

Universidad de Las Palmas de Gran Canaria

Máster en Sistemas Inteligentes y Aplicaciones Numéricas en Ingeniería

Asignatura: Ciencia de Datos en Ingeniería

Sistema de recomendación de revistas científicas mediante técnica clásica y conexionista

Autora

Paula Rosa Rodríguez Morales

Curso académico 2025–2026

Índice

1. Introducción	2
2. Planteamiento del problema	2
3. Conjunto de datos	2
4. Metodología	3
4.1. Preprocesamiento del texto	3
4.2. Modelo clásico: TF-IDF y SVM	4
4.3. Modelo basado en Transformers	4
4.4. Evaluación de los modelos	4
5. Resultados y análisis	5
5.1. Resultados del modelo clásico (TF-IDF + SVM)	5
5.2. Resultados del modelo Transformer (DistilBERT)	6
5.3. Comparación global entre modelos	8
6. Conclusiones	8
6.1. Conclusiones sobre el modelo clásico (TF-IDF + SVM)	8
6.2. Conclusiones sobre el modelo Transformer (DistilBERT)	9
6.3. Discusión general	10

1. Introducción

En la actualidad, el número de revistas científicas es muy elevado, lo que ha provocado que dentro de un mismo campo de investigación existan muchas publicaciones diferentes. Como consecuencia, uno de los problemas habituales a los que se enfrentan los investigadores es decidir en qué revista resulta más adecuado enviar un artículo para su publicación. Esta decisión suele basarse en la experiencia personal del investigador o en búsquedas manuales de artículos similares, lo que puede resultar un proceso lento.

En este contexto, el objetivo de este trabajo es desarrollar un sistema inteligente que recomiende una revista científica a partir del contenido de un artículo. Para ello, el sistema utiliza información textual como el título, el *abstract* y las palabras clave, y aprende a realizar la recomendación a partir de artículos publicados previamente en distintas revistas.

2. Planteamiento del problema

El problema se plantea como un problema de clasificación de textos, en el que cada artículo se representa mediante su contenido textual, y la etiqueta asociada corresponde a la revista en la que fue publicado. A partir de estos datos, se entrena modelos capaces de predecir la revista más adecuada para un nuevo artículo.

Para desarrollar este trabajo, en primer lugar se utiliza una aproximación clásica basada en la representación de los textos mediante TF-IDF y un clasificador lineal, que sirve como modelo base. Posteriormente, se implementa una aproximación más avanzada utilizando un modelo Transformer preentrenado, con el objetivo de analizar si este tipo de modelos mejora los resultados obtenidos.

Finalmente, se comparan los resultados obtenidos por ambos enfoques mediante distintas métricas de evaluación, analizando las diferencias de rendimiento entre el modelo clásico y el modelo basado en Transformers, así como sus ventajas y limitaciones en este contexto.

3. Conjunto de datos

El conjunto de datos utilizado en este trabajo está formado por artículos científicos publicados en distintas revistas de la editorial Elsevier. Los datos se han obtenido a partir de ficheros en formato JSON, organizados por revistas, y posteriormente se han procesado para construir un único conjunto de datos estructurado.

Cada artículo contiene información textual relevante para la tarea de recomendación, en concreto:

- El título del artículo.
- El *abstract*.
- Las palabras clave asociadas al artículo.

A partir de esta información, se ha generado un campo de texto unificado que combina el título, el resumen y las palabras clave, y que será utilizado como entrada de los modelos de clasificación.

Las revistas consideradas en el estudio son las siguientes:

- *Expert Systems with Applications*
- *Journal of Visual Communication and Image Representation*
- *Neural Networks*
- *Robotics and Autonomous Systems*

Tras el proceso de limpieza y filtrado de los datos, el conjunto final contiene un total de 13 904 artículos, con una distribución no equilibrada, ya que existen más artículos pertenecientes a algunas revistas frente a otras. Por este motivo, durante el entrenamiento de los modelos se han tenido en cuenta técnicas para mitigar el desbalance de clases.

Para la evaluación de los modelos, el conjunto de datos se ha dividido en dos subconjuntos:

- Un conjunto de entrenamiento, que representa el 80 % de los datos.
- Un conjunto de prueba, que representa el 20 % restante.

La división se ha realizado de forma estratificada, garantizando que la proporción de artículos de cada revista se mantenga tanto en el conjunto de entrenamiento como en el de prueba, así se asegura una evaluación más justa y representativa del rendimiento de los modelos.

4. Metodología

La metodología seguida en este trabajo se estructura en varias etapas, que abarcan desde el preprocesamiento de los datos hasta el entrenamiento y la evaluación de los modelos.

4.1. Preprocesamiento del texto

Antes de entrenar los modelos, fue necesario realizar un preprocesamiento del contenido textual de los artículos. Para cada artículo se construyó un único texto concatenando el título, el resumen y las palabras clave, ya que esta información conjunta aporta un mayor contexto sobre la temática del trabajo.

En la aproximación clásica, el preprocesamiento incluye:

- Conversión del texto a minúsculas.
- Eliminación de palabras vacías (*stopwords*) en inglés.
- Representación del texto mediante n-gramas de una y dos palabras.

En el caso del modelo Transformer, el preprocesamiento es realizado de forma automática por el tokenizador del modelo, que se encarga de dividir el texto en subpalabras, truncar textos largos y aplicar relleno (*padding*) hasta una longitud fija.

4.2. Modelo clásico: TF-IDF y SVM

Como primera aproximación, y como ya se ha comentado anteriormente, se implementó un modelo clásico de clasificación de textos. En este enfoque, los documentos se representan mediante la técnica TF-IDF (*Term Frequency–Inverse Document Frequency*), que permite convertir el texto en una matriz numérica término-documento.

A partir de esta representación, se ha entrenado un clasificador lineal basado en Máquinas de Vectores de Soporte (SVM), que es habitual en tareas de clasificación de texto debido a su buen rendimiento y bajo coste computacional.

Dado que el conjunto de datos presenta un cierto desbalance entre las clases, se han utilizado pesos de clase ajustados automáticamente para compensar este efecto durante el entrenamiento del modelo.

Este modelo sirve como línea base frente a la cual comparar los resultados obtenidos con el enfoque más avanzado.

4.3. Modelo basado en Transformers

Como segunda aproximación se ha utilizado un modelo Transformer preentrenado, concretamente *DistilBERT*, que es una versión más ligera y eficiente del modelo BERT original.

Se ha escogido este modelo Transformer y no otros más grandes principalmente por razones prácticas, ya que DistilBERT mantiene un rendimiento cercano al de BERT, pero con un menor número de parámetros, lo que reduce significativamente el tiempo de entrenamiento y el consumo de memoria.

El modelo se ha ajustado (*fine-tuning*) para esta tarea concreta añadiendo una capa de clasificación sobre el modelo preentrenado. Durante el entrenamiento, el modelo aprende a asociar el contenido textual del artículo con la revista correspondiente.

El entrenamiento se ha realizado durante un número limitado de épocas, utilizando una división estratificada del conjunto de datos y aprovechando aceleración por GPU cuando ha sido posible.

4.4. Evaluación de los modelos

Para evaluar el rendimiento de los modelos se han utilizado distintas métricas de clasificación:

- Precisión (*precision*)
- Exhaustividad (*recall*)
- Medida F1 (*f1-score*)
- Exactitud global (*accuracy*)

Además, se han calculado matrices de confusión para analizar de forma más detallada los errores cometidos por cada modelo y estudiar qué revistas tienden a confundirse entre sí.

5. Resultados y análisis

Antes de mostrar los resultados es importante mencionar que la evaluación se ha realizado sobre el conjunto de test.

5.1. Resultados del modelo clásico (TF-IDF + SVM)

En la Figura 1 se muestra la matriz de confusión normalizada correspondiente al modelo clásico. Cada fila representa la revista real del artículo, mientras que cada columna indica la revista predicha por el modelo.

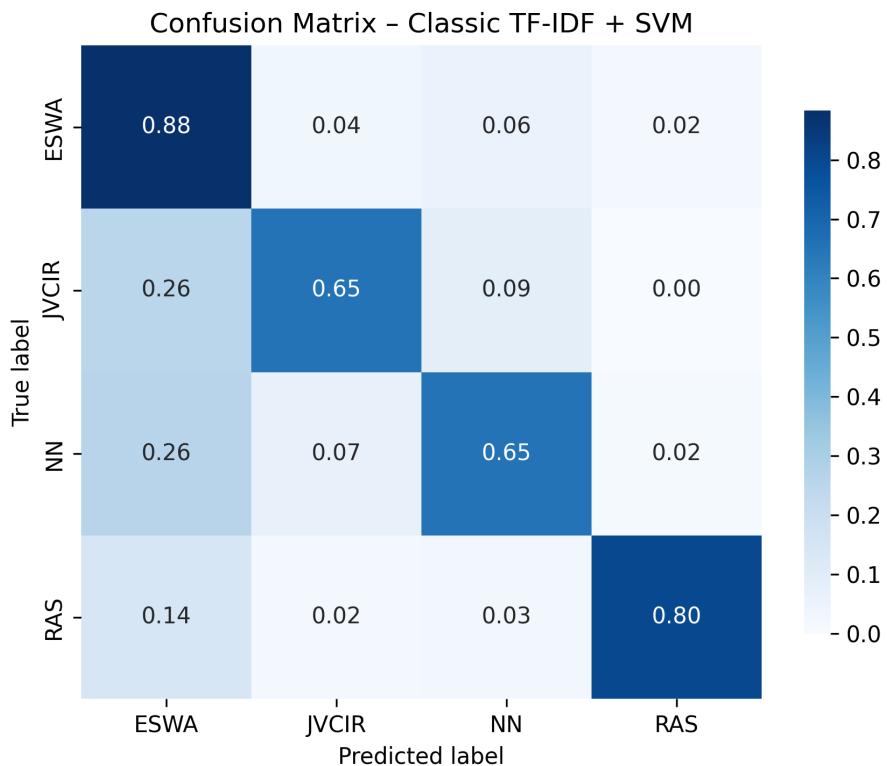


Figura 1: Matriz de confusión normalizada del modelo clásico TF-IDF + SVM.

A partir de esta matriz se observa que la revista *Expert Systems with Applications* (*ESWA*) es la mejor clasificada, con un porcentaje elevado de aciertos. Sin embargo, se aprecia cierta confusión entre las revistas *Journal of Visual Communication and Image Representation* (*JVCIR*) y *Neural Networks* (*NN*), lo que indica que el contenido textual de estas revistas presenta similitudes que el modelo clásico no siempre logra diferenciar correctamente.

La Figura 2 muestra el F1-score obtenido para cada revista en el modelo clásico, lo que permite analizar el rendimiento de forma individual por clase.

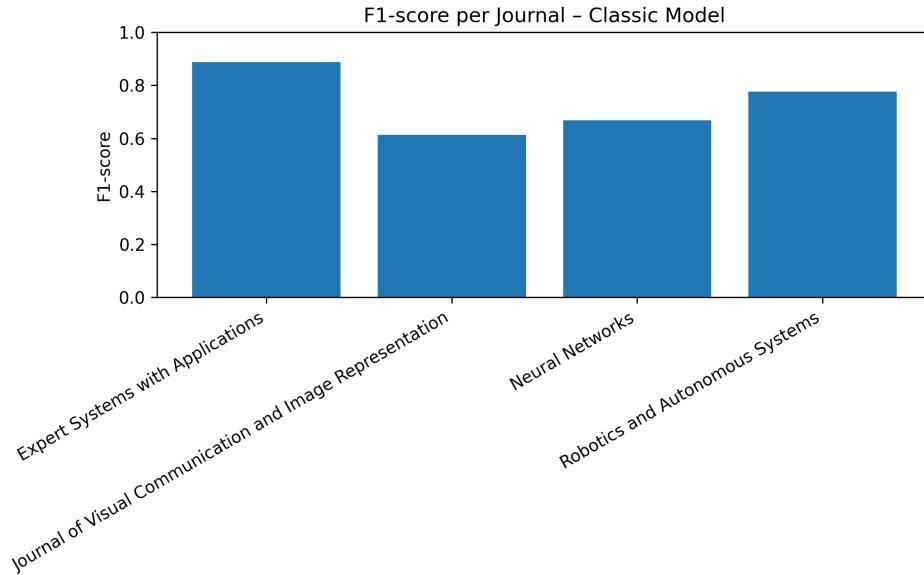


Figura 2: F1-score por revista para el modelo clásico TF-IDF + SVM.

Se observa que el F1-score es más alto para ESWA y *Robotics and Autonomous Systems* (*RAS*), mientras que JVCIR y NN presentan valores inferiores. Esto refleja el efecto del desbalanceo del conjunto de datos y la dificultad del modelo clásico para capturar relaciones semánticas más complejas en textos similares.

5.2. Resultados del modelo Transformer (DistilBERT)

La Figura 3 muestra la matriz de confusión normalizada correspondiente al modelo basado en Transformers. En comparación con el modelo clásico, se aprecia una mejora general en la diagonal principal, lo que indica un mayor número de predicciones correctas.

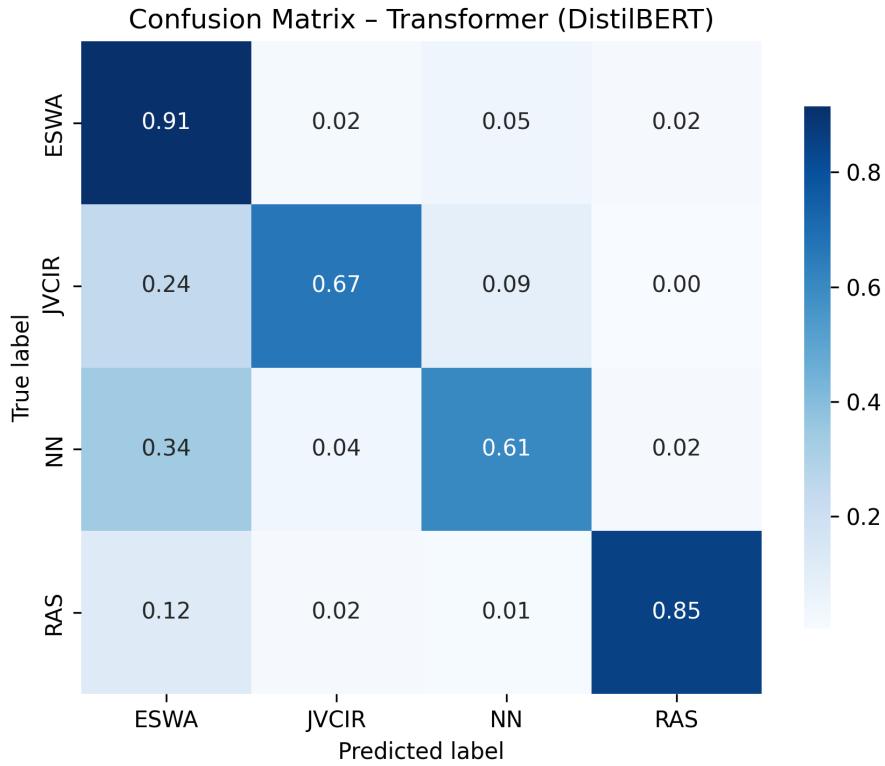


Figura 3: Matriz de confusión normalizada del modelo Transformer (DistilBERT).

El modelo Transformer presenta una mejora notable en la clasificación de ESWA y RAS, y reduce parcialmente la confusión entre JVCIR y NN. No obstante, sigue existiendo cierto solapamiento entre estas dos revistas, lo que sugiere que incluso modelos más avanzados pueden encontrar dificultades cuando los temas tratados son cercanos.

En la Figura 4 se muestran los F1-score por revista obtenidos por el modelo Transformer.

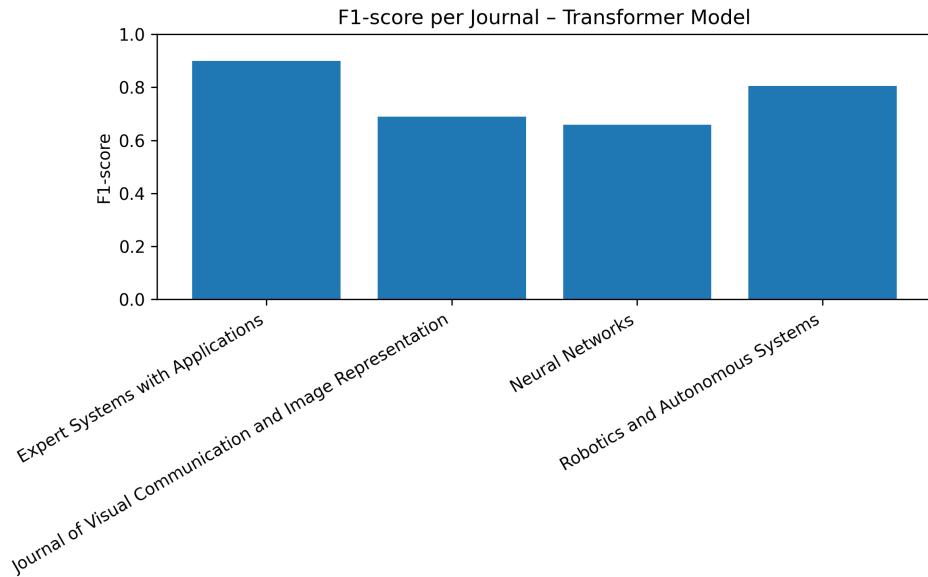


Figura 4: F1-score por revista para el modelo Transformer (DistilBERT).

En este caso, el F1-score mejora de forma generalizada respecto al modelo clásico, especialmente en JVCIR y RAS. Esto indica que el modelo Transformer es capaz de capturar mejor el contexto semántico de los textos, gracias al preentrenamiento sobre grandes corpus de lenguaje natural.

5.3. Comparación global entre modelos

Para facilitar la comparación entre ambos enfoques, en la Figura 5 se muestran las métricas globales de *accuracy* y *macro F1* para el modelo clásico y el modelo Transformer.

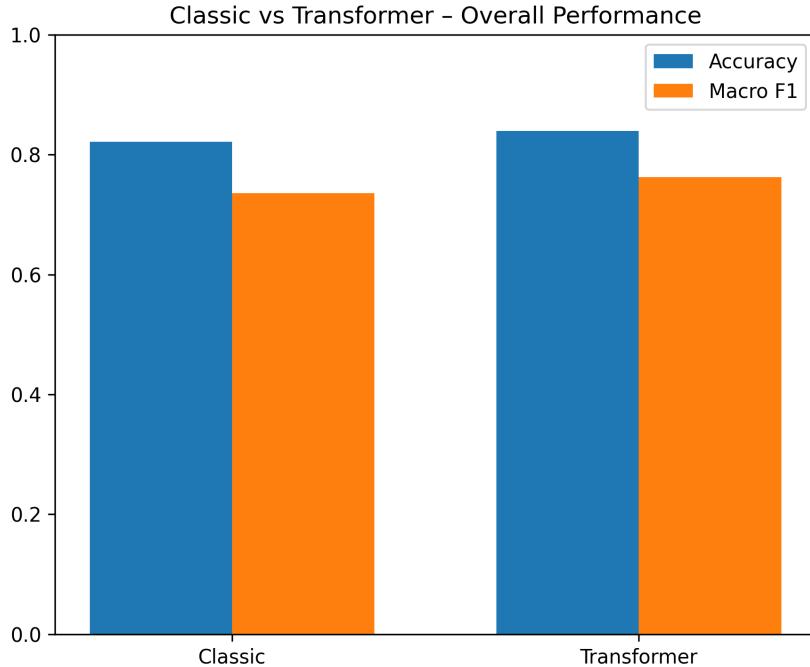


Figura 5: Comparación global de rendimiento entre el modelo clásico y el modelo Transformer.

Los resultados muestran que el modelo Transformer obtiene mejores valores tanto de precisión global como de F1 macro. No obstante, la mejora no es extremadamente grande, lo que resulta razonable teniendo en cuenta el tamaño del conjunto de datos y el número limitado de épocas de entrenamiento.

Se puede razonar que el modelo basado en Transformers ofrece un rendimiento superior y más equilibrado entre clases, aunque el modelo clásico sigue siendo una alternativa válida debido a su menor complejidad computacional y su facilidad de implementación.

6. Conclusiones

6.1. Conclusiones sobre el modelo clásico (TF-IDF + SVM)

El enfoque clásico ha obtenido un rendimiento global bueno, con una **accuracy** de **0.8213** y un **macro F1** de **0.7359**. Esto indica que, aun tratándose de un modelo relativamente simple, es capaz de capturar información relevante del texto para diferenciar entre las distintas revistas científicas.

Analizando los resultados por clase, se observa que la revista *Expert Systems with Applications* presenta un rendimiento especialmente alto ($F1 \approx 0.8879$), lo que es coherente, ya que se trata de la revista con mayor número de artículos en el conjunto de datos y con una temática amplia, lo que permite al modelo aprender patrones léxicos bien definidos.

Por el contrario, las revistas con menor número de muestras, como *Journal of Visual Communication and Image Representation*, muestran un rendimiento inferior ($F1 \approx 0.6124$). Este comportamiento es esperable incluso habiendo aplicado técnicas de balanceo de clases, ya que dicho balanceo compensa el peso de cada clase durante el entrenamiento, pero no aumenta la diversidad real de ejemplos disponibles. En consecuencia, el modelo dispone de menos variabilidad temática y léxica para aprender representaciones discriminativas en estas clases minoritarias.

La matriz de confusión del modelo clásico revela además que una parte significativa de los artículos pertenecientes a clases minoritarias tienden a ser clasificados como *Expert Systems with Applications*. Esto sugiere que esta revista actúa como una clase “generalista” dentro del conjunto de datos, probablemente debido a su amplio alcance temático y al solapamiento de vocabulario con otras revistas. Además, al basarse en una representación TF-IDF, el modelo tiende a apoyarse en términos frecuentes y compartidos, lo que favorece la clase dominante cuando existen fronteras difusas entre categorías.

6.2. Conclusiones sobre el modelo Transformer (DistilBERT)

El modelo basado en Transformers ha conseguido mejorar el rendimiento global con respecto al enfoque clásico. En concreto, se obtiene una **accuracy de 0.8393** y un **macro F1 de 0.7628**. Aunque la mejora no es muy grande, sí es consistente y relevante, especialmente si se tiene en cuenta que el entrenamiento se ha realizado con un número reducido de épocas y utilizando una configuración relativamente estándar del modelo.

Este comportamiento es coherente con las características de DistilBERT, que no representa el texto como una simple bolsa de palabras, sino que genera representaciones contextualizadas. De este modo, el modelo es capaz de captar mejor el significado de las palabras en función de su contexto y las relaciones entre términos, lo que resulta especialmente útil cuando distintas revistas comparten vocabulario similar pero lo emplean en contextos diferentes.

Se ha podido ver que el impacto del modelo Transformer no es igual en todas las clases. Hubo una mejora clara en revistas que presentaban un rendimiento más bajo en el modelo clásico, como *Journal of Visual Communication and Image Representation*, cuyo F1 pasa de aproximadamente 0.6124 a 0.6887. También se aprecia una mejora en *Robotics and Autonomous Systems*, que alcanza un F1 cercano a 0.8056. Estos resultados sugieren que el modelo Transformer es capaz de capturar mejor las particularidades semánticas de clases con menor número de ejemplos o con vocabulario más específico.

Sin embargo, no todas las clases se benefician de la misma forma. En el caso de *Neural Networks*, el rendimiento no mejora claramente respecto al modelo clásico ($F1 \approx 0.6674$ frente a ≈ 0.6582). Esto puede explicarse por el fuerte solapamiento temático de esta revista con otras del conjunto de datos, como artículos sobre aprendizaje automático aplicado a robótica o a sistemas expertos. Cuando las fronteras entre clases son difusas y el conjunto de datos no contiene suficientes ejemplos claramente diferenciados, incluso modelos avanzados pueden presentar dificultades para separar correctamente las categorías.

La matriz de confusión del modelo Transformer muestra que, aunque el patrón general de errores es similar al del modelo clásico, la diagonal principal es más marcada, indicando

una mayor proporción de aciertos por clase. Este comportamiento explica el aumento del macro F1 y sugiere que el modelo logra un reparto más equilibrado del rendimiento entre las distintas revistas, aunque sin eliminar completamente los efectos del solapamiento temático y del tamaño limitado de algunas clases.

6.3. Discusión general

En conjunto, los resultados muestran que:

- El enfoque clásico TF-IDF + SVM ofrece un rendimiento sólido y es una buena línea base, con la ventaja de ser rápido y fácil de entrenar.
- El modelo Transformer obtiene mejores métricas globales y mejora especialmente en algunas clases más difíciles, lo cual sugiere que el contexto aporta información adicional útil.
- El desbalanceo de clases afecta claramente a ambos modelos: las clases con más ejemplos tienden a obtener mejores métricas, mientras que las minoritarias sufren más confusiones.

Un punto importante es que la diferencia entre ambos modelos no es “enorme”. Esto no significa que el Transformer no sea bueno, sino que hay varios factores que limitan su ventaja potencial. De este modo, vemos que pueden existir tareas para las que modelos avanzados pueden tener límites de rendimiento, debido a que los datos no permiten separar etiquetas claramente.