Polo del Conocimiento

Pol. Con. (Edición núm. 103) Vol. 10, No 2
Febrero 2025, pp. 839-860
ISSN: 2550 - 682X
DOI: https://doi.org/10.23857/pc.v10i2.8920

Deep Learning y la arquitectura Transformer: Evaluación del Desempeño de RoBERTa-large-bne en la Predicción de la Complejidad Léxica en Textos Estatales Ecuatorianos

Deep Learning and Transformer Architecture: Evaluating the Performance of RoBERTa-large-bne in Predicting Lexical Complexity in Ecuadorian State Texts

Aprendizagem profunda e arquitetura de transformadores: avaliando o desempenho de RoBERTa-large-bne na previsão da complexidade lexical em textos estaduais equatorianos

Jenny Alexandra Ortiz-Zambrano [I]
jenny.ortizz@ug.edu.ec
https://orcid.org/0000-0001-6708-4470

Arturo Montejo-Ráez [II]
amontejo@ujaen.es
http://orcid.org/0000-0002-8643-2714

Correspondencia: jenny.ortizz@ug.edu.ec

Ciencias Técnicas y Aplicadas
Artículo de Investigación

*** Recibido:** 01 de diciembre de 2024 ***Aceptado:** 24 de enero de 2025 *** Publicado:** 11 de febrero de 2025

I. Doctorante en Tecnologías de la Información y Comunicación por la Universidad de Jaén, España.
II. Doctor en Ciencias de la Computación, Ecuador.

Resumen

En el ámbito de las instituciones públicas, la accesibilidad y comprensión de los documentos gubernamentales suelen verse afectadas por la complejidad léxica y el uso de lenguaje técnico especializado. Este problema impacta tanto a ciudadanos como a funcionarios, generando malentendidos que pueden reducir la transparencia y dificultar la participación ciudadana. En este estudio, se analiza el desempeño del modelo RoBERTa-large-bne, basado en la arquitectura Transformer, en la predicción de la complejidad léxica en textos estatales ecuatorianos. Para ello, se implementó un ajuste fino del modelo con el fin de optimizar su rendimiento en esta tarea específica. Se llevó a cabo una evaluación comparativa con otros modelos de lenguaje pre-entrenados aplicados en el corpus GovAIEc, cuyos textos corresponden a instituciones estatales del Ecuador. Los resultados obtenidos buscan sentar las bases para el desarrollo de herramientas que faciliten la simplificación de documentos públicos, mejorando su accesibilidad y promoviendo una interacción más eficiente entre la ciudadanía y las instituciones gubernamentales.

Palabras clave: Aprendizaje profundo; Complejidad léxica; Textos estatales; Arquitectura Transforme; Predicción.

Abstract

In the field of public institutions, the accessibility and understanding of government documents are often affected by lexical complexity and the use of specialized technical language. This problem impacts both citizens and officials, generating misunderstandings that can reduce transparency and hinder citizen participation. In this study, the performance of the RoBERTa-large-bne model, based on the Transformer architecture, is analyzed in the prediction of lexical complexity in Ecuadorian state texts. To do so, a fine-tuning of the model was implemented in order to optimize its performance in this specific task. A comparative evaluation was carried out with other pre-trained language models applied to the GovAIEc corpus, whose texts correspond to state institutions in Ecuador. The results obtained seek to lay the foundations for the development of tools that facilitate the simplification of public documents, improving their accessibility and promoting a more efficient interaction between citizens and government institutions.

Keywords: Deep learning; Lexical complexity; State texts; Transformer architecture; Prediction.

Resumo

No domínio das instituições públicas, a acessibilidade e a compreensão dos documentos governamentais são frequentemente afetadas pela complexidade lexical e pela utilização de linguagem técnica especializada. Este problema afeta tanto os cidadãos como as autoridades, gerando mal-entendidos que podem reduzir a transparência e dificultar a participação dos cidadãos. Neste estudo, o desempenho do modelo RoBERTa-large-bne, baseado na arquitetura Transformer, é analisado na previsão da complexidade lexical em textos estaduais equatorianos. Para tal, foi implementado um ajuste fino do modelo de forma a otimizar o seu desempenho nesta tarefa específica. Foi realizada uma avaliação comparativa com outros modelos de linguagem pré-treinados aplicados no corpus GovAIEc, cujos textos correspondem a instituições estatais do Equador. Os resultados obtidos procuram lançar as bases para o desenvolvimento de ferramentas que facilitem a simplificação dos documentos públicos, melhorando a sua acessibilidade e promovendo uma interação mais eficiente entre os cidadãos e as instituições governamentais.

Palavras-chave: Aprendizagem profunda; Complexidade lexical; Textos do Estado; Transformar Arquitetura; Previsão.

Introducción

Si bien los textos son los principales portadores de información para la toma de decisiones gubernamentales, pocos estudios han examinado el papel de la complejidad textual en la comunicación entre el gobierno y los ciudadanos (Lu et al., 2023). Sin embargo, la complejidad léxica y el uso de un lenguaje técnico especializado representan un obstáculo significativo para muchas personas. Esta barrera afecta tanto a los ciudadanos, quienes pueden tener dificultades para comprender información clave sobre sus derechos y obligaciones, como a las instituciones públicas, que buscan comunicar de manera clara y efectiva sus políticas y procedimientos. (Wold et al., 2024).

La dificultad de comprensión no solo disminuye la claridad de la comunicación, sino que también puede generar malentendidos, reduciendo la transparencia y limitando la participación ciudadana. Tal como lo señala la investigación, "la complejidad de las palabras, la lógica y las emociones en el texto aumentan la dificultad del procesamiento de la información, por lo que la complejidad a menudo se considera la encarnación de la baja calidad de la información" (Alter y Oppenheimer, 2009; Graf et al., 2018, como se citó en Lu et al. 2023). Estos factores dificultan que los ciudadanos

comprendan los documentos gubernamentales, lo que genera una desconexión en la interacción entre la población y las instituciones públicas. Además, los estudios han mostrado que "un texto simple puede lograr una mayor participación", mientras que la complejidad puede reducir la interacción y el entendimiento en contextos donde es crucial que la comunicación sea clara (Markowitz y Shulman, 2021, como se citó en Lu et al., 2023).

A pesar de estos hallazgos, en Ecuador no se ha explorado suficientemente el uso de tecnologías avanzadas, como los modelos basados en la arquitectura Transformer, para predecir y abordar la complejidad léxica en textos públicos. Los modelos Transformer, como GPT-3 y BERT, han demostrado un rendimiento sobresaliente en tareas de Procesamiento de Lenguaje Natural, incluyendo la generación de textos realistas y coherentes, y la clasificación de secuencias textuales. Estos modelos pueden ser aplicados para identificar términos complejos y ayudar a simplificar el lenguaje de los documentos gubernamentales, lo que mejoraría su accesibilidad. Dada la capacidad de los Transformers para capturar dependencias de largo alcance en los textos mediante mecanismos de atención multi-cabezal, estos modelos ofrecen una herramienta prometedora para abordar la complejidad léxica de manera efectiva (Mo et al., 2024).

Este estudio pretende evaluar la complejidad léxica de los textos públicos ecuatorianos mediante la ejecución de los modelos BERT, y RoBERTa para determinar la complejidad de las palabras, y analizar el performance (rendimiento) de los modelos en un conjunto de datos en español. Esta investigación es una contribución al campo del Procesamiento del Lenguaje Natural como apoyo de la accesibilidad de los ciudadanos a la información gubernamental, fomentando así una interacción más inclusiva y efectiva entre las instituciones y la ciudadanía.

La investigación de Azucena y Yanet (2021) se enfoca en la educación inclusiva en Ecuador, examinándola tanto desde la perspectiva legal como educativa. El estudio analiza las barreras del lenguaje y la terminología presentes en las leyes actuales, mostrando cómo la complejidad del lenguaje legal puede entorpecer la comprensión y aplicación de las leyes, especialmente para personas con niveles educativos más bajos. Los autores enfatizan la necesidad de simplificar el lenguaje legal y crear herramientas que faciliten la interpretación de las leyes, buscando un acceso igualitario a la educación y a la información legal. Adicionalmente, proponen estrategias pedagógicas y metodológicas para optimizar la interacción entre los ciudadanos y las instituciones gubernamentales.

De acuerdo con Ortiz et al. (2022) presentan una contribución a la predicción de la complejidad de palabras simples en español mediante la combinación de múltiples características. Se emplearon modelos basados en Transformers, como BERT, XLM-RoBERTa y RoBERTa-large-BNE, ejecutados en algoritmos de regresión. Los mejores resultados se obtuvieron con el modelo BERT refinado y el algoritmo Random Forest Regressor, logrando un MAE de 0.1598 y un coeficiente de Pearson de 0.9883. Como trabajo futuro, se propone experimentar con más conjuntos de datos en español y modelos Transformers avanzados para mejorar la predicción de la complejidad léxica.
El Ministerio de Telecomunicaciones de la República del Ecuador - (MINTEL) emitió en 2022 la "Norma Técnica para la Priorización y Simplificación de Trámites". Este proyecto buscó optimizar los procesos administrativos mediante la eliminación de redundancias y la digitalización de servicios públicos. Aunque se centró en la estructura de los procedimientos, incluyó la necesidad de evaluar la claridad lingüística de las normativas. Sin embargo, no se incorporaron herramientas avanzadas de Procesamiento de Lenguaje Natural (PLN). Esto evidencia un área de oportunidad para implementar tecnologías como Transformers, que podrían automatizar la identificación de términos complejos y contribuir a una mayor accesibilidad de estos textos (Ministerio de Telecomunicaciones y de la Sociedad de la Información, 2022).
Según Ortiz (2023) en este estudio introducen a LegalEc, un nuevo corpus anotado de léxico complejo basado en textos legales en español ecuatoriano, con detalles sobre su proceso de compilación y anotación. Como recurso para avanzar en la investigación sobre simplificación léxica en español, se realizaron experimentos de predicción de palabras complejas utilizando 23 características lingüísticas combinadas con codificaciones generadas por modelos como XLM-RoBERTa y RoBERTa-BNE. Los resultados demuestran que esta combinación mejora la predicción de la complejidad léxica.
El estudio "BERT for Legal Texts: Training and Fine-tuning in a New Language" abordó el desafío de aplicar BERT en el ámbito legal mediante su ajuste a lenguajes específicos y la adaptación al contexto jurídico, entrenando el modelo en un corpus extenso de documentos legales en múltiples idiomas, como inglés, alemán y francés. La metodología incluyó el fine-tuning del modelo BERT en datos legales para tareas como la clasificación de sentencias y la segmentación de contratos, además de una evaluación basada en métricas como F1-score y precisión para medir su efectividad en tareas específicas. Los resultados mostraron una mejora del 12% en la clasificación de documentos legales en comparación con métodos tradicionales, con ajustes que permitieron

identificar términos complejos y relaciones semánticas propias del ámbito legal. Esto valida el uso de Transformers en textos especializados y demuestra que los ajustes contextuales pueden mejorar significativamente la precisión en tareas específicas, lo que resulta especialmente relevante para proyectos enfocados en la aplicación de modelos de lenguaje en dominios técnicos (Soneji et al., 2024).

Si bien modelos como GPT-3 y GPT-4 han mostrado un gran potencial en la predicción de términos complejos mediante las técnicas de zero-shot learning y few-shot learning, BERT y RoBERTa ofrecen oportunidades únicas en la clasificación de secuencias y predicción de complejidad léxica mediante ajuste fino. Como lo demuestran (Devlin, et al., 2019) estos modelos, basados en la arquitectura Transformer, han demostrado ser altamente efectivos en tareas de clasificación y predicción de complejidad léxica, lo que podría mejorar significativamente la accesibilidad de los textos gubernamentales mediante técnicas de simplificación y adaptación léxica de acuerdo a los estudios realizados por (Ortiz et al., 2024).

Para Moscoso y Pacheco (2024) desarrollaron en la Universidad de Cuenca proyectos enfocados en la adaptación de modelos de PLN al español ecuatoriano. En particular, han trabajado con modelos como BERT para el análisis de sentimientos y clasificación de textos, demostrando su efectividad en contextos locales. Aunque no específicamente orientados a textos públicos, estos avances han puesto en evidencia la necesidad de ajustar los modelos a las particularidades lingüísticas y culturales del español hablado en Ecuador, como variaciones léxicas y construcciones sintácticas únicas.

Según Ortiz y Montejo (2024) en este estudio, se presenta un método innovador para la predicción de la complejidad léxica (LCP) que integra un conjunto diverso de propiedades lingüísticas con codificaciones de redes neuronales profundas. Para ello, se combinan 23 características lingüísticas artesanales junto con las representaciones generadas por dos modelos de lenguaje de amplia adopción: BERT y XLM-RoBERTa. El procedimiento consiste en concatenar dichas características antes de introducirlas en diversos algoritmos de aprendizaje automático, que abarcan desde SVM y Random Forest hasta modelos transformadores ajustados.

Para Soneji et al. (2024) en su estudio utilizó RoBERTa para analizar y simplificar políticas de privacidad y términos legales complejos mediante su entrenamiento con un corpus especializado de términos legales y políticas de privacidad. La metodología incluyó la implementación de mecanismos de atención para identificar las frases más relevantes en cada documento y la

validación de las predicciones al compararlas con resúmenes generados por expertos. Los resultados mostraron que RoBERTa logró un 85 % de concordancia con resúmenes manuales, además de reducir la redundancia en los textos legales y mejorar su accesibilidad para usuarios no especializados. Este antecedente subraya la capacidad de los Transformers para abordar el lenguaje técnico en documentos legales y facilitar su comprensión, lo que resulta relevante para proyectos enfocados en simplificar contenidos complejos.

Materiales y métodos

En el marco de esta investigación, se utilizan métodos estadísticos e informáticos avanzados para la recopilación y el análisis de datos científicos. Como destacan Zhang, et al. (2016):

La evaluación de modelos de lenguaje en tareas específicas, como la comprensión de textos, requiere no solo la recolección de datos, sino también la aplicación de técnicas analíticas rigurosas que permitan identificar patrones complejos y medir la eficacia de los modelos. Este enfoque asegura la obtención de resultados confiables y una interpretación contextualizada de los hallazgos. Este enfoque es fundamental para nuestra investigación, ya que nos permitió evaluar sistemáticamente el rendimiento de modelos como BERT o RoBERTa y otros sistemas basados en arquitecturas Transformer en un conjunto representativo de documentos públicos. Esto no solo garantizó la confiabilidad de los resultados, sino que también facilitó su generalización a un espectro más amplio de textos institucionales y gubernamentales.

Materiales

Conjunto De Datos

Para el tratamiento de los datos, se empleó el corpus GovAIEc como fuente documental, permitiendo obtener acerca de los textos públicos ecuatorianos. Durante este proceso, se aplicó el código necesario para analizar la complejidad de las palabras en los documentos oficiales, así como la generación de métricas asociadas a la complejidad léxica.

Adicionalmente, el proceso incluyó la división del conjunto de datos en dos proporciones: una destinada a la fase de entrenamiento del modelo y otra reservada para la fase de evaluación. Esto permitió medir de manera efectiva el desempeño de los modelos en contextos controlados. Como el objetivo de esta investigación es explorar el impacto de las características lingüísticas añadidas, se llevaron a cabo ejecuciones tanto con el conjunto de datos original (sin características

lingüísticas) como con una versión enriquecida que incorpora 17 nuevas características lingüísticas adicionales (LF) a las 23 características que contenía GovAIEc al inicio de su creación. Estas características buscan proporcionar al modelo un mejor entendimiento del contexto, optimizando su capacidad predictiva y mejorando los resultados generales.

Estructura del Corpus

GovAIEc es un corpus especializado en textos administrativos y legales provenientes de cinco fuentes gubernamentales ecuatorianas:

1. ATM[i] - Autoridad de Tránsito Municipa. Sitio web oficial: https://www.atm.gob.ec/
2. CNE[ii] - Consejo Nacional Electoral. Sitio web oficial: https://www.cne.gob.ec/
3. CNT[iii] - Corporación Nacional de Telecomunicación. Sitio web oficial: https://www.cnt.com.ec/
4. Muy ilustre municipalidad de guayaquil[iv]. Sitio web oficial: https://guayaquil.gob.ec/
5. SRI[v] - Servicio de Rentas Internas. Sitio web oficial: https://www.sri.gob.ec/web/intersri/home

El archivo principal del dataset contiene 7.813 registros, cada uno etiquetado con información relevante para la tarea de predicción de la complejidad léxica.

Estructura del Dataset

Cada registro dentro de GovAIEc está compuesto por los siguientes campos:

- ID: Identificador único asignado a cada registro.
- CORPUS: Fuente específica del registro dentro del dataset.
- SENTENCE: Párrafo donde se encuentra la palabra identificada como compleja.
- TOKEN: Palabra específica marcada para el análisis de complejidad léxica.
- COMPLEXITY: Valor numérico que representa el nivel de complejidad de la palabra según los etiquetadores.

Criterios de Evaluación y Comparación

Las experimentaciones realizadas con el modelo RoBERTa-large-bne pretenden analizar el rendimiento de bajo dos configuraciones:

1. Sin características lingüísticas, donde el modelo aprende exclusivamente de los datos textuales sin información adicional.
2. Con características lingüísticas (LF), incorporando información lingüística adicional para mejorar la predicción de la complejidad léxica.

Tabla1: *Registro*

ID	CORPUS	SENTENCE	TOKEN	COMPLEXITY
5667	Municipio - Tramites - TEXTO 0028 REQUISISTOS PARA LA CREACION DE URBANIZACIONES.txt	REQUISISTOS PARA LA CREACION DE URBANIZACIONES Tal situación debe contemplarse en el Plano Definitivo del Proyecto Urbanístico, así como en el Reglamento Interno de la Urbanización correspondiente, que se anexará en las escrituras de transferencia de dominio de los solares afectados, a efectos de consolidar la certeza jurídica de las propiedades a adquirirse	contemplarse	0.333
2502	CNE - Tramites - TEXTO 0073 REGLAMENTO DE PROMOCION ELECTORAL.txt	REGLAMENTO DE PROMOCION ELECTORAL, Además, se prohíbe durante la campaña electoral la contratación y difusión de propaganda y publicidad por parte de sujetos de derecho privado referente al proceso electoral en prensa escrita, radio, televisión, vallas publicitarias, medios digitales y cualquier otro medio de comunicación social	sujetos	0.333
7364	SRI - Tramites - TEXTO 0055 GUÍA PARA CONTRIBUYENTES INGRESO DE TRÁMITES Y ANEXOS A TRAVÉS DE SRI EN LÍNEA.txt	GUÍA PARA CONTRIBUYENTES INGRESO DE TRÁMITES Y ANEXOS A TRAVÉS DE SRI EN LÍNEA Paso 4 Notificación A continuación, ingrese los campos de dirección si no son correctos los que vienen precargados y si es necesario active la opción de notificación en el Casillero judicial	precargados	0.666

Nota: Esta tabla ofrece una descripción detallada del tipo de datos empleados en el entrenamiento y evaluación de los modelos, así como de la metodología utilizada para asignar los valores de complejidad léxica a las palabras dentro del corpus. Elaboración: Molina Vargas Jorge y Villota Viteri Kendrick. Fuente: Propia

El ID corresponde a un identificador único asignado a cada registro. El campo CORPUS indica la fuente de origen de cada ejemplo. SENTENCE representa el párrafo en el que se encuentra la

palabra etiquetada como compleja, mientras que TOKEN identifica la palabra seleccionada para el análisis de complejidad. Finalmente, COMPLEXITY es un valor numérico que representa el grado de complejidad asignado por los etiquetadores. La siguiente tabla presenta la escala de complejidad léxica utilizada para clasificar los textos en el dataset según su nivel de dificultad. Se empleó una escala de Likert, la cual asigna valores numéricos a las palabras o frases en función de su complejidad dentro del contexto del texto. La clasificación se organiza en tres categorías principales:

- **Moderadamente Difícil (Moderately Difficult):** Rango de 0 a 0.333. Las palabras o frases en esta categoría presentan cierta complejidad, pero suelen ser comprensibles dentro del contexto en el que aparecen.
- **Difícil (Difficult):** Rango de 0.333 a 0.666. Estas palabras o frases poseen un nivel de complejidad considerable y pueden requerir un mayor grado de comprensión o conocimientos específicos para su adecuada interpretación.

Muy Difícil (Very Difficult): Rango de 0.666 a 1. Las palabras o frases en esta categoría presentan un alto grado de complejidad, lo que puede representar un desafío significativo para la comprensión.

Además, el corpus GovAIEc incorpora un total de 40 características lingüísticas adicionales diseñadas para mejorar las predicciones de los modelos de aprendizaje. Estas características proporcionan información contextual y estructural que permite a los modelos comprender mejor la complejidad léxica de las palabras. Las características lingüísticas son las siguientes: Frecuencia absoluta, Frecuencia relativa, Longitud de la palabra, Número de silabas, Posición del token, Número de palabras en la oración, Part of speech, Frecuencia relativa de la palabra antes de la palabra objetivo (token), Frecuencia relativa de la palabra después de la palabra objetivo (token), Longitud de la palabra anterior, Longitud de la palabra que sigue, Medida de diversidad léxica textual, Número de sinónimos, Número de hipónimos, Número de hiperónimos, Número de sustantivos singular o plural, Número de verbos auxiliares, Número de adverbios, Número de símbolos, Número de expresiones numéricas, Número de verbos, Número de sustantivos, Número de pronombres, Número de morfemas, Longitud del lema, Is stopword (Es una palabra vacía), Número de sentidos de una palabra, Índice de legibilidad de Flesch, Índice de Gunning-Fog, Índice de SMOG, Índice RIX, n-gramas de caracteres, WordNet synset size, WordNet nomber of synset, Language model sentence probability, Average n-gram frecuency, Degree of Polyseny o número

de sentidos de la palabra objetivo en WordNet, Número de vocales, Word complexity lexicón, Phrase lenght in terms of words and characters.

Procesamiento de la información

Para el análisis de la complejidad léxica, se empleó el corpus GovAIEc junto con los modelos RoBERTa (con sus respectivas configuraciones) para generar predicciones de complejidad léxica. Con los valores obtenidos de las predicciones y los valores asignados a cada palabra en el corpus, se utilizaron métricas de evaluación como Validation Loss, Training Loss, MAE (Error Absoluto Medio), MSE (Error Cuadrático Medio), RMSE (Raíz del Error Cuadrático Medio) y R^2 (Coeficiente de Determinación) para contrastar los resultados obtenidos con los valores de referencia.

Métricas Utilizadas

Validation Loss

La pérdida de validación (validation loss) es una métrica utilizada en el entrenamiento de modelos de aprendizaje automático y aprendizaje profundo para evaluar el rendimiento del modelo en un conjunto de datos de validación. Se calcula aplicando la función de pérdida al conjunto de validación después de cada iteración o época de entrenamiento. Su propósito es monitorear si el modelo está generalizando correctamente a datos no vistos y detectar problemas como el sobreajuste (overfitting) (Baeldung, Training and Validation Loss in Deep Learning, 2024).

Training Loss

La pérdida de entrenamiento (training loss) es una métrica utilizada para evaluar qué tan bien un modelo de aprendizaje automático se ajusta a los datos de entrenamiento. Se calcula utilizando una función de pérdida específica después de cada iteración o época del entrenamiento. Una disminución en el training loss indica que el modelo está aprendiendo patrones a partir de los datos (Goodfellow et al., 2017).

Mean Absolute Error (MAE)

El Error Absoluto Medio (Mean Absolute Error, MAE) es una métrica utilizada en modelos de regresión para medir el promedio de los errores absolutos entre las predicciones del modelo y los valores reales. Se calcula mediante la siguiente fórmula:

$$MAE = \frac{1}{n} = \sum_{i=1}^{n} \gamma i - \gamma \, \mathrm{i}$$

Donde:

$\bar{\Upsilon} i$ representa el valor real.

$\bar{\Upsilon} i$ es la predicción del modelo.

N es el número total de observaciones.

El MAE proporciona una medida directa de la magnitud del error en unidades de la variable de salida sin considerar la dirección del error (positiva o negativa) (Willmott y Matsuura, 2005).

Mean Squared Error (Mse)

El Error Cuadrático Medio (Mean Squared Error, MSE) es una métrica de evaluación utilizada en modelos de regresión que mide el promedio de los errores al cuadrado entre los valores reales y las predicciones del modelo. Se define mediante la siguiente fórmula:

$$MSE = \frac{1}{n} = \sum_{i=1}^{n} (\gamma 1 - \bar{\Upsilon} i)^2$$

Donde:

γi representa el valor real.

$\bar{\Upsilon} i$ es la predicción del modelo.

n es el número total de observaciones.

El MSE eleva los errores al cuadrado, lo que penaliza más los errores grandes en comparación con los errores pequeños. (Chai & Draxler, 2014)

Root Mean Squared Error (Rmse)

El Error Cuadrático Medio de Raíz (Root Mean Squared Error, RMSE) es una métrica utilizada para evaluar la precisión de modelos de regresión, proporcionando una medida de la diferencia promedio entre los valores reales y las predicciones. Se calcula como la raíz cuadrada del Error Cuadrático Medio (Mean Squared Error, MSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\gamma 1 - \bar{\Upsilon} i)}$$

Donde:

γi representa el valor real.

$\bar{Y}i$ es la predicción del modelo.

n es el número total de observaciones.

El RMSE mide el error en las mismas unidades que la variable objetivo, lo que facilita su interpretación. (Chai y Draxler, 2014)

Coeficiente de Determinación (R^2)

El Coeficiente de Determinación (R^2), también conocido como el coeficiente de explicación, es una métrica utilizada para evaluar el desempeño de modelos de regresión. Indica qué proporción de la variabilidad en la variable dependiente (γ) es explicada por las variables independientes (X) en el modelo. (Palma, 2022)

Resultados y discusión

El análisis se centró en la comparación del modelo evaluado, examinando su desempeño en distintas configuraciones de épocas de entrenamiento (30, 50, 70) para determinar cómo el número de iteraciones impacta su rendimiento en la predicción de la complejidad léxica dentro del corpus GovAIEc. A continuación, la tabla 2 presenta los resultados finales alcanzados tras la ejecución del modelo roberta-large-bne alcanzando su mejor rendimiento con el conjunto de datos conformado por las características lingüísticas.

Tabla 2: *Modelos Predictivos Aplicados*

EPOCHS	MODELO	MAE	MSE	RMSE	R2
50	roberta-large-bne +LF	0,204512	0,053927	0,232223	0,039591
70	roberta-large-bne + LF	0,205496	0,053931	0,23223	0,039526
30	roberta-large-bne +LF	0,206888	0,053906	0,232176	0,039977

Nota*: El modelo incluido es RoBERTa-large-bne con la inclusión de características lingüísticas (LF). La tabla ilustra cómo el número de épocas (30, 50, 70) y la inclusión de LF impactan en el rendimiento de los modelos.*

A continuación, se describen los resultados alcanzados:

- Ejecución con 50 épocas (epochs): Esta configuración logra el mejor rendimiento general con los menores valores de MAE (0,204512) y MSE (0,053927). Además, alcanza un RMSE de 0.232223 y un R^2 de 0,039591, lo que indica un buen equilibrio entre precisión y capacidad explicativa.

- Ejecución con 70 épocas (epochs): Aunque mantiene un rendimiento similar al de 50 épocas, no logra mejorar significativamente las métricas clave. El MAE aumenta levemente a 0,205496, y el R^2 disminuye a 0,039526, lo que sugiere que el modelo alcanza una saturación al entrenar por más tiempo.
- Ejecución con 30 épocas (epochs): Presenta resultados competitivos, con el mejor R^2 (0,039977) y un MSE ligeramente menor (0,053906). Sin embargo, el MAE más alto (0,206888) indica una menor precisión en las predicciones.

Tabla 3: *Resultados de la mejor configuración del modelo alcanzados con 50 épocas*

Epoch	Training Loss	Validation Loss	MAE	MSE	RMSE	R2
1	0,1186	0,071007	0,232025	0,070902	0,266274	-0,262713
2	0,1013	0,083054	0,250094	0,082998	0,288093	-0,478131
3	0,0941	0,067434	0,227909	0,067317	0,259455	-0,198867
4	0,0893	0,062287	0,223068	0,062135	0,249269	-0,106581
5	0,0811	0,068457	0,228509	0,068336	0,261411	-0,217009
6	0,0802	0,070329	0,230048	0,070216	0,264983	-0,250499
7	0,0781	0,062365	0,222919	0,062209	0,249417	-0,107896
8	0,0722	0,055097	0,211709	0,05487	0,234243	0,022806
9	0,0728	0,060072	0,219854	0,059905	0,244754	-0,066861
10	0,0689	0,057378	0,216287	0,057189	0,239142	-0,018492
…	…	…	…	…	…	…
33	0.0572	0.054196	0.204512	0.053927	0.232223	0.039591
…	…	…	…	…	…	…
48	**0,0567**	**0,05474**	**0,210431**	**0,054511**	**0,233476**	**0,029197**
49	0,0563	0,054621	0,209963	0,054389	0,233215	0,031367
50	0,0562	0,054595	0,209852	0,054362	0,233157	0,031851

Nota*: Se ilustra en la tabla los resultados del entramiento del mejor modelo.*

Los resultados obtenidos muestran que el modelo roberta-large-bne + LF (FEATURES) con 50 épocas logró un desempeño consistente y eficiente a lo largo del entrenamiento. Durante las primeras 10 épocas, el modelo experimenta una reducción significativa en las métricas de error, con el MAE disminuyendo de 0,232025 (en la época 1) a 0,216287 (en la época 10). Esto refleja un aprendizaje rápido en las etapas iniciales. De manera similar, el MSE muestra una disminución notable, pasando de 0,070902 a 0,057189 en el mismo intervalo.

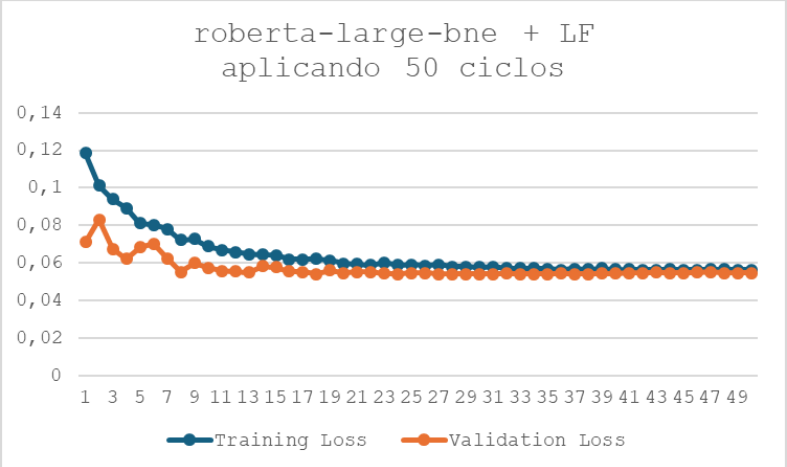
A partir de la época 30, las métricas clave como MAE y MSE comienzan a estabilizarse. El MAE oscila alrededor de 0,206-0,208, mientras que el MSE se mantiene cercano a 0,0539. Esto indica que el modelo alcanza un equilibrio entre precisión y generalización. En particular, la época 33 representa el punto óptimo del modelo, alcanzando un MAE de 0,204512 y un MSE de 0,053927, los valores más bajos registrados. Además, el coeficiente R^2 en esta época es de 0,039591, lo que sugiere una ligera mejora en la capacidad de explicar la variabilidad de los datos.

El rendimiento global del modelo demuestra robustez, con el RMSE permaneciendo estable alrededor de 0,232 durante las últimas épocas. Aunque los valores de R^2 no son altos, la mejora progresiva desde valores negativos (-0,262713 en la época 1) hasta positivos (0,039591 en la época 33) indica un refinamiento constante en la capacidad explicativa del modelo. Entrenar más allá de la época 33 no aporta mejoras significativas, como lo reflejan las métricas consistentes en las últimas epocas, lo que también sugiere que continuar el entrenamiento podría resultar en un uso ineficiente de recursos computacionales.

El modelo roberta-large-bne + LF (FEATURES) se destacó por su capacidad de lograr un equilibrio entre precisión y estabilidad. Su rendimiento óptimo, alcanzado en la época 33, demuestra su eficacia para la predicción de complejidad léxica, justificando su elección como la configuración final para esta tarea.

Análisis de la pérdida de entramiento y pérdida de validación aplicando 50 épocas

Figura 1: *Training Loss y Validation Loss (roberta-large-bne + LF, 50 Epochs)*



Nota*: Se ilustra un gráfico comparativo de los resultados del training y validation loss en el modelo roberta-large-bne con características linguisticas aplicando 50 ciclos.*

Comportamiento General de los Valores de Loss

Training Loss:

- Comienza en **0,1186** en la primera época y desciende progresivamente hasta estabilizarse en torno a **0,056** hacia las últimas épocas.
- Esto indica que el modelo está aprendiendo de manera consistente durante el entrenamiento, reduciendo su error en los datos de entrenamiento con cada ciclo.

Validation Loss

Inicia en **0,071007** en la primera época y desciende hasta estabilizarse cerca de **0,054**, mostrando un comportamiento similar al Training Loss. El Validation Loss se estabiliza aproximadamente a partir de la época 15, lo que sugiere que el modelo alcanza un punto donde mejora solo marginalmente en los datos de validación.

Convergencia entre Training y Validation Loss

Diferencia Inicial

En las primeras épocas, hay una notable diferencia entre el Training Loss y el Validation Loss. Por ejemplo:

Época 1: Training Loss = **0,1186**, Validation Loss = **0,071007**.

Esto es normal al inicio del entrenamiento, ya que el modelo aún está ajustándose a los datos.

Reducción de la Brecha

A medida que avanzan los ciclos, las diferencias entre ambos se reducen. Por ejemplo:

- Época 30: Training Loss = **0,0573**, Validation Loss = **0,054255**.
- Época 50: Training Loss = **0,0562**, Validation Loss = **0,054595**.

Esto indica que el modelo logra un buen equilibrio entre su capacidad para ajustarse a los datos de entrenamiento y generalizar a los datos de validación.

Pol. Con. (Edición núm. 103) Vol. 10, No 2, Febrero 2025, pp. 839-860, ISSN: 2550 - 682X

Ausencia de Sobreajuste

Si el Training Loss disminuyera constantemente mientras el Validation Loss comenzara a aumentar, indicaría que el modelo está sobreajustándose (memoriza los datos de entrenamiento, pero no generaliza bien). En este caso:

- Ambos valores (Training Loss y Validation Loss) se estabilizan en niveles similares hacia el final del entrenamiento (50 épocas).
- Esto sugiere que el modelo no muestra signos significativos de sobreajuste, lo que es un comportamiento deseable.

Estabilización del Modelo

Ambos valores de pérdida se estabilizan a partir de la época 30, con cambios marginales en las últimas épocas.

Por ejemplo:

- Época 30: Training Loss = **0,0573**, Validation Loss = **0,054255**.
- Época 50: Training Loss = **0,0562**, Validation Loss = **0,054595**.

Esto sugiere que entrenar más allá de 30-40 épocas podría no aportar mejoras significativas y, por tanto, podría ser una oportunidad para reducir el tiempo de entrenamiento.

Interpretación de los Valores Finales

En la época 50:

- **Training Loss**: **0,0562**.
- **Validation Loss**: **0,054595**.

Estos valores bajos y cercanos entre sí indican que el modelo tiene una alta precisión en los datos de entrenamiento y generaliza bien en los datos de validación.

Conclusiones

Las mejores ejecuciones se obtuvieron ejecutando roberta-large-bne + LF con 50 épocas, el cual tuvo el mejor rendimiento general, logrando un MAE = 0,204512 y un MSE = 0,053927, lo que indica su capacidad para predecir la complejidad léxica con mayor precisión. La inclusión de características lingüísticas (+LF) mejora significativamente el rendimiento del modelo, al proporcionar información adicional sobre la estructura y el contexto de las palabras. Esto permite

al modelo capturar patrones complejos que no son evidentes únicamente a partir del texto, incrementando la precisión y estabilidad de las predicciones. Además, entrenar por más de 50 epochs no proporciona beneficios significativos y, en algunos casos, puede llevar a un rendimiento subóptimo debido al sobreajuste.

Los resultados muestran que el modelo roberta-large-bne con 50 épocas y características lingüísticas (+LF) logra el mejor equilibrio entre precisión y generalización. Mientras que un mayor número de épocas (70) puede reducir levemente el MAE, también puede generar una ligera degradación en la estabilidad del modelo, como lo evidencia el comportamiento del R^2. Estos hallazgos refuerzan la importancia de ajustar cuidadosamente el número de ciclos de entrenamiento y la inclusión de características lingüísticas para optimizar el rendimiento del modelo.

Recomendaciones

Recomendamos la optimización del número de épocas por modelo y configuración, ya que, consideramos que entrenar por más de 50 épocas no mostró mejoras significativas y, en algunos casos, llevó al sobreajuste, se recomienda realizar experimentos adicionales para determinar de forma más precisa el número óptimo de ciclos de entrenamiento para diferentes arquitecturas y configuraciones del modelo. Esto podría incluir análisis adaptativos donde el entrenamiento se detenga automáticamente al alcanzar una convergencia en las métricas de validación.

Sugerimos la exploración de nuevas características lingüísticas, Aunque las características lingüísticas utilizadas (+LF) demostraron mejorar significativamente el rendimiento de los modelos, se sugiere explorar nuevas características relacionadas con semántica, sintaxis o complejidad cognitiva de los textos. Esto permitiría enriquecer aún más las representaciones de los modelos y podría contribuir a una mayor precisión en la predicción de la complejidad léxica.

Es necesario la ampliación del corpus y validación en diversos contextos, dado que el corpus utilizado (GovAIEc) se basa en textos públicos ecuatorianos, sería valioso ampliar el análisis a textos de otros países o contextos institucionales. Esto permitiría evaluar la capacidad de generalización de los modelos en diferentes dominios lingüísticos y validar la efectividad de las características lingüísticas en otros escenarios.

Referencias

1. Azucena, H., & Yanet, S. (2021). La educación inclusiva desde el marco legal educativo en el Ecuador. 6(3). Obtenido de https://doi.org/10.5281/ZENODO.5512949
2. Baeldung. (2024). Training and Validation Loss in Deep Learning. Obtenido de https://www.baeldung.com/cs/training-validation-loss-deep-learning
3. Beltagy, I., Peters, M., & Cohan, A. (2020). Longformer: The Long-Document Transformer. Obtenido de https://arxiv.org/pdf/2004.05150
4. Bender, E. (2023). Transformer Models: From Architecture to Impact in NLP. SADIO Electronic Journal. Obtenido de https://publicaciones.sadio.org.ar/index.php/EJS/article/download/465/393/.
5. Calero Sánchez, M., González González, J., Sánchez Berriel, I., Burillo-Putze, G., & Roda García, J. (2024). El Procesamiento de Lenguaje Natural en la revisión. Obtenido de https://www.reue.org/wp-content/uploads/2024/07/184-195.pdf?utm_source
6. Cesteros, J. (2023). Aproximaciones a la simplificaci´on l´exica mediante. Obtenido de https://apidspace.linhd.uned.es/server/api/core/bitstreams/24152488-5e9b-4185-904d-9e0b0346162b/content
7. Clark, K., Luong, M., Le, Q., & Maning, C. (2020). ELECTRA: PRE-TRAINING TEXT ENCODERS. Obtenido de https://arxiv.org/pdf/2003.10555
8. Cornell University. (26 de Febrero de 2021). Exploring Transformers in Natural Language Generation: GPT, BERT, and XLNet. Obtenido de https://arxiv.org/abs/2102.08036?utm_source
9. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Obtenido de https://arxiv.org/abs/1810.04805
10. Duchitanga, R., & León-Paredes, G. (21 de Mayo de 2023). An Approach to the Presumptive Detection of Road Incidents in Cuenca, Ecuador Using the Data from the Social Media Twitter and Spanish Natural Language Processing. Obtenido de https://link.springer.com/chapter/10.1007/978-3-031-32213-6_17
11. Face, H. (2020). Transformers. Obtenido de https://huggingface.co/docs/transformers/index

12. Geng, S., Lebret, R., & Aberer, K. (2023). Legal Transformer Models May Not Always Help. Obtenido de https://ugye-my.sharepoint.com/personal/kendrick_villotav_ug_edu_ec/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fkendrick%5Fvillotav%5Fug%5Fedu%5Fec%2FDocuments%2FBIBLIOGRAFIAS%2FBIBLIOGRAFIAS%2Ffiles%2F110%2FGeng%20et%20al%2E%20%2D%202021%20%2D%20Legal%20T
13. Grimmelikhujisen, S., & Welch, E. (8 de Junio de 2012). Developing and Testing a Theoretical Framework for Computer-Mediated Transparency of Local Governments. Obtenido de https://onlinelibrary.wiley.com/doi/10.1111/j.1540-6210.2011.02532.x
14. IBM TechXchange. (21 de Octubre de 2024). ¿Qué es el PLN (procesamiento del lenguaje natural)? Obtenido de https://www.ibm.com/es-es/topics/natural-language-processing?utm_source
15. Lenin, M. (2024). Aplicación de Modelos Transformers para Clasificar Textos en Idioma Español [Universidad Estatal Península de Santa Elena]. Obtenido de https://repositorio.upse.edu.ec/bitstream/46000/11875/1/UPSE-TTI-2024-0035.pdf?utm_source
16. Lu, Xu, & Wei. (2023). Understanding the effects of the textual complexity on government communication: Insights from China's online public service platform. Obtenido de https://www.sciencedirect.com/science/article/abs/pii/S0736585323000928
17. Ministerio de Telecomunicaciones y de la Sociedad de la Información. (2022). Obtenido de MINTEL-MINTEL: https://ugye-my.sharepoint.com/personal/kendrick_villotav_ug_edu_ec/_layouts/15/onedrive.aspx?ga=1&id=%2Fpersonal%2Fkendrick%5Fvillotav%5Fug%5Fedu%5Fec%2FDocuments%2FBIBLIOGRAFIAS%2FBIBLIOGRAFIAS%2Ffiles%2F109%2FMINTEL%2DMINTEL%2D2022%2D0034%2Epdf&parent=
18. Mo, Y., Qin, H., Dog, Y., Zhu, Z., & Li, Z. (24 de Abril de 2024). Large Language Model (LLM) AI text generation detection based on transformer deep learning algorithm. Obtenido de https://arxiv.org/abs/2405.06652
19. Moscoso Lozano, D. F., & Pacheco Fares, J. O. (2024). Trabajo de Titulación. Obtenido de https://ugye-my.sharepoint.com/personal/kendrick_villotav_ug_edu_ec/_layouts/15/onedrive.aspx?id=

%2Fpersonal%2Fkendrick%5Fvillotav%5Fug%5Fedu%5Fec%2FDocuments%2FBIBLIOGRAFIAS%2FBIBLIOGRAFIAS%2Ffiles%2F108%2FSistema%20de%20recomendaci%C3%B3n%20de%20cursos%

20. Nasimba, F. (2023). "Attention is all you need". Arquitectura Transformers: descripción y aplicaciones. Obtenido de https://dspace.umh.es/bitstream/11000/30273/1/TFG-Nasimba%20Tipan%2c%20Alexis%20Fabian.pdf
21. Olmos, M. (2021). PROCESAMIENTO DE LENGUAJE NATURAL APLICADO A LOS DISCURSOS DE JUAN DOMINGO PERÓN ENTRE 1943 Y 1955. Obtenido de https://ri.itba.edu.ar/server/api/core/bitstreams/b2074780-d8af-4326-beb2-2830b39ff56b/content
22. Ormaechea, L., Tsourakis, N., Schwab, D., Bouillon, P., & Lecouteux, B. (2023). Simple, Simpler and Beyond: A Fine-Tuning BERT-Based Approach to Enhance Sentence Complexity Assessment for Text Simplification.
23. Ortiz Zambrano, J., & Montejo-Ráez, A. (2017). A corpus of videos and transcriptions for research in the Reading Comprehension of University Students. Obtenido de https://doi.org/10.1007/978-3-030-32022-5_16
24. Ortiz Zambrano, J., & Varela Tapia, E. (2019). Reading Comprehension in University Texts: The Metrics of Lexical Complexity in Corpus Analysis in Spanish. Obtenido de https://doi.org/10.1007/978-3-030-12018-4_9
25. Ortiz-Zambrano, J., & Montejo-Raez, A. (2021). CLexIS2: A New Corpus for Complex Word Identification Research in Computing Studies. Obtenido de https://doi.org/10.26615/978-954-452-072-4_121
26. Ortiz-Zambrano, J., & Montejo-Raez, A. (2021). SINAI at SemEval-2021 Task 1: Complex word identification using Word-level features. Obtenido de https://ugye-my.sharepoint.com/personal/kendrick_villotav_ug_edu_ec/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fkendrick%5Fvillotav%5Fug%5Fedu%5Fec%2FDocuments%2FBIBLIOGRAFIAS%2FBIBLIOGRAFIAS%2Ffiles%2F20%2FOrtiz%2DZambrano%20y%20Montejo%2DR%C3%A1ez%20%2D%
27. Ortiz-Zambrano, J., Espin-Riofrio, C., & Montejo-Ráez, A. (2022). Transformers for Lexical Complexity Prediction in Spanish Language. Obtenido de https://doi.org/10.26342/2022-69-15

28. Ortiz-Zambrano, J., Espin-Riofrio, C., & Montejo-Ráez, A. (2023). Combining Transformer Embeddings with Linguistic Features for Complex Word Identification. Obtenido de https://doi.org/10.3390/electronics12010120
29. Ortiz-Zambrano, J., Espín-Riofrio, C., & Montejo-Ráez, A. (2023). LegalEc: Un nuevo corpus para la investigación de la identificación de palabras complejas en los estudios de Derecho en español ecuatoriano. Obtenido de https://doi.org/10.26342/2023-71-19
30. Ortiz-Zambrano, J., Espín-Riofrío, C., & Montejo-Ráez, A. (2024). Deep Encodings vs. Linguistic Features in Lexical Complexity Prediction. Obtenido de https://doi.org/10.1007/s00521-024-10662-9
31. Ortiz-Zambrano, J., Espín-Riofrío, C., & Montejo-Ráez, A. (2024). Enhancing Lexical Complexity Prediction Through Few-Shot Learning with GPT-3. Obtenido de https://ugye-my.sharepoint.com/personal/kendrick_villotav_ug_edu_ec/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fkendrick%5Fvillotav%5Fug%5Fedu%5Fec%2FDocuments%2FBIBLIOGRAFIAS%2FBIBLIOGRAFIAS%2Ffiles%2F24%2FOrtiz%2DZambrano%20et%20al%2E%20%2D%202024%20%2D%
32. Soneji, S., Hoesing, M., Koujalgi, S., & Dodge, J. (17 de Abril de 2024). Demystifying Legalese: An Automated Approach for Summarizing and Analyzing Overlaps in Privacy Policies and Terms of Service. Obtenido de https://arxiv.org/abs/2404.13087
33. Wold, S., Maehlum, P., & Hove, O. (1 de Abril de 2024). Estimating Lexical Complexity from Document-Level Distribution. Obtenido de https://arxiv.org/abs/2404.01196
34. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., & Moi, A. (2020). HuggingFace's Transformers: State-of-the-art Natural Language Processing. Obtenido de https://arxiv.org/abs/1910.03771
35. Yamada, I., Asai, A., Shindo, H., Takeda, H., & Matsumoto, Y. (2023). LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. Obtenido de https://arxiv.org/pdf/2010.01057