLoRA) para equilibrar desempeño y eficiencia.