Polo del Conocimiento

Pol. Con. (Edición núm. 103) Vol. 10, No 2
Febrero 2025, pp. 861-884
ISSN: 2550 - 682X
DOI: https://doi.org/10.23857/pc.v10i2.8921

GPT-4 y sus modelos sucesores en la predicción de la complejidad léxica en textos públicos Ecuatorianos mediante Few-Shot Learning

GPT-4 and its successor models in the prediction of lexical complexity in Ecuadorian public texts using Few-Shot Learning

GPT-4 e seus modelos sucessores na previsão da complexidade lexical em textos públicos equatorianos usando Few-Shot Learning

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:** 03 de diciembre de 2024 ***Aceptado:** 25 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

Evaluar la complejidad léxica en documentos utilizando los modelos Generative Pre-trained Transformer (GPT-4, GPT-4o y GPT-4 Turbo) permitió analizar su impacto en la comprensión del lenguaje específicamente en textos estatales ecuatorianos en español. Se aplicó la técnica de *few-shot learning* en todos los modelos, tomando el conjunto de datos GovAIEc. El enfoque aplicado de la investigación es cuantitativo y descriptivo, con un diseño cuasi experimental. Los modelos se evalúan mediante métricas del error común: MAE, MSE, RMSE. El objetivo de esta investigación es evaluar el rendimiento del modelo GPT-4 y sus modelos sucesores en la predicción de la complejidad léxica en textos públicos ecuatorianos mediante *Few-Shot Learning*. Los resultados indican que el modelo GPT-4 obtiene un mayor rendimiento frente a sus sucesores con un MAE = 0.2464, mientras que GPT-4o presenta un MAE = 0.3889, y finalmente los resultados aplicando el modelo GPT-4 Turbo dieron un MAE = 0.2540.

Palabras clave: GPT-4; GPT-4 Turbo; GPT-4º; few-shot learning; predicción; complejidad léxica.

Abstract

Evaluating lexical complexity in documents using the Generative Pre-trained Transformer models (GPT-4, GPT-4o and GPT-4 Turbo) allowed us to analyze its impact on language comprehension specifically in Ecuadorian state texts in Spanish. The few-shot learning technique was applied to all models, taking the GovAIEc data set. The applied research approach is quantitative and descriptive, with a quasi-experimental design. The models are evaluated using common error metrics: MAE, MSE, RMSE. The objective of this research is to evaluate the performance of the GPT-4 model and its successor models in predicting lexical complexity in Ecuadorian public texts using Few-Shot Learning. The results indicate that the GPT-4 model obtains greater performance compared to its successors with a MAE = 0.2464, while GPT-4o presents a MAE = 0.3889, and finally the results applying the GPT-4 Turbo model gave a MAE = 0.2540.

Keywords: GPT-4; GPT-4 Turbo; GPT-4º; few-shot learning; prediction; lexical complexity.

Resumo

A avaliação da complexidade lexical em documentos utilizando os modelos Generative Pre-trained Transformer (GPT-4, GPT-4o e GPT-4 Turbo) permitiu analisar seu impacto na compreensão da

linguagem especificamente em textos estaduais equatorianos em espanhol. A técnica de aprendizagem de poucos disparos foi aplicada a todos os modelos, utilizando o conjunto de dados GovAIEc. A abordagem da pesquisa aplicada é quantitativa e descritiva, com desenho quase experimental. Os modelos são avaliados usando métricas de erro comuns: MAE, MSE, RMSE. O objetivo desta pesquisa é avaliar o desempenho do modelo GPT-4 e seus modelos sucessores na previsão da complexidade lexical em textos públicos equatorianos usando Few-Shot Learning. Os resultados indicam que o modelo GPT-4 obtém maior desempenho em relação aos seus sucessores com um MAE = 0,2464, enquanto o GPT-4o apresenta um MAE = 0,3889, e por fim os resultados aplicando o modelo GPT-4 Turbo deram um MAE = 0,2540.

Palavras-chave: GPT-4; GPT-4Turbo; GPT-4º; aprendizagem em poucas tentativas; previsão; complexidade lexical.

Introducción

La comprensión de textos en documentos públicos es esencial para la participación ciudadana y la transparencia en sociedades democráticas. El acceso a la información es un derecho fundamental que fortalece la democracia y promueve la rendición de cuentas, lo que es crucial para combatir la corrupción (Roque, 2024).

Sin embargo, el lenguaje complejo de estos documentos dificulta su comprensión, especialmente para personas con bajos niveles de alfabetización. En Ecuador, el censo de 2022 reportó 472,228 personas en condición de analfabetismo (El Universo, 2024), lo que limita su acceso a información clara y afecta su participación en decisiones de interés general, como establece el artículo 4 de la Ley Orgánica de Transparencia y acceso a la

Información Publica (LOTAIP, 2004).

Es necesario explorar herramientas que evalúen la complejidad del lenguaje en estos documentos. La predicción de la complejidad léxica mediante modelos de GPT-4 se presenta como una solución prometedora. Estudios han demostrado que estos modelos son efectivos para la simplificación léxica y pueden identificar términos complejos, mejorando así la comprensión de textos (Cesteros, 2023; Ortiz et al., 2024). Además, la complejidad textual se ve influenciada por factores culturales y el conocimiento previo del lector, lo que resalta la importancia de considerar tanto características lingüísticas como contextos culturales al evaluar la complejidad léxica (Ortiz y Montejo, 2020).

En este contexto, el objetivo de este estudio es evaluar el rendimiento del modelo GPT-4 y sus sucesores en la predicción de la complejidad léxica en textos públicos ecuatorianos mediante la técnica de *Few-Shot Learning*. Esta evaluación permitirá determinar la efectividad de estos modelos en la mejora de la comprensión de textos, contribuyendo así a la inclusión de personas con bajos niveles de alfabetización en el acceso a la información pública.

En estos últimos años, ha habido un considerable volumen de investigación en el área de la predicción de la complejidad textual y el procesamiento del lenguaje natural. A continuación, se mencionan investigaciones significativas que tratan estos aspectos:

Para el estudio de Ortiz et al. (2020) en donde se mostró la creación de un corpus multimodal que fusiona vídeos educativos y sus transcripciones, anotándolos con un nivel de complejidad del texto. Se concluyó que, el corpus "VYTEDU" se presentó como una herramienta útil para analizar la complejidad de los textos en contextos educativos, facilitando un estudio comparativo del discurso oral y escrito.

En este contexto, el estudio de Ortiz y Montejo (2021) se utilizaron datos de SemEval-2020 Task 1 para identificar palabras complejas en inglés, extrayendo características como longitud, frecuencia y embeddings preentrenados. Un clasificador Random Forest realiza la predicción, evaluando el rendimiento con el F1-Score. El modelo alcanzó un rendimiento competitivo en la identificación de palabras complejas, logrando una puntuación F1 de 0,85 en el conjunto de datos de evaluación, destacándose por la importancia de las características de frecuencia y embeddings.

Asimismo, el estudio de Ortiz et al. (2022) propone un enfoque para predecir la complejidad léxica en español utilizando modelos transformadores como BERT, XLM-RoBERTa y RoBERTa-large-BNE, entrenados con el corpus CLexIS2. Se combinan características manuales (frecuencia, longitud, categorías POS) con embeddings de modelos preentrenados. El modelo BERT ajustado alcanzó el rendimiento más destacado, con un MAE de 0.1592 y una correlación de Pearson de 0.9883. XLM-RoBERTa y RoBERTa-large-BNE igualmente mejoraron después del ajuste, aunque BERT resultó ser superior.

De la misma manera, la investigación de Ortiz et al. (2023) en donde se aplicó los sucesores davinci-002 y davinci-003 del Modelo GPT-3 para la clasificación de la complejidad de las palabras y se utilizó el enfoque de aprendizaje few-shot, donde se proporcionaron ejemplos limitados al modelo para ayudar en la clasificación. El mejor rendimiento fue del modelo davinci-003 con un MAE de 0.0882, en la predicción de la complejidad se observaron coincidencias y

discrepancias entre las categorías asignadas por GPT-3 y las del corpus Complex. Por último, se identificaron oportunidades para explorar nuevos modelos como Claude 2 y GPT-4 en la predicción de la complejidad léxica.

Además, el estudio de Ortiz et al. (2024) en donde se empleó el modelo GPT-4 Turbo centrándose en dos subtareas: Sub-task 2.1 para la identificación de términos y asignaciones de niveles de dificultad y Sub-task 2.2 para la generación de definiciones y explicaciones de términos considerados como difíciles. Los resultados indicaron que GPT-4 Turbo mostro un rendimiento notable en la evaluación de la complejidad léxica sin necesidad de entrenamiento adicional. Para los resultados de Sub-task 2.1 se logró una buena capacidad para identificar términos relevantes y su dificultad y para Sub-task 2.2 las definiciones y explicaciones generadas fueron efectivas en términos difíciles mejorando la comprensión de los textos científicos.

En este mismo año, Ortiz et al. (2024) en su estudio combinó características lingüísticas con codificaciones de modelos de lenguaje profundos (BERT, XLM-RoBERTa) en datatsets en inglés y español, en los cuales se aplicaron varios algoritmos de aprendizaje automático. El modelo en inglés logró un MAE de 0.0683, mejorando un 29.2%, en cambio para el modelo en español se alcanzó un MAE de 0.1323, con una mejora del 19.4%.

En el presente año, el estudio de Prada et al. (2025) se desarrolló un sistema de calificación automática de textos académicos utilizando técnicas de procesamiento de lenguaje natural y modelos de aprendizaje profundo. La investigación incluyó tres iteraciones: exploración de representaciones de texto con Word Embeddings y Transformers, entrenamiento directo con Transformers en un flujo unificado, y Fine-tuning del modelo RoBERTa evaluando clasificación y regresión. En la primera iteración, RoBERTa alcanzó un QWK de 0.7479 en regresión ordinal. En la segunda iteración, logró un QWK de 0.796, y en la tercera con el enfoque de clasificación obtuvo un QWK de 0.80238, teniendo dificultades en la categoría 6. Finalmente, con el enfoque de regresión mejoró a un QWK de 0.81639, clasificando correctamente algunos textos de la categoría 6.

Para Taboada (2024) realizó una revisión histórica de la evolución del PLN en ciencias sociales, una guía práctica con pasos para su aplicación y un análisis de los desafíos que enfrentan estas disciplinas al implementar PLN. Se identificaron herramientas y software accesibles para investigadores sociales, como R, Python, Orange Data Mining y RapidMiner. Las fuentes de datos se clasifican en analógicos, transcritos y digitales, subrayando la importancia de la digitalización.

También se presentan técnicas de análisis como la tokenización, eliminación de las palabras vacías y algoritmos de aprendizaje automático.

Además, Godínez y Rosas (2024a) realizaron un estudio cuantitativo y cualitativo con 12 estudiantes universitarios para analizar la relación entre perfil lingüístico, autoeficacia y complejidad textual en la producción escrita en español, utilizando cuestionarios y herramientas de análisis de texto. Los hablantes con la lengua de herencia tuvieron una mayor facilidad temática, mientras que los no hablantes enfrentaron desafíos gramaticales. La autoeficacia se correlacionó con la complejidad textual, sugiriendo estrategias pedagógicas para atender las necesidades de cada perfil lingüístico.

Asimismo, la investigación de Salgado y Trujillo (2024) en donde se realizó una investigación de la literatura sobre el análisis de los sentimientos en datos de redes sociales utilizando técnicas de procesamiento de LPN y ML, y una búsqueda en BD académicas claves, con el objetivo de identificar y analizar aplicaciones, desafíos y tendencias emergentes en estas tecnologías. Se destaco la necesidad de adaptar continuamente los modelos a los cambios en la dinámica lingüística y culturas, con respecto a las fuentes de datos, se identificaron como primordiales las redes sociales, representando hasta un 85% de las interacciones analizadas.

Finalmente, el estudio de Emanuel et al. (2024) en donde se comparan algoritmos de machine learning para el LPN en tareas de clasificación y análisis de texto, se evaluaron cuatro algoritmos; regresión lógica, arboles de decisión, máquinas de vectores de soporte (SVM) y redes neuronales, se emplearon métricas estándar de evaluación (precisión, exhaustividad, puntuación F1 y exactitud) para comparar el rendimiento de los algoritmos en el conjunto de datos de tweets etiquetados. El mejor rendimiento lo tuvo Random Forest con una precisión del 98.17% y una puntuación de F1 de 0.9813, con respecto a la Regresión Logística su precisión fue del 87.74% y un F1 de 0.885, para el Árbol de Decisión la precisión fue del 96.22% y su F1 de 0.9606, y por último Naive Bayes con el menor rendimiento, con una precisión del 71.75% y una puntuación F1 de 0.7755.

Por los motivos anteriormente expuestos, el objetivo de la investigación se centró en GPT-4 y sus modelos sucesores en la predicción de la complejidad léxica en textos públicos ecuatorianos mediante Few-Shot Learning, la cual respondió a la interrogante: ¿Cómo pueden GPT-4 y sus modelos sucesores, mediante el uso de Few-Shot Learning, predecir eficazmente la complejidad léxica en textos públicos ecuatorianos para mejorar la comprensión del público?

Materiales y métodos

La metodología de investigación empleada en este estudio sobre la predicción de la complejidad léxica de textos en documentos públicos, utilizando GPT-4 y modelos sucesores, ha sido diseñada para abordar de manera integral el objetivo planteado. Esta investigación es de carácter aplicado, ya que busca utilizar conocimientos existentes en procesamiento de lenguaje natural y análisis de modelos de inteligencia artificial para resolver un problema práctico: evaluar la complejidad léxica de documentos públicos y mejorar su accesibilidad para los ciudadanos.

Según Godínez y Rosas (2024) esta modalidad incluye cualquier esfuerzo sistemático y socializado para resolver problemas o intervenir en situaciones, abarcando tanto la innovación técnica como la investigación científica. De este modo, la investigación vincula la teoría con la práctica, generando un impacto directo en la comprensión y uso de textos administrativos emitidos por las instituciones públicas de Guayaquil. El tipo de investigación es cuantitativa y descriptiva, con un enfoque cuasi-experimental. Se considera cuantitativa porque busca medir y analizar numéricamente la complejidad léxica de los textos, así como evaluar el desempeño de los modelos GPT-4 y sus sucesores a través de métricas específicas como precisión, fluidez y coherencia.

El enfoque cuasi-experimental se utiliza para examinar las relaciones entre una o más variables independientes y la variable dependiente o de respuesta (Bono, 2012). Este enfoque es adecuado, ya que se realizarán pruebas controladas con un corpus específico de documentos, simulando escenarios reales para evaluar los modelos en condiciones controladas.

La unidad de estudio corresponde a cada registro individual dentro del dataset *GovAIEc* que contiene un total de 7,813 registros, el cual está compuesto por notificaciones e instrucciones relacionadas con trámites legales, en general oraciones seleccionadas de documentos de las entidades públicas gubernamentales que pertenecen a Ecuador, específicamente de la ciudad de Guayaquil, las cuales son: CNT, SRI, CNE, Municipio y ATM.

Cada registro tiene los siguientes campos:

- id: Identificador único para cada registro.
- corpus: Institución pública gubernamental (Fuente).
- sentence: Oración que contiene la palabra etiquetada como compleja.
- token: Palabra identificada como compleja por etiquetadores.
- complexity: Valor numérico que representa la complejidad asignada por los etiquetadores.

Tabla 1

Dataset GovAIEc

id	corpus	Sentence	token	complexity
6075	Municipio - Tramites - TEXTO 0060 TRAMITES EN LA BIBLIOTECA MUNICIPAL.txt	TRAMITES EN LA BIBLIOTECA MUNICIPAL Si en el stock existe el libro pedido, se emite Comprobante de la Publicación para que el comprador proceda a …	comprobante	0,333333333
7719	SRI - Tramites - TEXTO 0130 REQUERIMIENTOS Y JUSTIFCACIONES DEL PROCESO INCONSITENCIAS.txt	REQUERIMIENTOS Y JUSTIFCACIONES DEL PROCESO INCONSITENCIAS Si el Servicio de Rentas Internas detectare inconsistencias en las declaraciones o en los anexos que presente el contribuyente, siempre que no generen …	sustitutivo	0,666666667
2734	CNE - Tramites - TEXTO 0091 REGLAMENTO PARA CONFORMACION DE ALIANZAS ELECTORALES.txt	REGLAMENTO PARA CONFORMACION DE ALIANZAS ELECTORALES la Constitución de la República dispone en el artículo 112, que los partidos y movimientos …	militantes	1

Nota: *La tabla muestra un conjunto de registros extraídos de documentos legales de diversas instituciones públicas de Guayaquil, con el objetivo de identificar palabras complejas dentro de los textos relacionados con trámites legales. Estos datos se utilizarán para calcular características lingüísticas y entrenar el modelo*

La escala de complejidad tiene los siguientes niveles:

- Moderately difficult: Rango de complejidad entre 0 y 0.3333. Las palabras u oraciones en este rango son algo complejas, pero se pueden entender en su contexto.
- Difficult: Desde 0.3334 a 0.6666. Las palabras u oraciones en este nivel son bastante complejas y pueden necesitar un mayor nivel de comprensión o conocimientos técnicos.
- Very difficult: Desde 0.6667 a 1. Las palabras u oraciones en esta categoría son muy complejas, lo que puede hacer que sean difíciles de entender.

Tabla 2

Escala de complejidad

Etiqueta	Rango
moderately difficult	(0, 0.3333)
difficult	(0.3334, 0.6666)
very difficult	(0.6667, 1)

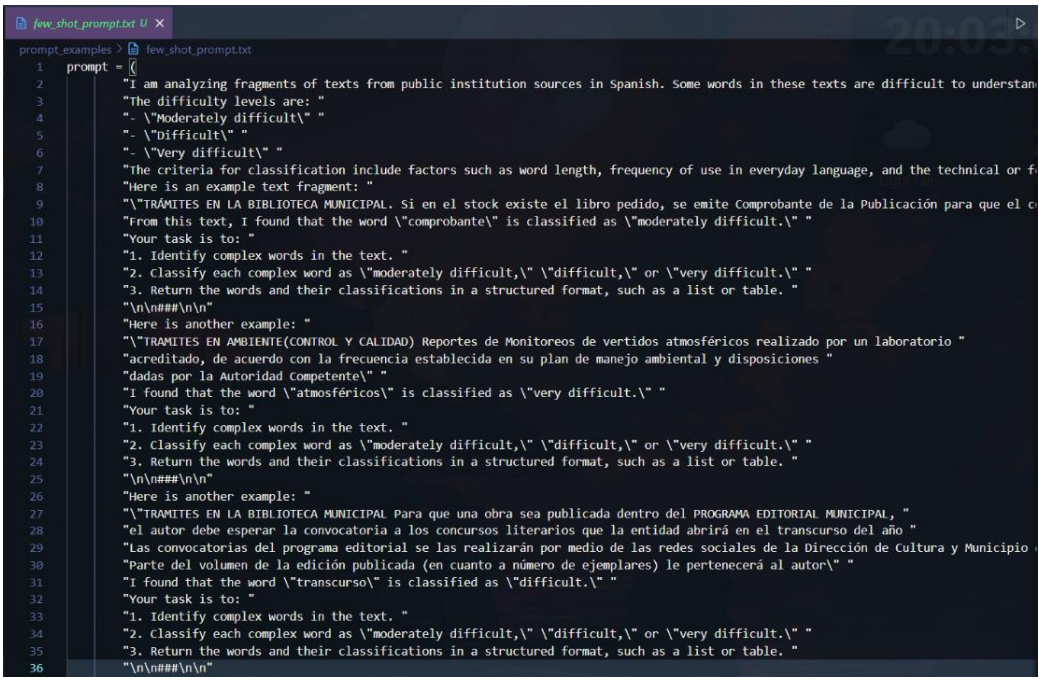
Nota: *Esta escala se utiliza para evaluar el nivel de complejidad de las palabras en el prompt, aplicando la técnica de few-shot learning en los modelos GPT-4. Además, esta escala fue empleada por los anotadores para asignar un valor a cada palabra identificada como compleja (token), el cual se registró en la columna complexity*

En el tratamiento de los datos para la identificación de palabras complejas del datatset *GovAIEc*, se utilizó un enfoque basado en el modelo de lenguaje GPT-4, aplicando la técnica de *few-shot learning*. Este enfoque se empleó para predecir la complejidad textual de las palabras en función de su contexto dentro de las oraciones extraídas de documentos legales gubernamentales. El proceso de tratamiento de los datos consistió en los siguientes pasos:

1. *Lectura y preprocesamiento de datos:* Los datos se leyeron desde el archivo GovAIEc.xlsx, que contenía oraciones y palabras identificadas como complejas en documentos legales de instituciones públicas. Las columnas relevantes para el análisis fueron *id, sentence, token* y *complexity*, las cuales se utilizaron para calcular las predicciones del modelo y mostrar los resultados.
2. *Generación de predicciones de complejidad:* Para clasificar las palabras según su complejidad, se emplearon modelos basados en la arquitectura Transformer, específicamente GPT-4, GPT-4 Turbo y GPT-4o. Estos modelos se aplicaron para identificar el nivel de complejidad de las palabras dentro de las oraciones de los documentos legales. Se utilizó el enfoque de *few-shot learning*, ya que permite a los modelos aprender con pocos ejemplos proporcionados en el *prompt*. En este caso, se incluyeron un total de 20 ejemplos, con el objetivo de mejorar la precisión de las predicciones. El modelo GPT-4 y sus sucesores clasificaron cada palabra identificada como compleja en una de las tres categorías de complejidad: *Moderately difficult*, *Difficult* y *Very difficult*.
3. *Formato del prompt de predicción:* El modelo recibió como entrada un *prompt* específico que estableció el contexto de la tarea. A continuación, se muestra el formato del *prompt* utilizado para realizar las predicciones de complejidad mediante *few-shot learning*:

Figura 1

Prompt Few-Shot Learning



Nota: *El fragmento de código muestra parte del prompt utilizado para clasificar palabras complejas en tres categorías de dificultad basándose en el contexto de la oración. Esta clasificación se realiza mediante Few-Shot Learning*

1. *Evaluación y resultados:* Una vez que el modelo genera las predicciones de complejidad, los resultados se almacenan en un archivo Excel para su posterior análisis.
2. *Resumen final:* Tras obtener todos los resultados, se genera un archivo Excel como resumen final, que incluye las métricas calculadas para cada modelo y la técnica aplicada. Este archivo se guarda con el nombre *resumen_metricas.xlsx.*

Métricas del error común

Las métricas aplicadas en esta investigación corresponden a métricas de evaluación utilizadas para medir la precisión y calidad de las predicciones realizadas por el modelo GPT-4 y sus sucesores, GPT-4 Turbo y GPT-4o. A continuación, se detalla cada una de ellas:

MAE (Mean Absolute Error): Puede utilizarse si los valores atípicos representan partes corruptas de los datos. (Chicco et al., 2021)

Ecuación 1 Mean Absolute Error

$$MAE = \frac{1}{m}\sum_{i=1}^{m} | (X_i - Y_i)^2 |$$

Tomado de: (Chicco et al., 2021)

MSE (Mean Squared Error): Se emplea el error cuadrático medio (MSE) para evaluar la exactitud de un modelo de predicción, ya que mide la diferencia entre los valores detectados y los valores estimados por el modelo. También puede utilizarse para detectar valores atípicos, ya que, debido a la norma L2, el MSE otorga un mayor peso a estos puntos. Si el modelo produce una única predicción muy mala, la parte cuadrática de la función incrementa el error. (Chicco et al., 2021)

Ecuación 2 Mean Squared Error

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2$$

Tomado de: (Chicco et al., 2021)

RMSE (Root Mean Squared Error): MSE y RMSE mantienen una relación monotónica, es decir, por medio de la raíz cuadrada. Una ordenación de los modelos de regresión basada en el MSE será idéntica a una ordenación de los modelos basada en el RMSE. (Chicco et al., 2021)

Ecuación 3 Root Mean Squared Error

$$RMSE = \sqrt{MSE}$$

Tomado de: (Reyes, 2024)

R² (R-squared): "El coeficiente de determinación puede interpretarse como la proporción de la varianza de la variable dependiente que puede predecirse a partir de las variables independientes." (Chicco et al., 2021, p. 5)

Ecuación 4 R-squared

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (X_i - \bar{Y}_i)^2}{\sum_{i=1}^{m} (\bar{Y} - \bar{Y}_i)^2}$$

Tomado de: (Chicco et al., 2021)

Se evalúa el modelo GPT-4 y sus sucesores utilizando métricas que permiten determinar su rendimiento en la identificación de la complejidad léxica. Estas métricas se calculan comparando los valores reales de complejidad con los valores predichos por el modelo. Esto es fundamental para entender la efectividad del modelo y realizar ajustes si es necesario. Además, la evaluación permite comparar el rendimiento de GPT-4 con sus sucesores, identificando cuál es el más efectivo en la tarea de identificación de palabras complejas.

Para evaluar la diferencia relativa entre los modelos en términos de porcentaje, se utilizará la fórmula de diferencia porcentual, mencionada por (Yuen et al., 2024), la cual se expresa de la siguiente manera:

Ecuación 5 Diferencia porcentual

$$\% = \left(\frac{x_1 - x_2}{x_2}\right) \times 100$$

En donde:

x_1 representa el valor final.

x_2 representa el valor inicial.

Resultados y discusión

Aplicación del modelo GPT-4 y sus sucesores GPT-4 Turbo y GPT-4o

A continuación, se presentan tablas que contienen un extracto de los resultados obtenidos al aplicar la técnica de *few-shot learning* a cada modelo. Se han seleccionado las palabras clasificadas como difíciles en la columna "*up to 5 difficult terms retrieved by GPT-4* " por el modelo GPT-4 y sus sucesores para cada una de las instituciones públicas gubernamentales. Es decir, el modelo evalúa la dificultad de comprensión del texto y selecciona hasta 5 palabras que considera difíciles de entender para una persona promedio. Estas palabras son las que se presentan en las tablas.

En primera instancia, se muestran las palabras complejas identificadas por el modelo GPT-4 utilizando la técnica de *few-shot learning*.

Tabla 3

Palabras complejas GPT-4 / Few-Shot Learning

Institución	Palabras complejas
Municipio	Ocurre, Registrada, Convenio, Débito, Vigente

CNE	Calificará, Organizaciones, Solicitudes, Contratación, Publicidad
SRI	Requerimientos, Justificaciones, Proceso, Contribuyente, Tributaria
CNT	Institución, Irrevocable, Verídica, Ostentar, Cesionaria
ATM	Tacómetro, Sonómetro, Frenómetro, Luxómetro, Regloscopio

Nota: *La tabla presenta algunas palabras que el modelo GPT-4 clasificó como complejas*

Para GPT-4 Turbo, se presentan las siguientes palabras:

Tabla 4

Palabras complejas GPT-4 Turbo / Few-Shot Learning

Institución	Palabras complejas
Municipio	Comisario, Municipal, Compromiso, Determinado, Subsanar
CNE	Ratifico, Autorización, Expiración, Notificar, Cancelar
SRI	Requerimientos, Justificaciones, Resolución, Carácter, Información
CNT	Telecomunicaciones, Tranquilidad, Autorización, Determinadas, Propaguen
ATM	Estupefacientes, Psicotrópicas, Terapéutico, Rehabilitación, Infracciones

Nota: *La tabla presenta algunas palabras que el modelo GPT-4 Turbo clasificó como complejas*

Por último, para el modelo GPT-4o, se identificaron las siguientes palabras como difíciles:

Tabla 5

Palabras complejas GPT-4o / Few-Shot Learning

Institución	Palabras complejas
Municipio	Préstamo, Consultado, Instalaciones, Reteniendo, Identificación
CNE	Desafiliación, Renuncia, Escaneado, Nulidad, Seguimiento
SRI	Conformación, Organizaciones, Discriminatorias, Funcionamiento, Garantizara
CNT	Prestación, Establecidos, Reparaciones, Situaciones, Fortuito
ATM	Inminente, Seguridad, Ocupantes, Obligación, Comprobar

Nota: *La tabla presenta algunas palabras que el modelo GPT-4o clasificó como complejas*

Resultados de las métricas de evaluación

A continuación, se presentan tablas que resumen los resultados de las predicciones de los modelos GPT-4, GPT-4 Turbo y GPT-4o en la identificación de palabras complejas, utilizando la técnica de *few-shot learning*. Cada fila corresponde a un registro del corpus, donde se comparan las predicciones del modelo con los valores reales de complejidad.

Las columnas de las tablas incluyen:

- id: Identificador único del registro.
- token: Palabra identificada como compleja.
- Respuesta GPT-4: Categoría de complejidad predicha por el modelo.
- Rango GPT-4: Rango de complejidad asignado por el modelo.
- Complejidad GPT-4: Valor numérico de la complejidad predicha.
- complexity: Valor numérico de la complejidad real.
- escala: Categoría de complejidad real.
- comparación: Indica si la predicción coincide con el valor real (Sí/No).

Tabla 6

Predicciones del modelo GPT-4

id	token	Respuesta GPT-4	Rango GPT-4	Complejidad GPT-4	complexity	escala	comparación
6075	comprobante	difficult	(0.3334, 0.6666)	0,5	0,333333333	moderately difficult	No
6076	recaudadoras	difficult	(0.3334, 0.6666)	0,5	1	very difficult	No
6077	stock	moderately difficult	(0, 0.3333)	0,16665	1	very difficult	No
6093	señalando	difficult	(0.3334, 0.6666)	0,5	0,333333333	moderately difficult	No
6094	bibliográfico	moderately difficult	(0, 0.3333)	0,16665	0,333333333	moderately difficult	Si
6095	autorización	moderately difficult	(0, 0.3333)	0,16665	0,666667	difficult	No
6096	donarse	moderately difficult	(0, 0.3333)	0,16665	0,333333	moderately difficult	Si
6059	devolución	difficult	(0.3334, 0.6666)	0,5	1	very difficult	No

6060	instalaciones	moderately difficult	(0, 0.3333)	0,16665	0,333333	moderately difficult	Si
6061	identificación	difficult	(0.3334, 0.6666)	0,5	0,333333	moderately difficult	No

Nota: *La tabla muestra un extracto de los resultados de las métricas de evaluación para el modelo GPT-4*

Tabla 7

Predicciones del modelo GPT-4 Turbo

id	token	Respuesta GPT-4	Rango GPT-4	Complejidad GPT-4	complexity	escala	comparación
6075	comprobante	moderately difficult	(0, 0.3333)	0,16665	0,333333	moderately difficult	Si
6076	recaudadoras	difficult	(0.3334, 0.6666)	0,5	1	very difficult	No
6077	stock	moderately difficult	(0, 0.3333)	0,16665	1	very difficult	No
6093	señalando	moderately difficult	(0, 0.3333)	0,16665	0,333333	moderately difficult	Si
6094	bibliográfico	difficult	(0.3334, 0.6666)	0,5	0,333333	moderately difficult	No
6095	autorización	difficult	(0.3334, 0.6666)	0,5	0,666667	difficult	Si
6096	donarse	moderately difficult	(0, 0.3333)	0,16665	0,333333	moderately difficult	Si
6059	devolución	moderately difficult	(0, 0.3333)	0,16665	1	very difficult	No
6060	instalaciones	moderately difficult	(0, 0.3333)	0,16665	0,333333	moderately difficult	Si

6061	identificación	difficult	(0.3334, 0.6666)	0,5	0,333333	moderately difficult	No

Nota: *La tabla muestra un extracto de los resultados de las métricas de evaluación para el modelo GPT-4 Turbo*

Tabla 8

Predicciones del modelo GPT-4o

id	token	Respuesta GPT-4	Rango GPT-4	Complejidad GPT-4	complexity	escala	comparación
6075	comprobante	moderately difficult	(0, 0.3333)	0,16665	0,333333	moderately difficult	Si
6076	recaudadoras	difficult	(0.3334, 0.6666)	0,5	1	very difficult	No
6077	stock	moderately difficult	(0, 0.3333)	0,16665	1	very difficult	No
6093	señalando	moderately difficult	(0, 0.3333)	0,16665	0,333333	moderately difficult	Si
6094	bibliográfico	difficult	(0.3334, 0.6666)	0,5	0,333333	moderately difficult	No
6095	autorización	moderately difficult	(0, 0.3333)	0,16665	0,666667	difficult	No
6096	donarse	moderately difficult	(0, 0.3333)	0,16665	0,333333	moderately difficult	Si
6059	devolución	moderately difficult	(0, 0.3333)	0,16665	1	very difficult	No
6060	instalaciones	moderately difficult	(0, 0.3333)	0,16665	0,333333	moderately difficult	Si
6061	identificación	moderately difficult	(0, 0.3333)	0,16665	0,333333	moderately difficult	Si

Nota: *La tabla muestra un extracto de los resultados de las métricas de evaluación para el modelo GPT-4º*

A continuación, se presentan los resultados obtenidos a manera de resumen de la aplicación de los sucesores GPT-4 Turbo, GPT-4o y el modelo GPT-4 en la identificación de palabras complejas.

Tabla 9

Resultado de los sucesores

Modelo	MAE	MSE	RMSE	R2	Coincidencia
GPT-4_few_shot	0.2464	0.0888	0.2980	-0.5935	37,59 %
GPT-4Turbo_few_shot	0.2540	0.0915	0.3025	-0.6420	51,67 %
GPT-4o_few_shot	0.2593	0.0963	0.3103	-0.7278	53,44 %

Nota: *Los resultados presentados en esta tabla muestran la precisión del modelo GPT-4 y sus sucesores GPT-4 Turbo y GPT-4o en la predicción de palabras complejas, evaluadas a través de métricas estadísticas como MAE, MSE, RMSE, R^2 y el porcentaje de coincidencia permitiendo una comparación cuantitativa de su rendimiento en la tarea de simplificación del lenguaje*

Evaluación de la predicción de la complejidad léxica

Para la evaluación de los resultados, se realizó una comparativa entre los diferentes modelos. A cada modelo se le proporcionaron 20 ejemplos en el prompt para entrenarlo en la clasificación de palabras según su nivel de dificultad: 1. *Moderately difficult*, 2. *Difficult* y 3. *Very difficult* (incluyendo 2 ejemplos de esta última categoría), lo que suma un total de 4 ejemplos por cada institución gubernamental pública.

Como se observa en la tabla de resultados, la diferencia del MAE entre los modelos es pequeña pero significativa. Según (Tatachar, 2021) el MAE es una métrica que proporciona el promedio de la diferencia absoluta, lo que la hace menos sensible a valores atípicos. Esto permite evaluar la precisión de los modelos en el contexto de la identificación de palabras complejas, ya que ofrece una mejor comprensión del error promedio de las predicciones sobre la complejidad léxica.

Con el objetivo de entender la diferencia del MAE entre los modelos, se calculó la diferencia porcentual utilizando la *Ecuación 5* Diferencia porcentual. Sabiendo que el modelo GPT-4 tuvo un MAE de 0.24641, el GPT-4 Turbo un MAE de 0.254051 y el GPT-4o un MAE de 0.259385, se obtuvieron los siguientes resultados:

Entre GPT-4 y GPT-4 Turbo:

$$\% = \left(\frac{0.254051 - 0{,}24641}{0{,}24641}\right) \times 100 \approx 3.10\%$$

Entre GPT-4 y GPT-4o:

$$\% = \left(\frac{0.259385 - 0{,}24641}{0{,}24641}\right) \times 100 \approx 5.27\%$$

Como resultado, se observa que el incremento porcentual en el MAE entre los modelos GPT-4 y GPT-4 Turbo es de aproximadamente 3.10%, mientras que entre los modelos GPT-4 y GPT-4o es de 5.27%. Esto indica que el GPT-4 realiza predicciones más precisas en promedio en la identificación de palabras complejas en comparación con el GPT-4 Turbo. En el caso del GPT-4o, la diferencia en precisión es más notable, lo que podría sugerir mejoras o variaciones en su

entrenamiento. En resumen, tanto el GPT-4 Turbo como el GPT-4o tienen un MAE mayor que el GPT-4, pero con incrementos relativamente pequeños, siendo el GPT-4o el que presenta el mayor incremento.

Con respecto al MSE, (Tatachar, 2021) menciona que esta métrica representa la diferencia al cuadrado entre los valores reales y los predichos. Es decir, el MSE indica cuán cerca está la línea de mejor ajuste de un conjunto de puntos. En este caso, el GPT-4 tiene un valor de 0.088819, el GPT-4 Turbo de 0.091523 y el GPT-4o de 0.096301, con diferencias porcentuales de:

Entre GPT-4 y GPT-4 Turbo:

$$\% = \left(\frac{0.091523 - 0.088819}{0.088819}\right) \times 100 \approx 3.04\%$$

Entre GPT-4 y GPT-4o:

$$\% = \left(\frac{0.096301 - 0.088819}{0.088819}\right) \times 100 \approx 8.43\%$$

Esto indica que la diferencia entre los valores reales y los predichos entre los modelos GPT-4 y GPT-4 Turbo es del 3.21%, mientras que entre los modelos GPT-4 y GPT-4o es del 8.43%. Esto significa que el MAE del GPT-4 Turbo es un 3.04% mayor que el del GPT-4, y el MAE del GPT-4o es un 8.13% mayor que el del GPT-4. En otras palabras, el GPT-4 tiene un rendimiento ligeramente superior en términos de precisión y un ajuste más cercano a los valores reales en comparación con el GPT-4o y el GPT-4 Turbo.

Para el caso con el RMSE, según el autor (Hodson, 2022) tomar la raíz no afecta los rangos relativos de los modelos, pero produce una métrica con las mismas unidades que (y), lo que representa convenientemente el error típico o estándar para errores distribuidos normalmente. En este contexto, el modelo GPT-4 tiene un RMSE de 0.298025, el GPT-4 Turbo de 0.302527 y el GPT-4o de 0.310325, con diferencias porcentuales de:

Entre GPT-4 y GPT-4 Turbo:

$$\% = \left(\frac{0.302527 - 0.298025}{0.298025}\right) \times 100 \approx 1.51\%$$

Entre GPT-4 y GPT-4o:

$$\% = \left(\frac{0.310325 - 0.298025}{0.298025}\right) \times 100 \approx 4.13\%$$

La diferencia porcentual entre los modelos GPT-4 y GPT-4 Turbo es del 1.51%, lo que indica que el GPT-4 tiene un rendimiento ligeramente superior al del GPT-4 Turbo. Por otro lado, la diferencia

entre el GPT-4 y el GPT-4o es del 4.13%, lo que sugiere que el GPT-4o realiza predicciones menos precisas en comparación con el GPT-4. Según (Reyes, 2024) el R^2 mide la proporción de varianza explicada por las variables independientes en un sentido estadístico. Sin embargo, esta medida no refleja necesariamente la importancia de las variables en el modelo. Un R^2 de 1.00 no implica que se haya encontrado una explicación válida para el fenómeno estudiado. En este caso, el modelo GPT-4 tiene un valor de R^2 de -0.593562, el GPT-4 Turbo de -0.64207 y el GPT-4o de -0.727808, con diferencias porcentuales de:

Entre GPT-4 y GPT-4 Turbo:

$$\% = \left(\frac{-0.64207 - (-0.593562)}{0.593562}\right) \times 100 \approx 8.18\%$$

Entre GPT-4 y GPT-4o:

$$\% = \left(\frac{-0.727808 - (-0.593562)}{0.593562}\right) \times 100 \approx 22.61\%$$

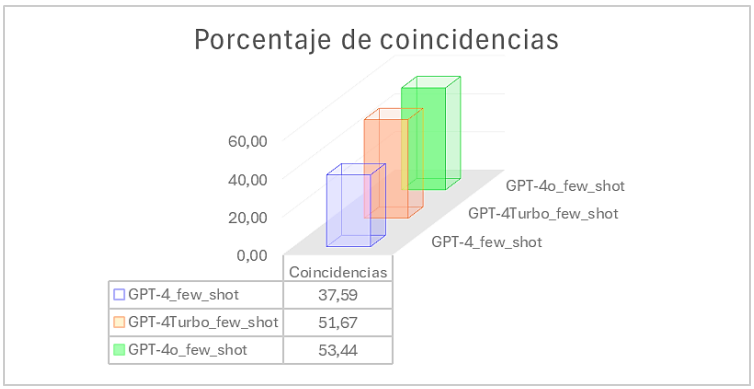
Los valores en ambas comparaciones sugieren que los modelos tienen predicciones deficientes y, aunque no están ajustando bien los datos, el modelo GPT-4 Turbo es un 8.18% peor en términos de ajuste en comparación con el modelo GPT-4. Además, la diferencia del 22.61% indica que el GPT-4o tiene un rendimiento significativamente inferior al del GPT-4. Por último, en cuanto a la coincidencia entre los modelos, el GPT-4o presenta un 53.44%, lo que sugiere que tiene una mayor precisión en sus predicciones en comparación con el GPT-4 y el GPT-4 Turbo. Esto indica que el GPT-4o es ligeramente más preciso al predecir valores cercanos a los reales en relación con los otros dos modelos.

Evaluación del porcentaje de coincidencia

En cuanto a los resultados de porcentaje de coincidencias entre los modelos GPT-4, GPT-4 Turbo y GPT-4o con *few-shot learning*, se presentan en el siguiente gráfico donde se puede visualizar el resultado de cada ejecución:

Figura 2

Porcentaje de coincidencias de los modelos



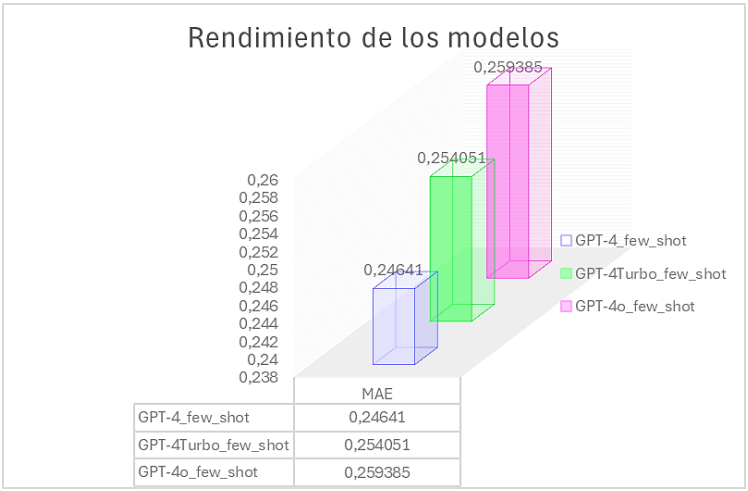
Nota: *Resultados de los porcentajes de coincidencias de las ejecuciones realizadas en el modelo GPT-4 y sus sucesores, utilizando la técnica de few-shot learning*

Como se puede visualizar en el gráfico anterior, GPT-4o utilizando *few-shot learning* es el modelo con más coincidencias con el corpus a diferencia de los otros modelos, con el 53.44% de coincidencias. Aunque las otras ejecuciones no están tan alejadas de este resultado, presentan un desempeño ligeramente inferior, seguido de GPT-4 Turbo que alcanzó un 51.67% de coincidencias, mientras que GPT-4 tiene el porcentaje de coincidencias más bajo con 37.59% de coincidencias.

Rendimiento de los modelos en función del MAE

Figura 3

Rendimiento de los modelos en función del MAE



Nota: *Resultados del rendimiento del modelo GPT-4 y sus sucesores en función al MAE con la técnica de few-shot learning*

Con respecto al rendimiento de los modelos en función del MAE, el modelo GPT-4_few_shot (0.2464) presenta el MAE más bajo, lo que indica una mayor precisión en la predicción de la complejidad léxica. El modelo GPT-4Turbo_few_shot, con un MAE de 0.2540, se sitúa en un punto intermedio, superando al GPT-4o_few_shot (0.2593) pero sin alcanzar la precisión del GPT-4. Esto sugiere que, aunque el modelo Turbo tiene un desempeño aceptable, aún no iguala al GPT-4 en términos de exactitud. Por otro lado, el GPT-4o_few_shot, con el MAE más alto, muestra un rendimiento inferior, evidenciando que la técnica *few-shot learing* no siempre garantiza mejoras en la precisión.

En resumen, los resultados se pueden sintetizar de la siguiente manera:

- Mejor modelo en precisión (MAE): GPT-4_few_shot (MAE: 0.2464).
- Mejor modelo en coincidencia: GPT-4o_few_shot (53.44%).
- Peor modelo en precisión (MAE): GPT-4o_few_shot (MAE: 0.2593).
- Peor modelo en ajuste (R^2): GPT-4o_few_shot (R^2: -0.7278).

Aunque el GPT-4_few_shot es el más preciso según el MAE, el GPT-4o_few_shot destaca en coincidencia, lo que sugiere un mayor acierto en sus predicciones. Sin embargo, todos los modelos presentan un R^2 negativo, indicando una mala explicación de la variabilidad de los datos y posible sobreajuste.

Conclusiones

Se observa una diferencia significativa en el desempeño de los modelos GPT-4, GPT-4o y GPT-4 Turbo al aplicar la técnica de *few-shot learning*, siendo el modelo GPT-4 el que presenta el mayor porcentaje de coincidencias con el corpus de referencia. En términos del error absoluto medio (Mean Absolute Error, MAE), los resultados indican que el modelo GPT-4 con *few-shot learning* obtuvo el MAE más bajo (0.2464), lo que sugiere que sus predicciones son más cercanas a los valores reales en comparación con los demás modelos evaluados. Estos hallazgos evidencian que la técnica *few-shot learning* mejora significativamente la precisión de los modelos en la tarea de predicción de complejidad léxica.

Finalmente, si bien el desarrollo de una herramienta basada en GPT-4 para la simplificación de documentos públicos resulta viable, su implementación efectiva requiere la combinación de modelos de IA con revisión humana. Este enfoque híbrido es fundamental para garantizar que la información generada sea accesible sin comprometer su precisión y relevancia.

Recomendaciones

Realizar ejecuciones variando el *prompt* original para explorar cómo estas modificaciones afectan el rendimiento de los modelos. Esto permitirá identificar qué formulaciones generan mejores resultados en términos de coincidencia y precisión.

Continuar evaluando el rendimiento de los modelos sucesores actuales de GPT en relación con la complejidad léxica de textos provenientes de instituciones públicas. Comparar estos resultados con textos de otros dominios para obtener una visión más amplia de su desempeño.

Finalmente, se recomienda ejecutar el corpus con otros modelos basados en la misma arquitectura Transformer. Esto facilitará el análisis del comportamiento de diferentes LLMs (Large Language Models) y permitirá identificar características que puedan mejorar la precisión y la relevancia de las predicciones.

Referencias

1. Bono Cabré, R. (2012). Diseños cuasi-experimentales y longitudinales. OMADO (Objectes i MAterials DOcents). https://diposit.ub.edu/dspace/handle/2445/30783
2. Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7, 1–24. https://doi.org/10.7717/PEERJ-CS.623
3. Emanuel, Y., Solis, C., & Rivas, H. H. (2024). Comparison of machine learning algorithms for natural language processing (Vol. 11). https://orcid.org/0000-0002-2650-8932
4. Godínez López, E. M., & Rosas-Mayen, N. (2024a). Producción Escrita en Español L2: Influencia de la Autoeficacia y el Perfil Lingüístico en la Complejidad Textual. Revista Veritas de Difusão Científica, 5(3), 1263–1287. https://doi.org/10.61616/rvdc.v5i3.267
5. Godínez López, E. M., & Rosas-Mayen, N. (2024b). Producción Escrita en Español L2: Influencia de la Autoeficacia y el Perfil Lingüístico en la Complejidad Textual. Revista Veritas de Difusão Científica, 5(3), 1263–1287. https://doi.org/10.61616/rvdc.v5i3.267

6. Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. In Geoscientific Model Development (Vol. 15, Issue 14, pp. 5481–5487). Copernicus GmbH. https://doi.org/10.5194/gmd-15-5481-2022
7. LOTAIP. (2004). LEY ORGANICA DE TRANSPARENCIA Y ACCESO A LA INFORMACION PUBLICA. 2004. https://www.educacionsuperior.gob.ec/wp-content/uploads/downloads/2014/09/LOTAIP.pdf
8. Ortiz Zambrano, J., MontejoRáez, A., Lino Castillo, K. N., Gonzalez Mendoza, O. R., & Cañizales Perdomo, B. C. (2020). VYTEDU-CW: Difficult Words as a Barrier in the Reading Comprehension of University Students. Advances in Intelligent Systems and Computing, 1066, 167–176. https://doi.org/10.1007/978-3-030-32022-5_16
9. Ortiz-Zambrano, J. A., Espín-Riofrío, C. H., & Montejo-Ráez, A. (2024). Deep Encodings vs. Linguistic Features in Lexical Complexity Prediction. Neural Computing and Applications. https://doi.org/10.1007/s00521-024-10662-9
10. Ortiz-Zambrano, J. A., & Montejo-Ráez, A. (2020). Overview of ALexS 2020: First Workshop on Lexical Analysis at SEPLN. https://www.ujaen.es/
11. Ortiz-Zambrano, J., Espin-Riofrio, C., & Montejo-Ráez, A. (2022). Transformers for Lexical Complexity Prediction in Spanish Language. Procesamiento Del Lenguaje Natural, 69, 177–188. https://doi.org/10.26342/2022-69-15
12. Ortiz-Zambrano, J., Espin-Riofrio, C., & Montejo-Ráez, A. (2023). SINAI Participation in SimpleText Task 2 at CLEF 2023: GPT-3 in Lexical Complexity Prediction for General Audience Notebook for the SimpleText Lab at CLEF 2023. http://ceur-ws.org
13. Ortiz-Zambrano, J., Espin-Riofrio, C., & Montejo-Ráez, A. (2024). SINAI Participation in SimpleText Task 2 at CLEF 2024: Zero-shot Prompting on GPT-4-Turbo for Lexical Complexity Prediction Notebook for the SimpleText Lab at CLEF 2024. https://openai.com/
14. Ortiz-Zambrano, J., & Montejo-Ráez, A. (2021). SINAI at SemEval-2021 Task 1: Complex word identification using Word-level features. https://pypi.
15. Prada, V., Santiago, D., Martinez, L., & Enrique, F. (2025). Optimización de la evaluación académica mediante procesamiento de lenguaje natural: desarrollo de un sistema de calificación automática para textos en educación superior.

16. Reyes, S. (2024). Aplicación de la espectroscopía NIR y herramientas Quimiométricas para la determinación de componentes químicos del café verde especial producido en la provincia de Charquí, Panamá.
17. Roque López Verónica Montserrat. (2024). Estudios multidisciplinarios: Transparencia y esquemas anticorrupción.
18. Salgado Reyes, N. I., & Elizabeth Trujillo Moreno, G. I. (2024). Sentiment Analysis in Social Network Data: Application of natural language processing and machine learning techniques to analyze opinions and feelings in social network data in the context of information systems. Núm. 1. Enero-Marzo, 10, 314–327. https://doi.org/10.23857/dc.v10i1.3714
19. Taboada Villamarín, A. (2024). Big data en ciencias sociales. Una introducción a la automatización de análisis de datos de texto mediante procesamiento de lenguaje natural y aprendizaje automático. Revista CENTRA de Ciencias Sociales, 3(1). https://doi.org/10.54790/rccs.51
20. Tatachar, A. V. (2021). Comparative Assessment of Regression Models Based On Model Evaluation Metrics. International Research Journal of Engineering and Technology. www.irjet.net