Predicción de notas ORES para articulos de wikipedia en inglés y euskera

Iñigo Berasategi Urdanpilleta

1 Introducción

Todos sabemos que Wikipedia es una de las plataformas más influyentes para el acceso al conocimiento en todo el mundo, con millones de artículos disponibles en una amplia variedad de idiomas. No solo es importante por su capacidad de democratizar el conocimiento (es accesible desde cualquier lugar siempre que haya acceso a internet, salvo excepciones), si no que su naturaleza le permite a cualquier usuario crear o editar contenido. Esto sin embargo, también tiene sus contras. El hecho de que todos podamos editarla plantea desafíos significativos a la hora de verificar la fiabilidad y calidad de los articulos. Reconocer automáticamente cuáles son los artículos bien estructurados, basados en fuentes verificables y que cumplen estándares editoriales, frente a aquellos incompletos, con errores o con información de dudosa calidad supondría sin duda una mejora en la confianza hacia la la plataforma.

En este contexto, ORES (Objective Revision Evaluation Service) surge como una herramienta clave. Este sistema utiliza modelos de inteligencia artificial para analizar y evaluar automáticamente la calidad de los artículos en Wikipedia. A través de estos modelos, ORES asigna puntuaciones a los artículos basándose en diversas métricas relacionadas con su contenido, estilo, estructura y confiabilidad.

Sin embargo, la eficacia de ORES flojea en algunos idiomas. En idiomas más minoritarios como el euskera el problema no radica tanto en la iformación falsa como en la falta de información. Los modelos que se van a estudiar en este análisis no son suficientemente complejos como para detectar información falsa, pero si lo suficiente como para poder acertar con relativa exactitud la calidad de los articulos tanto en inglés como en euskera (o eso es lo que se busca).

El euskera, al ser un idioma con mucho menos recursos digitales y corpus disponibles, plantea desafíos extra a la hora de realizar buenos etiquetados. ¿son los modelos multilingües capaces de adaptarse con igual precisión a idiomas menos representados? ¿Se beneficia la evaluación de la calidad al traducir artículos a un idioma con más recursos, como el inglés, o se pierden matices importantes en la traducción? Estudiaremos algunos de estos temas a lo largo del documento.

El proyecto se ha dividio en 3 fases como estandar para la asignatura de PLN:

Z1: Entrenar, evaluar y fine-tunear los modelos en inglés (con DistilBert) y en inglés/euskera (con mBert). Elegir los modelos a utilizar, elegir el dataset con el que trabajar, recabar datos de articulos de wikipedia en euskera y sus puntuaciones, y por último pre-procesar toda esta información para poder alimentar al modelo. Esta parte ha sido sin duda la que más tiempo ha llevado.

Z2: Traducir los articulos del dataset del euskera al inglés y evaluar el modelo monolongüe

Z3: Evaluar el modelo multilingüe con los articulos directamente en Euskera. Comparar los resultados entre Z3 y Z2

2 Trabajos relacionados

Ya hemos comentado que en el ámbito de la evaluación de la calidad de artículos en Wikipedia, el desarrollo de ORES es uno de los trabajos más relevantes. ORES utiliza modelos de aprendizaje automático e IA para clasificar y evaluar la calidad de los artículos basándose en características como la longitud del texto, la cantidad de referencias y la diversidad de palabras clave. Algunos de los trabajos relacionados más relevantes a comentar son:

Wulczyn et al. (2016) exploraron las dinámicas de calidad en Wikipedia desde una perspectiva

multilingüe. Su estudio reveló que los artículos en idiomas con menos recursos, como aquellos con menor número de editores o datos disponibles, tienden a ser de calidad inferior en comparación con idiomas dominantes como el inglés. Este análisis subraya las desigualdades en la calidad del contenido según la disponibilidad de recursos lingüísticos.

Tiedemann (2012) investigó el impacto de la traducción automática en la preservación de la calidad del texto. Sus hallazgos destacaron cómo las traducciones pueden alterar tanto el significado como las características lingüísticas de los textos, afectando la precisión de cualquier evaluación basada en modelos.

Estos trabajos constituyen una base sólida para comprender cómo se evalúa la calidad en Wikipedia y cómo se enfrentan los retos derivados del multilingüismo y la escasez de recursos en ciertos idiomas.

3 Sistema

El sistema utilizado para este proyecto se ha diseñado con el objetivo de entrenar y evaluar varios modelos y comparar sus rendimientos entre sí. Inicialmente, se utilizó Google Colab como entorno principal para el entrenamiento de los modelos, pero por la alta demanda computacional que requería el entrenamiento para el DistilBERT (y supongo que por tanto para el mBERT) tuve que utilizar Jupyter Notebook para el entrenamiento. Como ya tenía el código ahí seguí utilizando Jupyter como principal herramienta de edición de código, utilizando mi propio ordenador y mi GPU para realizar las tareas de entrenamiento.

En cuanto a las arquitecturas empleadas, el modelo principal fueron DistilBert para el entrenamiento del modelo en inglés y el mBert (BERT Multilingüe). Empecé con el BERT más tradicional hasta que ví que el DistilBert tenía un desempeño muy similar mientras que el coste computacional se reducía muchísimo. Integré un pipeline de traducción automática para convertir artículos en euskera a inglés, utilizando modelos preentrenados disponibles en la biblioteca Hugging Face, con el objetivo de analizar si los modelos entrenados en inglés obtenían resultados más consistentes en comparación con los entrenados en euskera e inglés (en el caso del multilingüe). Sin embargo al final me decanté por utilizar la API del traductor de Google para crear 2 datasets separados, uno en inglés y otro en euskera. Esto no solo hacía más cómodo el uso de los datos si no que las traducciones eran en general de mayor calidad.

Durante el diseño del sistema, evalué la posibilidad de utilizar Big Bird, una arquitectura optimizada para manejar secuencias más largas de texto. Esto podría haber resultado muy útil por la naturaleza del problema. Al entrenar los modelos con articulos de wikipedia sería recomendable tener el contexto de TODO el articulo, que a veces rebasa por mucho los 512 tokens que BERT suele permitir. Sin embargo trás hablar con los tutores llegué a dos conclusiones que me hicieron decantarme por los modelos mBERT y DistilBert:

1- Muchas veces los 512 primeros tokens son suficientes para determinar cómo de bueno es un articulo. Si bien para diferenciar articulos buenos de articulos excelentes o muy buenos es necesario tener un amplio contexto del articulo, muchas veces los 512 primeros tokens nos sirven para determinar si un articulo es bueno, malo o muy malo. Es un análisis más simple del que se podría haber hecho con Big Bird, pero leer todos los articulos habría supuesto muchísimo coste computacional.

2- En este trabajo no se busca encontrar modelos que realicen su trabajo de forma excepcional, el estudio del problema y las conclusiones sacadas de los modelos es más importante.

4 Datos

En cuanto a los datos se han utilizado principalmente dos conjuntos de datos en inglés: wiki60k y 2017 english wikipedia quality dataset, ambos centrados en la evaluación de calidad de artículos de Wikipedia. Inicialmente tenía pensado trabajar con el segundo conjunto de datos o realizar una combinación de ambos, pero opté por el primero debido a la simplicidad en el preprocesamiento. Finalmente utilicé el dataset wiki60k, seleccionando únicamente los primeros 16,000 artículos. Consideré que era una buena cifra que balanceaba bien el coste computacional con el rendimiento (por si acaso le metí un dropout por miedo a que el número de datos fuese menor del necesario para un buen rendimiento. Sorprendentemente, el dropout bajó drásticamente el porcentaje de articulos bien etiquetados).

En cuanto al euskera, creé un conjunto de datos propio utilizando la Wikipedia en euskera como fuente principal y me apoyé en la API de ORES para obtener las puntuaciones de calidad de los

artículos. Este corpus consta de un total de 1,500 artículos (para la evaluación) y 3500 para el entrenamiento del modelo multilingüe. Introduje datos en euskera a este modelo mBert porque consideré que el modelo podría aprender ciertos patrones que solo salen en articulos en euskera. Los datasets presentan un desbalance significativo, ya que la mayoría de los artículos en euskera son de baja calidad y están clasificados como stub, mientras que las categorías de mayor calidad (b, ga, fa, etc.) están representadas por un número considerablemente menor de ejemplos.

En cuanto al pre-procesamiento, para ambos idiomas se realizó un proceso para limpiar los artículos, eliminando elementos como metadatos, encabezados, referencias y cualquier contenido no textual, de modo que los datos consistieran principalmente en texto plano (con la información relevante). Además tuve combinar y separar los distintos datasets varias veces para ir realizando las pruebas. También tuve que cuadrar las puntuaciones de ORES con los artículos correspondientes, asegurando la coherencia entre las características textuales y las etiquetas de calidad. Gracias a estos pasos pude crear distintos datasets con los que poder alimentar a las redes directamente.

5 Resultados

Resultados durante el entrenamiento.

[1944/1944 18:15:19, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Accuracy
1	1.196300	1.124836	0.520833
2	1.019300	1.052727	0.560442

Figure 1: Métricas en el entrenamiento del modelo multilingüe (DEV).

Epoch	Training Loss	Validation Loss	Accuracy
1	1.513600	1.238145	0.491563
2	1.172600	1.195667	0.510312
3	1.061200	1.182609	0.518437
4	0.924000	1.209466	0.519375

Figure 2: Métricas en el entrenamiento del modelo monolingüe (DEV).

Resultados durante el TEST.

```
Clasificación usando el modelo:
Classification Report:
              precision    recall  f1-score   support

          FA       0.00      0.00      0.00         3
           C       0.23      0.50      0.31       221
        Stub       0.00      0.00      0.00       811
           B       0.00      0.00      0.00        13
          GA       0.00      0.33      0.00         6
       Start       0.00      0.00      0.00       420

    accuracy                           0.08      1474
   macro avg       0.04      0.14      0.05      1474
weighted avg       0.03      0.08      0.05      1474


Confusion Matrix:
[[  0   2   0   0   1   0]
 [  0 110   0   0 111   0]
 [  0  38   0   0 773   0]
 [  0   9   0   0   4   0]
 [  0   4   0   0   2   0]
 [  0 322   0   0  98   0]]
```

Figure 3: Métricas del test del modelo multilingüe.

```
Clase mayoritaria: Stub

Clasificación usando la clase mayoritaria:
Classification Report:
              precision    recall  f1-score   support

          FA       0.00      0.00      0.00         3
           C       0.00      0.00      0.00       221
        Stub       0.55      1.00      0.71       811
           B       0.00      0.00      0.00        13
          GA       0.00      0.00      0.00         6
       Start       0.00      0.00      0.00       420

    accuracy                           0.55      1474
   macro avg       0.09      0.17      0.12      1474
weighted avg       0.30      0.55      0.39      1474


Confusion Matrix:
[[  0   0   3   0   0   0]
 [  0   0 221   0   0   0]
 [  0   0 811   0   0   0]
 [  0   0  13   0   0   0]
 [  0   0   6   0   0   0]
 [  0   0 420   0   0   0]]
```

Figure 4: Métricas utilizando la clase mayoritaria.

```
Clasificación usando el modelo:
Classification Report:
              precision    recall  f1-score   support

        Stub       0.83      0.01      0.01       811
          FA       0.00      0.00      0.00         3
          GA       0.00      0.00      0.00         6
           B       0.00      0.00      0.00        13
       Start       0.00      0.00      0.00       420
           C       0.37      0.20      0.26       221

    accuracy                           0.03      1474
   macro avg       0.20      0.03      0.05      1474
weighted avg       0.51      0.03      0.05      1474


Confusion Matrix:
[[  5   0 802   0   0   4]
 [  0   0   0   0   0   3]
 [  0   0   0   0   0   6]
 [  0   0   0   0   0  13]
 [  0   2 366   0   0  52]
 [  1   0 175   0   0  45]]
```

Figure 5: Métricas del test del modelo monolingüe.

6 Análisis

6.1 Modelo Multilingüe

Podemos ver claramente que el modelo multilingüe entrenado con datos en inglés y euskera, muestra un desempeño muy pobre al evaluarse exclusivamente con artículos en euskera, alcanzando una precisión global de apenas el 8% y métricas como el f1-score prácticamente nulas para la mayoría de las clases. Esto se puede deber a varios factores, pero lo más seguro es que se deba al desbalance que existe entre los datos en inglés y euskera utilizados durante el entrenamiento, lo que podría haber sesgado al modelo hacia el idioma predominante, en este caso el inglés. Los modelos multilingües tienden a favorecer los idiomas con mayor representación en los datos, ya que tienen que repartir su capacidad para aprender patrones de múltiples lenguas. Además la complejidad morfológica del euskera también representa un desafío en sí mismo. Este idioma tiene una estructura altamente aglutinante, con palabras largas y una morfología rica que puede ser difícil de capturar para un modelo que no este 100% especializado en este idioma si no se le proporcionan suficientes datos representativos (que en este caso, no tenemos, pues solo había 3500 articulos en euskera en la fase de entrenamiento).

Por eso mismo podemos concluir que otro factor importante a considerar es que aunque el modelo es multilingüe, no está específicamente diseñado para el euskera, y por tanto no aprovecha de manera óptima las características únicas de este idioma. Aunque los modelos multilingües son útiles para manejar varios idiomas simultáneamente, lo cual nos ayuda a aprovecharnos de la gran cantidad de datos que hay en inglés para entrenar un modelo que también sirva para el euskera, su capacidad para aprender patrones específicos de un idioma se diluye al intentar abarcar demasiados lenguajes. Esto se refleja en los resultados de las métricas, donde clases como "Stub" y "Start", a pesar de tener más ejemplos en el conjunto de evaluación, no son correctamente clasificadas. Esto podría sugerirnos que el modelo no ha aprendido a diferenciar adecuadamente entre las clases, probablemente debido al desbalance o a una falta de señales claras en los datos de entrenamiento. El hecho de que tanto el dataset de entrenamiento como de evaluación tuviese tantos articulos de categoría 'STUB' y tan pocos de otras categorías puede desembocar en esto mismo, que el modelo no haya aprendido a diferenciar distintos artículos.

Para mejorar estos resultados podríamos tomar varias medidas: Aumentar la cantidad de datos en euskera disponibles para el entrenamiento sería una de las soluciones más efectivas, pero dado que no hay datasets de articulos de wikipedia en euskera (y mucho menos datasets que incluyan las puntuaciones ORES, y mucho menos aún datasets en los que estas puntuaciones estén relativamente balanceadas) crear un dataset tan grande por mi cuenta hubiese sido muy costoso. También se podrían aplicar técnicas de aumento de datos (data augmentation) para incrementar la representación de clases minoritaas y patrones lingüísticos. También podríamos entrenar un modelo especializado únicamente en euskera, lo que permitiría al modelo centrarse exclusivamente en las características únicas del idioma. Este fine-tuning podría realizarse partiendo de un modelo previamente entrenado como mBERT, pero utilizando únicamente datos en euskera. En principio tenía pensado entrenar el modelo solo en inglés, pero daba aún peores resultados que el modelo que estamos discutiendo.

Por supuesto el mayor problema, como hemos comentado, es el desbalance entre clases. Esto puede hacerse mediante técnicas como el remuestreo de clases minoritarias o la ponderación en la función de pérdida (weighted loss). El funcionamiento del tokenizador también podría estar influyendo, ya que podría estar fragmentando demasiado las palabras por la estructura morfológica rica del euskera, haciendo que el modelo no aprenda las relaciones semánticas debidamente.

Por último podríamos optimizar los hiperparámetros del modelo. He intentado entrenar el modelo a lo largo de varias epochs reduciendo el tamaño de los lotes, pero los resultados eran muy similares. Aumentar las epochs, cambiar el tamaño de los lotes y los dropouts no cambiaban mucho el rendimiento de los modelos. Otras técnicas utilizadas por ORES también podrían utilizarse para realizar mejores predicciones. Un conteo de la cantidad de secciones que tiene un articulo podría haber sido una buena mejora a la hora de realizar las predicciones, pero una vez más, como el entrenamiento y el test tienen tantísimos articulos STUB con solo 2 o 3 secciones (y solo 1 de texto 'útil', el resto siendo de enlaces y referencias) esto prácticamente no serviría, ya que se produciría un overfitting entre la puntuación ORES y el número de secciones por artículo. Una vez más, el problema es la cantidad tan pequeña de datos de entrenamiento

y el desbalanceo de clases.

6.2 modelo monolingüe

El modelo monolingüe DistilBERT, entrenado con datos traducidos del euskera al inglés y evaluado en artículos en inglés nos presenta resultados mixtos, con un buen desempeño en la clase mayoritaria ("Stub"), donde alcanza una precisión del 83%, pero un recall extremadamente bajo (1%), y métricas nulas o muy bajas en otras clases como "GA" y "FA".

Esto se podría deber a la calidad de las traducciones, que aunque correctas en su mayoría (dentro de lo que cabe), pueden no ser del todo exactas (sobre todo por la complejidad del euskera). A su vez, seguimos con el mismo problema que antes, el desbalance de clases en los datos, donde "Stub" domina con 811 ejemplos frente a muy pocos en clases como "GA" y "FA". Esto afecta la capacidad del modelo para generalizar. Además, DistilBERT, al ser una versión reducida de BERT, puede carecer de la capacidad necesaria para manejar un dominio multietiqueta con patrones complejos. Los resultados entre BERT y distilBERT no variaban mucho, pero si que algo puede haber influido. Al igual que en el anterior caso, sería esencial aumentar la cantidad de datos traducidos y balancear las clases mediante técnicas de re-muestreo o aumento de datos.

6.3 Clase mayoritaria

Este es un enfoque extra, no basado en modelos si no la técnica de la clase mayoritaria. En este enfoque, como era de esperarse, se muestra un desempeño altamente sesgado hacia la clase dominante "Stub", alcanzando un f1-score de 0.71 para esta clase, con un recall perfecto (1.00) ya que predice siempre "Stub" y, por tanto, clasifica correctamente todos los ejemplos de esta clase. Sin embargo, el desempeño en las demás clases es nulo, con métricas de precisión, recall y f1-score de 0.00 en todas las demás. La precisión global (accuracy) es del 55%, lo que corresponde al porcentaje de ejemplos que pertenecen a la clase mayoritaria en el conjunto de datos. Aquí podemos ver por qué en el entrenamiento los resultados de las predicciones son tan 'buenos'. La macro media (macro avg), que pondera por igual todas las clases, es extremadamente baja (0.17 en recall y 0.09 en precisión), evidenciando el desbalance severo en la representación de clases. Este comportamiento muestra claramente las limitaciones de este enfoque basado únicamente en la clase mayoritaria, que no captura ninguna de las relaciones entre características ni diferencias entre las clases.

7 Conclusiones y comparaciones

Los tres métodos utilizados (el modelo multilingüe entrenado con inglés y euskera, el modelo monolingüe entrenado con textos traducidos al inglés, y la técnica de asignar la clase mayoritaria) presentan varias diferencias. El modelo multilingüe mostró el peor desempeño general con una precisión del 8%, mostrando una clara falta de especialización en euskera y posiblemente un desbalance de datos entre idiomas durante el entrenamiento, lo que lo sesgó hacia patrones no aplicables al euskera. Probablemente los resultados hubiesen sido mejores si todo el dataset de entrenamiento hubiese estado en euskera a pesar de la naturaleza multilingüe del modelo. El modelo monolingüe entrenado con textos traducidos al inglés logró un desempeño ligeramente mejor, con una precisión del 51% en el weighted average, gracias a que trabajó con datos más homogéneos en inglés, aunque su efectividad se vio limitada por las inconsistencias semánticas introducidas por la traducción y por el desbalance de clases, que penalizó fuertemente las clases minoritarias. Este segundo factor es más importante que el primero, pues como hemos dicho, las traducciones estaban bastante decentes. Para terminar, la técnica de asignar la clase mayoritaria destacó como una línea base al alcanzar una precisión global del 55%. Es curioso que en este caso una medida tan simple sirva para clasificar mejor los documentos que los demás modelos, pero esto solo ocurre por la gran cantidad de articulos "STUB" que existen los datasets.

En generañ, el modelo monolingüe tiene la ventaja de capturar algunos patrones de los datos traducidos, lo que le permite superar al modelo multilingüe en desempeño, aunque sigue limitado por problemas inherentes de los datos y la arquitectura reducida de DistilBERT. La técnica de la clase mayoritaria, aunque útil como referencia, carece de capacidad para clasificar correctamente más allá de la clase dominante. En resumen, el modelo multilingüe es el menos efectivo debido a su falta de ajuste al idioma objetivo. Esto va en contra de lo que yo creía inicialmente, pues pensaba que la traducción iba a suponer una mayor reto del esperado, mientras que pensaba que el modelo multilingüe iba a adaptarse mejor al euskera aún con los datos

en inglés. El modelo monolingüe logra un balance intermedio pero sufre un poco por las traducciones y el constante desbalance, mientras que la técnica de la clase mayoritaria, aunque sea simple y efectiva en este caso, es completamente inaplicable para un uso realista con clases balanceadas (solo la he usado para tener una baseline para el estudio del rendimiento de los modelos). Con un mejor preprocesado, mejores equipos que soporten entrenamientos de redes más grandes y mayores datasets, probablemente el modelo multilingüe hubiese funcionado bastante mejor que el monolingüe. Un estudio interesante añadido podría ser ver si cogiendo todos los articulos del inglés, realizando una traducción al euskera de todos ellos y añadiendo todos esos articulos traducidos a wikipedia las puntuaciones de las páginas mejoran o empeoran. Claramente la redacción sería peor, ¿pero merecería eso la pena si pudiesemos tener muchísima más información a cambio? Es un proyecto a futuro interesante que no solo ayudaría a aumentar la cantidad de datos para el PLN en euskera, si no a la propia comunidad de wikipedia en sí.