



Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Dirección General de Estudios de Posgrado

Facultad de Ingeniería de Sistemas e Informática

Unidad de Posgrado

**Modelo de machine learning para predicción de la
complejidad léxica de palabras en instituciones
secundarias en un distrito de Cajamarca**

TESIS

Para optar el Grado Académico de Magíster en Ingeniería de
Sistemas e Informática con mención en Gestión de Tecnología de
Información y Comunicaciones

AUTOR

Luis Iván MERA DÁVILA

ASESOR

Dr. Pedro Segundo CASTAÑEDA VARGAS

Lima, Perú

2025



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

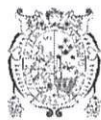
Referencia bibliográfica

Mera, L. (2025). *Modelo de machine learning para predicción de la complejidad léxica de palabras en instituciones secundarias en un distrito de Cajamarca*. [Tesis de maestría, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, Unidad de Posgrado]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Luis Iván Mera Dávila
Tipo de documento de identidad	DNI
Número de documento de identidad	46396150
URL de ORCID	https://orcid.org/0009-0002-7765-2691
Datos de asesor	
Nombres y apellidos	Pedro Segundo Castañeda Vargas
Tipo de documento de identidad	DNI
Número de documento de identidad	10744358
URL de ORCID	https://orcid.org/0000-0003-1865-1293
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Jose César Piedra Isusqui
Tipo de documento	DNI
Número de documento de identidad	25628915
Miembro del jurado 1	
Nombres y apellidos	Virginia Vera Pomalaza
Tipo de documento	DNI
Número de documento de identidad	09813434
Miembro del jurado 2	
Nombres y apellidos	César Augusto Alcántara Loayza
Tipo de documento	DNI
Número de documento de identidad	09132297
Datos de investigación	

Línea de investigación	C.0.5.8. Inteligencia artificial
Grupo de investigación	No aplica
Agencia de financiamiento	Sin financiamiento
Ubicación geográfica de la investigación	País: Perú Departamento: Cajamarca Provincia: Chota Distrito: Lajas Latitud: -6.560470 Longitud: -78.739328
Año o rango de años en que se realizó la investigación	2023 - 2024
URL de disciplinas OCDE	Ingeniería de sistemas y comunicaciones https://purl.org/pe-repo/ocde/ford#2.02.04



UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú. Decana de América
Facultad de Ingeniería de Sistemas e Informática
Vicedecanato de Investigación y Posgrado
Unidad de Posgrado

**ACTA DE SUSTENTACIÓN DE TESIS PARA OPTAR EL GRADO ACADÉMICO DE
MAGÍSTER EN INGENIERÍA DE SISTEMAS E INFORMÁTICA CON MENCIÓN EN
GESTIÓN DE TECNOLOGÍA DE INFORMACIÓN Y COMUNICACIONES**

A los veintidós (22) días del mes de julio de 2025, siendo las 11:00 am., se reunieron en el Auditorio Alfredo Celso Alva Bravo, el Jurado de Tesis conformado por los siguientes docentes:

Mg. Jose César Piedra Isusqui (Presidente)
Mg. Virginia Vera Pomalaza (Miembro)
Mg. César Augusto Alcántara Loayza (Miembro)
Dr. Pedro Segundo Castañeda Vargas (Miembro Asesor)

Se inició la Sustentación invitando al candidato a Magíster **LUIS IVÁN MERA DÁVILA**, para que realice la exposición oral de la tesis para optar el Grado Académico de Magíster en Ingeniería de Sistemas e Informática con mención en Gestión de Tecnología de Información y Comunicaciones, siendo la Tesis intitulada:

**“MODELO DE MACHINE LEARNING PARA PREDICCIÓN DE LA COMPLEJIDAD LÉXICA
DE PALABRAS EN INSTITUCIONES SECUNDARIAS
EN UN DISTRITO DE CAJAMARCA”**

Concluida la exposición, los miembros del Jurado de Tesis procedieron a formular sus preguntas que fueron absueltas por el graduando; acto seguido se procedió a la evaluación correspondiente, habiendo obtenido la siguiente calificación:

DIECISEIS (16) BUENO

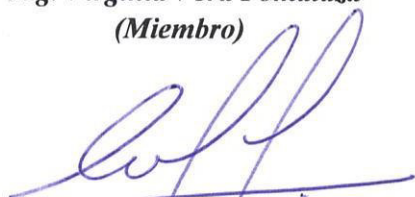
Por tanto, el presidente del Jurado, de acuerdo con el Reglamento General de Estudios de Posgrado, otorga al Bachiller **LUIS IVÁN MERA DÁVILA** el Grado Académico de Magíster en Ingeniería de Sistemas e Informática con mención en Gestión de Tecnología de Información y Comunicaciones.

Siendo las 11:58 horas, el presidente del Jurado de Tesis, da por concluido el acto académico de Sustentación de Tesis.


Mg. Jose César Piedra Isusqui
(Presidente)


Mg. Virginia Vera Pomalaza
(Miembro)


Mg. César Augusto Alcántara Loayza
(Miembro)


Dr. Pedro Segundo Castañeda Vargas
(Miembro Asesor)



UNIVERSIDAD NACIONAL MAYOR DE

SAN MARCOS

Universidad del Perú, Decana de América

CERTIFICADO DE SIMILITUD

Yo, *Pedro Segundo Castañeda Vargas* en mi condición de asesor acreditado con Dictamen N° 000054-2024-UPG-VDIP-FISI/UNMSM de la tesis, cuyo título es "*Modelo de machine learning para predicción de la complejidad léxica de palabras en instituciones secundarias en un distrito de Cajamarca*", presentado por el bachiller *Luis Iván Mera Dávila* para optar el grado académico de *Magíster en Ingeniería de Sistemas e Informática con mención en Gestión de Tecnología de Información y Comunicaciones*.

CERTIFICO que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante la plataforma PIDS, el documento evaluado cuenta con el **4%** de similitud, nivel **PERMITIDO** para continuar con los trámites correspondientes y para su **publicación en el repositorio institucional**.

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del grado académico de *Magíster en Ingeniería de Sistemas e Informática con mención en Gestión de Tecnología de Información y Comunicaciones*.

Lima, 03 de febrero de 2025

Firma del Asesor:

Nombres y Apellidos del asesor: Pedro Segundo Castañeda Vargas

DNI: 10744358

DEDICATORIA

Dedico este trabajo a toda mi familia,
uno de los principales pilares de mi vida,
en gratitud por su apoyo y amor incondicional.

AGRADECIMIENTOS

Agradezco a mis padres quienes siempre han sido y siguen siendo uno de los pilares más importantes en todos los proyectos de mi vida.

Mi agradecimiento al Dr. Pedro Segundo Castañeda Vargas por su asesoría en esta investigación, así como a todos los profesores de este programa de maestría.

Finalmente, agradezco también a todas las personas que han contribuido de forma directa e indirecta a la realización de esta investigación.

Índice general

<i>CAPÍTULO 1: INTRODUCCIÓN</i>	9
1.1 Situación problemática	9
1.2 Formulación del problema	15
1.2.1 Problemas específicos	15
1.3 Justificación teórica	15
1.4 Justificación práctica	19
1.5 Objetivos de la investigación	20
1.5.1 Objetivo general	20
1.5.2 Objetivos específicos	20
<i>CAPÍTULO 2: MARCO TEÓRICO</i>	21
2.1 Marco filosófico o epistemológico de la investigación	21
2.1.1 Empirismo	21
2.1.2 Positivismo.....	21
2.1.3 Racionalismo.....	21
2.1.4 Pragmatismo.....	21
2.2 Antecedentes de la investigación	22
2.2.1 Predicción de la complejidad léxica en el idioma español	22
2.2.2 Predicción de la complejidad léxica de palabras en el idioma inglés.....	24
2.2.3 Vacíos de la literatura revisada, justificación de la originalidad y contribución de la tesis.	25
2.3 Bases teóricas	26
2.3.1 Machine learning.....	26
2.3.1.1 Modelo de machine learning	26
2.3.1.2 Algoritmo de machine learning	27
2.3.1.3 Variables de modelos de machine learning	27
2.3.1.4 Arquitectura de modelo de machine learning	27
2.3.1.5 Tipos aprendizaje en machine learning	29
2.3.1.6 Métricas de evaluación de machine learning.....	30
2.3.1.7 Procesamiento del lenguaje natural (NLP)	31
2.3.2 Complejidad léxica de palabras.....	33
2.3.2.1 Complejidad absoluta	33
2.3.2.2 Complejidad relativa	34
2.3.2.3 Tipos de predicción de la complejidad	35
2.3.2.4 Predicción binaria de la complejidad léxica de palabras (Identificación de palabras complejas)	36
2.3.3 Simplificación léxica.....	36
2.3.4 Comprensión lectora	37
2.3.4.1 Dimensiones de la comprensión lectora	38
<i>CAPÍTULO 3: METODOLOGÍA</i>	39
3.1 Tipo y diseño de la investigación	39
3.2 Unidad de análisis	39
3.3 Población de estudio	39
3.4 Tamaño de la muestra	40
3.5 Selección de la muestra	40
3.5.1 Conjunto de entrenamiento del Dataset CWI Shared Task 2018	41

3.5.2 Conjunto de prueba basado en textos escolares	42
3.6 Técnicas de recolección de datos	42
<i>CAPÍTULO 4: RESULTADOS Y DISCUSIÓN.....</i>	<i>45</i>
4.1 Desarrollo del modelo de machine learning	45
4.1.1. Preprocesamiento de Datos e Ingeniería de Características	45
4.1.2 Selección de la combinación óptima de características	50
4.1.3 Algoritmo de Machine Learning	51
4.1.4 Hiperparámetros	53
4.1.5 Arquitectura del modelo	53
4.2 Análisis, interpretación y discusión de resultados	53
4.2.1 Objetivo Específico 1: Identificar las variables seleccionadas basadas en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, entrenadas en el dataset CWI 2018, que son más importantes al evaluarlas en el dataset de Textos Escolares de instituciones secundarias de un distrito de Cajamarca.	53
a) Análisis de resultados	54
b) Interpretación de resultados	55
c) Discusión de resultados	59
4.2.2 Objetivo Específico 2: Identificar el algoritmo de machine learning evaluado y basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, que obtiene el mejor rendimiento en el dataset CWI 2018 para proponerse a instituciones secundarias de un distrito de Cajamarca.	60
a) Análisis de resultados	60
b) Interpretación de resultados	60
c) Discusión de resultados	61
4.2.3 Objetivo Específico 3: Determinar el rendimiento y la eficiencia en memoria de la arquitectura del modelo basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, al compararlo con un baseline del CWI 2018, considerando los requisitos de memoria RAM viables para las instituciones secundarias de un distrito de Cajamarca.	61
a) Presentación de resultados	62
b) Interpretación de resultados	63
c) Discusión de resultados	64
4.3. Pruebas de hipótesis	65
4.3.1 Contrastación de hipótesis específica H1	65
4.3.2 Contrastación de hipótesis específica H2	65
4.3.3 Contrastación de hipótesis específica H3	66
4.3.4 Contrastación de hipótesis general	67
4.4 Presentación de resultados.....	67
<i>CONCLUSIONES.....</i>	<i>69</i>
<i>RECOMENDACIONES.....</i>	<i>72</i>
<i>REFERENCIAS BIBLIOGRÁFICAS.....</i>	<i>74</i>
<i>ANEXO A – CUESTIONES METODOLÓGICAS.....</i>	<i>83</i>
<i>ANEXO B – ACERCA DE LOS CONJUNTOS DE DATOS</i>	<i>86</i>
<i>ANEXO C – PIPELINE DE MACHINE LEARNING.....</i>	<i>89</i>
<i>ANEXO D – SOLICITUD DE PRÉSTAMO DE MATERIALES E INFORMACIÓN PARA LA INVESTIGACIÓN.....</i>	<i>90</i>

LISTA DE TABLAS

Tabla 1 Conjuntos de datos y propósitos en el desarrollo del modelo de este estudio.....	40
Tabla 2 Resultados de la validación cruzada 5 pliegues para las combinaciones de características en el conjunto de entrenamiento del dataset CWI Shared Task 2018	50
Tabla 3 Resultados de la validación cruzada de 5 de pliegues en el conjunto de entrenamiento para los algoritmos con la selección óptima de características.....	52
Tabla 4 Características seleccionadas durante el entrenamiento para el modelo de machine learning	54
Tabla 5 Resultados de la evaluación en el dataset de prueba vs el baseline de la tarea compartida del CWI Shared Task 2018	62

LISTA DE FIGURAS

Figura 1 Resultados por medida promedio en lectura de la evaluación PISA del año 2018	9
Figura 2 Rendimiento lector y gasto en educación según evaluación PISA 2018	11
Figura 3 Evolución de la pobreza regional 2004-2022 en el Perú	12
Figura 4 Lectura en 2º secundaria: medida promedio y niveles de logro según nivel socioeconómico	12
Figura 5 Lectura en 2º secundaria: medida promedio y niveles de logro en el 2019 y 2022 en regiones	13
Figura 6 Proceso de simplificación léxica de 4 etapas	37
Figura 7 Importancia de características mediante el método de permutación (en el dataset de prueba basado en Textos Escolares)	54
Figura 8 Rendimiento de los algoritmos identificados	60
Figura 9 Matriz de confusión de la evaluación en el dataset basado en textos escolares para el modelo desarrollado	62
Figura 10 Medidas del consumo de memoria en MB para carga del modelo en 10 iteraciones	62

LISTA DE ACRÓNIMOS Y ABREVIATURAS

ACL	Association for Computational Linguistics
BEA	Workshop on the Innovative Use of NLP for the Building Educational Applications
BERT	Bidirectional Encoder Representations from Transformer
BOW	Bag of words
CNN	Convolutional neural network
CWI	Complex word identification
EBR	Educación básica regular
GNN	Graph neural networks
LCP	Lexical complexity prediction
LightGBM	Light gradient boosting machine
LLM	Large language model
LR	Logistic regression
LS	Lexical simplification
MAE	Mean absolute error
MSE	Mean squared error
MWE	Multi-Word expressions
NB	Naive Bayes
NER	Named entity recognition
NLP	Natural language processing
ODS	Objetivos de Desarrollo Sostenible
PLMs	Pre-trained language models
POS	Part of speech
RF	Bosques aleatorios
RMSE	Root mean square error
RoBERTa	Robustly optimized BERT approach
SVM	Support vector machine
TFIDF	Term frequency-inverse document frequency
XGBoost	Extreme Gradient Boosting

RESUMEN

El Perú presenta un bajo desempeño en comprensión lectora a nivel secundario, especialmente en regiones vulnerables como Cajamarca, donde la complejidad léxica podría ser una barrera significativa para el aprendizaje.

En este sentido, el objetivo de este estudio es determinar la efectividad de un modelo propuesto de machine learning para predecir binariamente la complejidad léxica de palabras en instituciones secundarias en un distrito de Cajamarca en comparación con el baseline del CWI Shared Task 2018 para el idioma español.

La metodología empleada usó el dataset en español del CWI Shared Task 2018 para el entrenamiento y un dataset de validación que se recolectó a partir de textos escolares para la evaluación del modelo en el contexto específico de las instituciones secundarias mencionadas.

Los resultados en el contexto educativo demostraron un desempeño superior a los baselines establecidos, en cuanto a la precisión, así como en la eficiencia del uso de recursos computacionales evaluados.

Se determinó la efectividad del modelo de machine learning para la predicción binaria de la complejidad léxica de palabras en el contexto de las instituciones secundarias de un distrito de Cajamarca, haciendo frente a la escasa literatura para esta tarea en el idioma español y representando un primer acercamiento hacia sistemas personalizados en dicho contexto educativo, lo cual podría contribuir en la mejora de la comprensión lectora y del aprendizaje del alumnado.

Se plantea la sugerencia al MINEDU e instituciones secundarias de implementar el modelo para ajustar la complejidad léxica del material educativo, crear glosarios y explorar el uso de inteligencia artificial en la simplificación léxica de forma más eficaz y eficiente.

Palabras clave: Machine Learning, Identificación de palabras complejas, Predicción binaria de la complejidad Léxica, Simplificación Léxica, Procesamiento del lenguaje Natural, Complejidad Léxica, Inteligencia artificial.

ABSTRACT

Peru shows low performance in reading comprehension at the secondary level, especially in vulnerable regions like Cajamarca, where lexical complexity could be a significant barrier to learning.

In this regard, the objective of this study is to determine the effectiveness of a proposed machine learning model to predict binary lexical complexity of words in secondary institutions in a district of Cajamarca compared to the baseline of the CWI Shared Task 2018 for the Spanish language.

The methodology employed used the Spanish dataset from the CWI Shared Task 2018 for training and a validation dataset that was collected from school textbooks for model evaluation in the specific context of the aforementioned secondary institutions.

The results in the educational context demonstrated superior performance to the established baselines, in terms of precision as well as efficiency in the use of computational resources evaluated.

The effectiveness of the machine learning model for binary prediction of lexical complexity of words in the context of secondary institutions in a district of Cajamarca was determined, addressing the scarce literature for this task in the Spanish language and representing a first approach toward personalized systems in said educational context, which could contribute to improving reading comprehension and student learning.

A suggestion is proposed to MINEDU and secondary institutions to implement the model to adjust the lexical complexity of educational material, create glossaries, and explore the use of artificial intelligence in lexical simplification in a more effective and efficient manner.

Keywords: Machine Learning, Complex Word Identification, Binary Prediction of Lexical Complexity, Lexical Simplification, Natural Language Processing, Lexical Complexity, Artificial Intelligence

CAPÍTULO 1: INTRODUCCIÓN

En este capítulo se describen algunos puntos acerca de la situación problemática, como el panorama de la educación, la comprensión lectora en el Perú, las instituciones educativas públicas de educación básica regular (EBR) de nivel secundario entorno a esta investigación, así como la formulación del problema, la justificación y los objetivos de esta investigación.

1.1 Situación problemática

Para describir de forma general la situación de la educación en el Perú se menciona que: “El panorama de la educación en el Perú es preocupante. Encuestas recientes han revelado cifras que presentan una visión alarmante de la educación en nuestro país” (Care, 2023).

Según evaluaciones de Smartick (2022), Perú se posiciona como el país de Latinoamérica con el desempeño más bajo en comprensión lectora, siendo los estudiantes de secundaria quienes obtuvieron las calificaciones más desfavorables.

Figura 1

Resultados por medida promedio en lectura de la evaluación PISA del año 2018



Nota. La figura muestra los resultados para 77 países, incluyendo el Perú. De “Resultados Evaluación Internacional PISA” por MINEDU, 2018 (<http://umc.minedu.gob.pe/resultadospisa2018/>).

En la Figura 1, se muestran los resultados por medida promedio en lectura en el año 2018 del Programa para la Evaluación Internacional de Alumnos (PISA) de la OCDE, donde nuestro país ocupó el lugar 64 de una población de 77 países, en dicha evaluación que mide principalmente la capacidad de los estudiantes para comprender, interpretar y reflexionar sobre textos, habilidades las cuales dependen fundamentalmente del dominio léxico. Como señalan Perfetti y Stafura (2014), la comprensión lectora está condicionada por el conocimiento del vocabulario (tanto de palabras simples como de palabras complejas) y la capacidad de procesamiento léxico. En este sentido, la complejidad léxica está directamente relacionada con el conocimiento del vocabulario: a mayor dominio léxico, mayor capacidad para manejar palabras de alta complejidad léxica (términos especializados, técnicos o poco frecuentes). Precisamente, en el caso peruano, se ha identificado que uno de los principales obstáculos para el desarrollo de competencias lectoras es precisamente la brecha entre el vocabulario (léxico) que poseen los estudiantes (Junyent, 2017) y la complejidad léxica de los materiales educativos a los que se enfrentan (MINEDU, 2013), por ende, esto puede afectar los resultados en evaluaciones estandarizadas como las de PISA.

Este estudio aborda la complejidad léxica desarrollando un modelo de machine learning para predecir si una palabra dentro de una frase dada es compleja en el contexto de instituciones secundarias de un distrito de Cajamarca. El modelo propuesto puede tener un empleo versátil dentro de dichas instituciones, tales como: formar parte de herramientas de simplificación léxica, personalizar el léxico empleado en los materiales de enseñanza-aprendizaje, facilitar la creación de glosarios adaptados al nivel educativo específico, apoyar a docentes en la identificación proactiva de vocabulario potencialmente difícil para sus estudiantes, servir como base para el desarrollo de aplicaciones educativas interactivas que refuercen el aprendizaje de nuevos términos, y contribuir a la evaluación objetiva del nivel de complejidad de los textos utilizados en el aula. Todo esto con el fin de ofrecer herramientas que apoyen la mejora de la comprensión lectora y ayuden a optimizar el proceso de enseñanza-aprendizaje en la región, atendiendo a las necesidades lingüísticas particulares de los estudiantes.

Por otro lado, según la Universidad de Guanajuato (2022), el problema del aprendizaje puede ser producto de diversos factores, entre los que destacan: el intelectual (relacionado con las capacidades mentales del individuo), el físico (donde se contempla la nutrición, desarrollo físico, entre otros), el emocional, el social, el ambiental, así como la personalidad del docente.

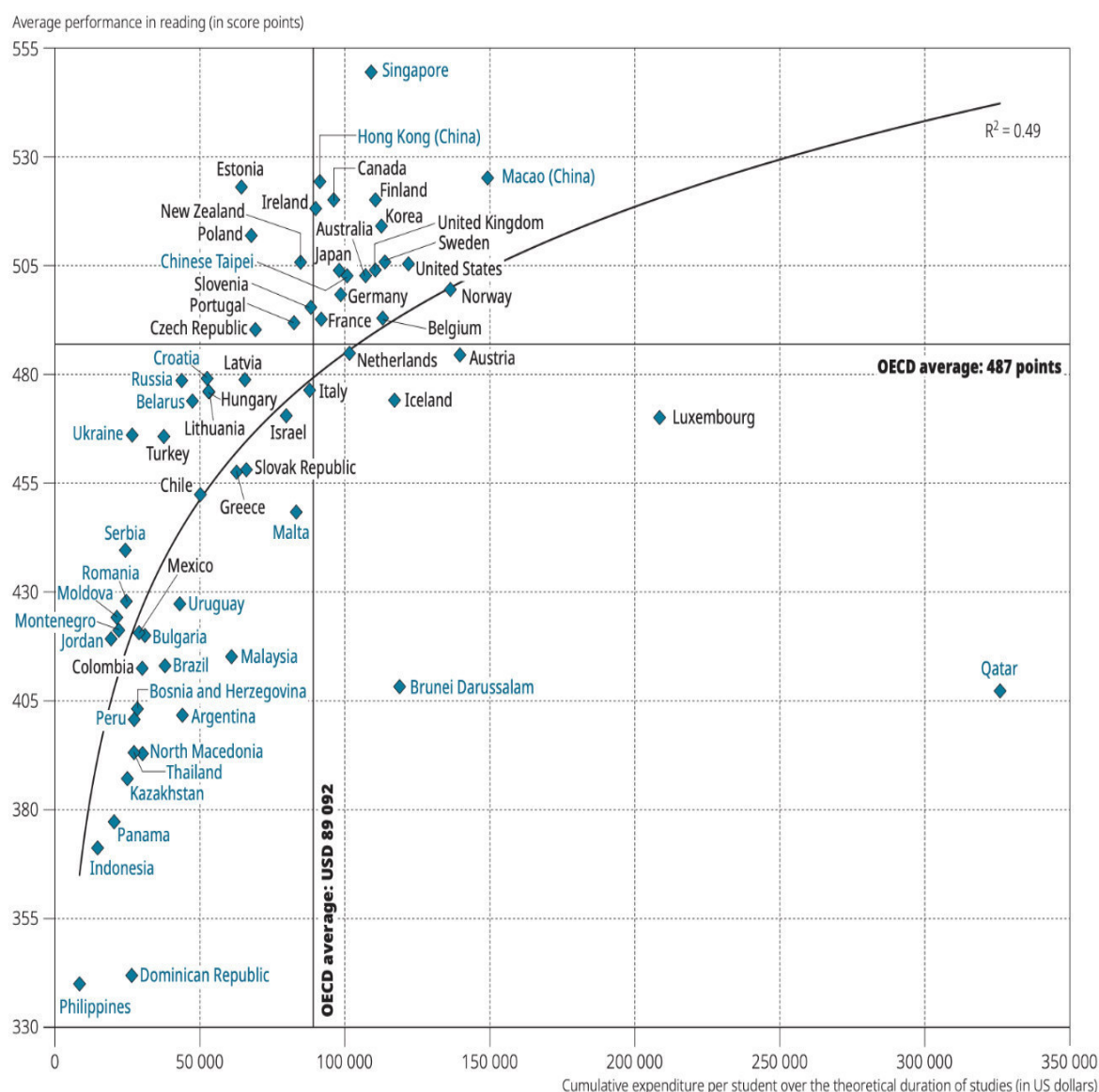
Además, de acuerdo con la Universidad de Guanajuato (2022), el factor aprendizaje está relacionado con la falta de conocimientos de los temas. Asimismo, menciona que el proceso de aprendizaje de cualquier estudiante puede verse afectado por métodos de

trabajo o estudio inadecuados. Así como, no tener buen rendimiento en historia podría ser causa a la baja capacidad lectora o a flaqueza en el dominio del lenguaje.

La Figura 2 muestra la relación entre el gasto en educación y el rendimiento en lectura de los países evaluados en el 2018 según las pruebas realizadas por PISA, en la que el Perú aparece entre las naciones que menos invierten y que además obtuvieron el menor puntaje en lectura. Es importante desarrollar métodos y herramientas accesibles que mejoren la comprensión lectora y se adapten a las necesidades de los estudiantes y escuelas peruanas.

Figura 2

Rendimiento lector y gasto en educación según evaluación PISA 2018

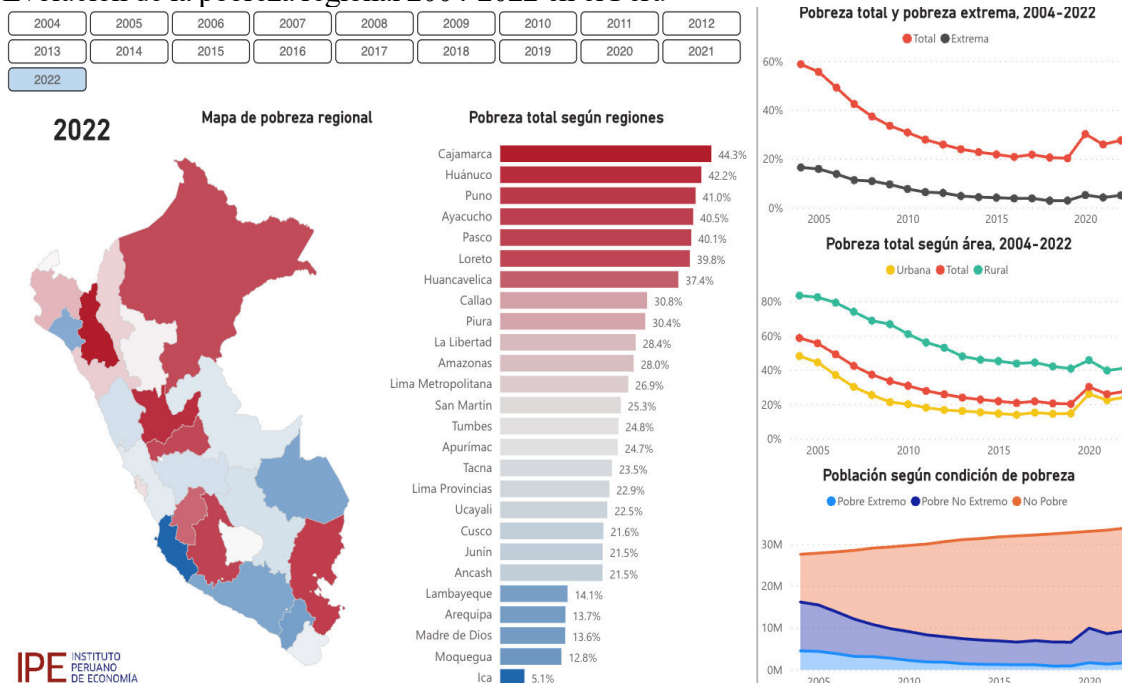


Nota. Relación del rendimiento promedio en lectura con el gasto acumulado por estudiante para varios países, basado en los resultados de la evaluación PISA. En “PISA 2018 Results. WHAT STUDENTS KNOW AND CAN DO VOLUME I” por OECD, 2019 (<https://doi.org/10.1787/7aa618c1-en>).

Cabe recalcar que Cajamarca ocupa el primer lugar de la lista de pobreza total regional del Perú según el Instituto Peruano de Economía (IPE) como se muestra en la Figura 3.

Figura 3

Evolución de la pobreza regional 2004-2022 en el Perú

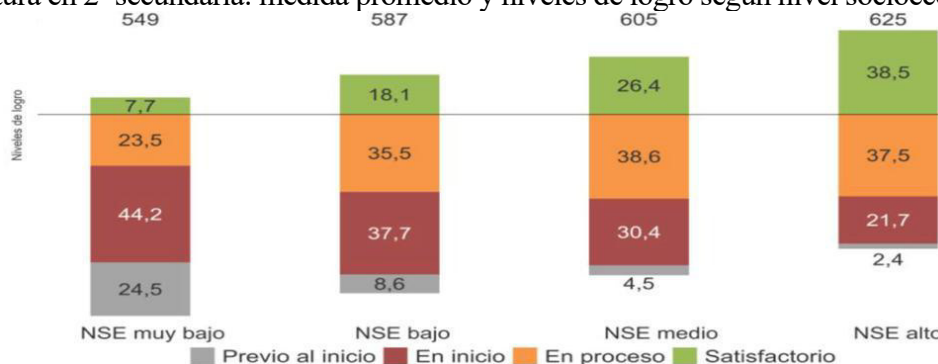


Nota. Mapa regional, estadísticas por región y tendencias nacionales de pobreza total y extrema del Perú en el periodo del 2004 al 2022. De “Evolución de la pobreza regional – Tablero Interactivo” por IPE, 2022 (<https://www.ipe.org.pe/portal/evolucion-de-la-pobreza-regional-tablero-interactivo/>).

Se podría sugerir una aparente relación entre los niveles socioeconómicos y los niveles de desempeño en lectura, ya que, de acuerdo con la evaluación muestral de estudiantes del MINEDU del año 2023, los niveles de logro en lectura para alumnos de segundo grado de educación secundaria de los niveles socioeconómicos (NSE) muy bajo y bajo son inferiores respecto a los cuatro NSE evaluados como se muestra en la Figura 4.

Figura 4

Lectura en 2º secundaria: medida promedio y niveles de logro según nivel socioeconómico

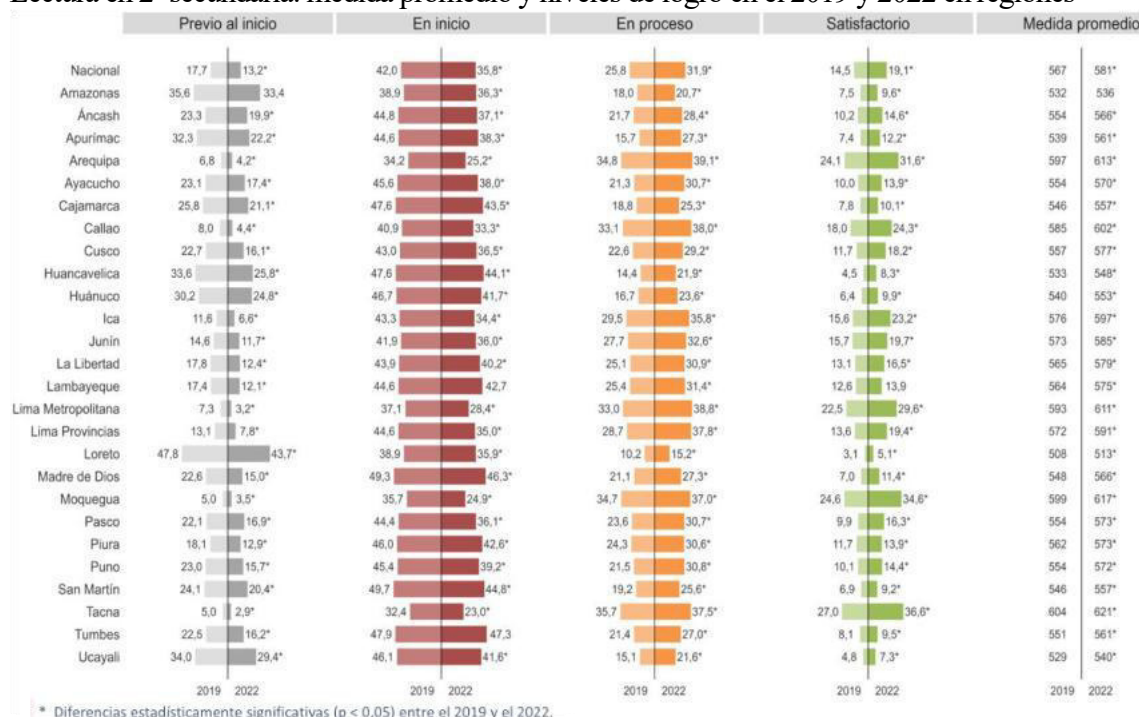


Note. Para cada NSE, se presentan los porcentajes de estudiantes en cuatro niveles de desempeño: Previo al inicio, En inicio, En proceso y Satisfactorio. De “Evaluación muestral de Estudiantes” por MINEDU, 2023 (<http://umc.minedu.gob.pe/resultados2022/>).

Por otro lado, las medidas promedio y niveles de logro en lectura de los años 2019 y 2022 de las regiones del Perú, que han sido obtenidas con la Evaluación Muestral (EM) de Estudiantes del Ministerio de Educación para el segundo grado de educación secundaria, son mostrados en la Figura 5. Se ve que Tacna, Moquegua y Arequipa son las regiones con mayor medida promedio en lectura, pero para Cajamarca se aprecia que va al puesto 19 en 2019 y en 2022 cae a la posición 20 de 25 regiones evaluadas en cuanto a medida de lectura. Además, se puede visualizar que Cajamarca se encuentra por debajo de la media nacional, situación que debería representar desafíos tanto para el estado como para cualquier investigador interesado en realizar estudios y encontrar posibilidades de mejora ante esta problemática.

Figura 5

Lectura en 2° secundaria: medida promedio y niveles de logro en el 2019 y 2022 en regiones



Nota. La figura muestra los resultados de la evaluación de lectura para estudiantes de segundo grado de secundaria en las diferentes regiones del Perú, comparando los años 2019 y 2022. De “Evaluación Muestral de Estudiantes 2022” por MINEDU, 2023 (<http://umc.minedu.gob.pe/resultadossem2022/>)

La complejidad léxica es un aspecto central para que la comprensión lectora se lleve a cabo. Las palabras con elevada complejidad léxica pueden provocar que la comprensión no avance y así dificultar el proceso de lectura. Por esta razón, expandir el vocabulario y trabajar la capacidad de procesamiento de palabras de alta complejidad léxica es adecuado para mejorar la comprensión lectora.

Laufer (2013) indica que la comprensión lectora está influenciada por la complejidad léxica de un texto. Cuanto más complejas sean las palabras utilizadas en un texto, mayor será el desafío para los lectores para comprenderlo. El conocimiento de vocabulario es un

predictor importante de la habilidad de lectura y los lectores necesitan tener un conocimiento suficiente de palabras para comprender un texto sin dificultad.

Teniendo en cuenta la relevancia que tiene la complejidad léxica en la comprensión lectora, resulta muy importante realizar su estudio, sobre todo en contextos educativos específicos y más aún en las zonas donde los resultados de la comprensión de lectura pueden presentar ciertos obstáculos (IPE, 2022; MINEDU, 2018, 2023) para lograr un buen desempeño. Dado el contexto de desafíos en comprensión lectora y la relevancia de la complejidad léxica, este estudio se enfoca en el distrito de Lajas, ubicado en la provincia de Chota, departamento de Cajamarca, Perú. Este distrito cuenta con cinco instituciones educativas públicas de educación básica regular (EBR) de nivel secundario, además de una institución de Educación Básica Alternativa (EBA). De dichas instituciones, dos se encuentran en la zona urbana y las otras cuatro se sitúan en la zona rural. Esta investigación se enfoca específicamente en los colegios de educación básica regular de la zona urbana, lo que nos lleva a enfocarnos en la institución educativa secundaria pública de educación básica regular: "Anaximandro Vega" - Lajas. Esta institución está ubicada en el ex fundo San Carlos, en el casco principal del distrito de Lajas y sirve como un importante pilar educativo para la comunidad local (Educación secundaria en Lajas, Chota, s/f; MINEDU, 2022).

El Colegio Secundario "Anaximandro Vega" - Lajas, identificado con el código modular 0453019, es una institución secundaria pública de educación básica regular perteneciente a la jurisdicción del distrito de Lajas, en la provincia de Chota, en el departamento de Cajamarca, en el Perú. La institución tiene una población estudiantil de aproximadamente 400 estudiantes distribuidos en 16 secciones desde el primer hasta el quinto grado de educación secundaria. Dicha institución proporciona una formación de cinco años reglamentarios de acuerdo con la normativa del sistema educativo del Perú.

Para la presente investigación, se consultaron los materiales bibliográficos disponibles en la biblioteca del Colegio Secundario "Anaximandro Vega" - Lajas, especialmente aquellos relacionados con la identificación binaria de la complejidad léxica en instituciones secundarias del distrito. Dichos materiales fueron fundamentales para el objetivo principal de desarrollar un modelo de machine learning para la predicción binaria de la complejidad léxica de palabras (también conocida como identificación de palabras complejas), con la finalidad de ofrecer una herramienta que podría servir como una alternativa más eficaz y eficiente que esté disponible para este propósito y contexto educativo. Dicho modelo podría emplearse en el apoyo de la mejora de la comprensión lectora de los estudiantes de dichas instituciones, ayudando a la identificación de términos complicados para ellos, lo cual sirve en el proceso de adaptación del material pedagógico que esté acorde con el nivel de lectura de los alumnos y, de esa forma, contribuir en la creación de materiales didácticos más accesibles e inclusivos para ellos. Esta herramienta se propone como una alternativa adicional que puede complementar a los métodos que comúnmente son utilizados en este entorno educativo local, ofreciendo de esta manera una solución de tecnología basada en inteligencia artificial que los profesores pueden usar de acuerdo con sus necesidades dentro del proceso de enseñanza-aprendizaje. Además, esta herramienta puede ser empleada en otras instituciones de educación secundaria

públicas del país, favoreciendo de esta manera el mejoramiento global de la calidad educativa y ofreciendo un recurso extra para enfrentar los retos en la comprensión de textos para los estudiantes y la educación que se imparte.

1.2 Formulación del problema

¿Cuál es la efectividad de un modelo de machine learning basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica de palabras, en comparación con un baseline del CWI Shared Task 2018, en instituciones secundarias en un distrito de Cajamarca?

1.2.1 Problemas específicos

- ¿Cuáles de las variables seleccionadas, basadas en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, entrenadas en el dataset CWI 2018, son más importantes al evaluarlas en el dataset de Textos Escolares de instituciones secundarias de un distrito de Cajamarca?
- ¿Qué algoritmo de machine learning evaluado y basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, obtiene el mejor rendimiento en el dataset CWI 2018 para proponerse a instituciones secundarias de un distrito de Cajamarca?
- ¿Cuál es el rendimiento y la eficiencia en memoria de la arquitectura del modelo basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, al compararlo con un baseline del CWI 2018, considerando los requisitos de memoria RAM viables para las instituciones secundarias de un distrito de Cajamarca?

1.3 Justificación teórica

La relevancia teórica del machine learning para la predicción de complejidad léxica radica en su capacidad para abordar un desafío multidimensional caracterizado por la interacción de factores como frecuencia de uso, polisemia, abstracción conceptual y dependencia contextual, que no pueden tratarse adecuadamente mediante enfoques tradicionales. Las fórmulas clásicas de legibilidad como Flesch-Kincaid (Kincaid et al., 1975) o los diccionarios de frecuencia por si solos resultan insuficientes, mientras que la evaluación manual de millones de palabras o textos es inviable en términos de consistencia y escalabilidad.

Además, la relevancia del Machine Learning se evidencia en su capacidad para identificar patrones que correlacionan características lingüísticas con niveles de complejidad. Los modelos de machine learning pueden analizar palabras en su contexto sintáctico-semántico, adaptarse a diferentes perfiles de lectores y procesar eficientemente grandes corpus textuales. La evidencia empírica respalda esta aproximación: competencias como

SemEval y estudios recientes (Bani Yaseen et al., 2021; Mosquera, 2021; Ortiz-Zambrano y Montejo-Ráez, 2021; G. Paetzold y Specia, 2016) demuestran correlaciones notablemente altas entre predicciones de ML y evaluaciones humanas de complejidad léxica. Esto tendría un impacto directo en el ámbito educativo, permitiendo desarrollar materiales mejor adaptados que pueden reducir significativamente las brechas de comprensión lectora identificadas en evaluaciones nacionales e internacionales.

La aplicación del machine learning a este campo comenzó a consolidarse con trabajos seminales como el de Shardlow (2013a), quien implementó los primeros algoritmos de aprendizaje supervisado para la identificación automática de palabras complejas. Este enfoque representó un punto de inflexión metodológico al permitir modelar computacionalmente un fenómeno previamente abordado desde perspectivas principalmente teóricas. Posteriormente, G. Paetzold y Specia (2016) expandieron significativamente el marco teórico al demostrar que los modelos de ML podían integrar múltiples dimensiones lingüísticas simultáneamente, validando empíricamente la naturaleza de múltiples factores de la complejidad léxica. Estos trabajos pioneros sentaron las bases para investigaciones posteriores que continúan refinando los métodos computacionales para la evaluación de complejidad léxica

Para la arquitectura del modelo desarrollado en el presente estudio se ha decidido emplear características tales como frecuencias de palabras basadas en recursos lingüísticos del español actual, embeddings de nueva generación y algunas propiedades léxicas. La selección de dichas variables para este modelo predictivo de complejidad léxica encuentra sólida justificación en diversas teorías lingüísticas, psicopedagógicas y del procesamiento del lenguaje natural.

Frecuencias basadas en recursos lingüísticos actuales

Desde una perspectiva lingüística la frecuencia léxica constituye un constructo fundamental, esta variable refleja cuán comúnmente aparece una palabra en un corpus determinado. Nation (2001) señala que las palabras frecuentes son más fáciles de reconocer, procesar y comprender, mientras que las palabras de baja frecuencia tienden a vincularse con mayor complejidad léxica, siendo generalmente más largas, menos conocidas o pertenecientes a dominios especializados.

En el ámbito psicopedagógico, la Teoría de la Carga Cognitiva de Sweller (1994), proporciona una explicación cognitiva para este fenómeno. Las palabras menos frecuentes y más complejas sobrecargan la limitada memoria de trabajo de los alumnos, dificultando su procesamiento y posterior recuperación. Cuando intentamos aprender vocabulario menos familiar, la cantidad de esfuerzo mental requerido aumenta significativamente, afectando directamente la disponibilidad léxica. Complementariamente, la Zona de Desarrollo Próximo de Vygotsky (1978), sugiere que la adquisición óptima de vocabulario debe seguir una progresión desde palabras de frecuencia media-alta hacia términos de baja frecuencia, construyendo el aprendizaje de forma escalonada y contextualizada.

Desde la perspectiva del procesamiento del lenguaje natural, la frecuencia léxica ha demostrado su poder predictivo en modelos computacionales. Investigaciones de Chen y

Meurers (2016) y Kauchak (2016) confirman la fuerte correlación entre la frecuencia de una palabra y su dificultad, estas convergencias teóricas y modelado computacional refuerzan la importancia de la frecuencia como variable predictora de la disponibilidad léxica, permitiendo el desarrollo de sistemas de IA que pueden evaluar y predecir con precisión qué palabras estarán más disponibles para un aprendiz en diferentes contextos comunicativos.

Embeddings de palabras

El ámbito psicopedagógico revela que los estudiantes encuentran palabras complejas no solo por su rareza, sino por su distancia semántica al vocabulario que ya dominan o encuentran en su entorno educativo. Los embeddings modelan esa proximidad semántica y ofrecen una representación data-driven para estimar qué tan ajena es una palabra para los estudiantes.

En este sentido, desde la Zona de desarrollo próximo de Vygotsky (1978), estos modelos computacionales permiten evaluar qué tan cerca o lejos está una palabra objetivo del vocabulario conocido y, en consecuencia, si una palabra estuviera en dicha zona de un estudiante o si requiriese un andamiaje adicional.

Complementariamente, al considerar la teoría de la carga cognitiva de Sweller (1994), observamos que las palabras que según los embeddings ocupan posiciones más alejadas de las regiones más densamente pobladas (que representan el léxico frecuente) presentan una mayor carga de procesamiento y, por ende, mayor complejidad para los aprendices. Esta conexión entre distancia vectorial y esfuerzo cognitivo ofrece un puente entre modelos computacionales y teorías del aprendizaje.

Desde el punto de vista del procesamiento del lenguaje natural, los embeddings codifican de manera eficiente y escalable grandes volúmenes de información semántica y contextual necesaria para predecir la complejidad léxica en corpus extensos (OPENAI, 2024). Esta capacidad representa una ventaja significativa frente a métodos tradicionales más limitados.

Particularmente relevante es que estos modelos trasladan el trabajo manual de ingeniería de atributos (morfología, familiaridad estimada, etc.) a una codificación automática y multidimensional de la información sobre cada palabra. De este modo, capturan sutilezas semánticas que otros enfoques no pueden detectar.

Propiedades léxicas

Desde la perspectiva psicopedagógica, la longitud de las palabras es un factor que incide directamente en la carga cognitiva y en los procesos de adquisición léxica. Diversos estudios señalan que las palabras más largas resultan más difíciles de decodificar y retener (Baddeley, 1992; Sweller, 1994), lo que incrementa la carga sobre la memoria de trabajo, puede dificultar la comprensión lectora y puede obstaculizar la adquisición y el uso eficaz del vocabulario. Esta relación entre la complejidad formal y el esfuerzo cognitivo resalta

la importancia de considerar variables cuantitativas, como la longitud de las palabras, al analizar el procesamiento y la enseñanza del léxico.

Por su parte, desde el ámbito del procesamiento del lenguaje natural y su aplicación en educación, estas mismas características de forma, como la longitud, el número de palabras en una frase y el uso de mayúsculas, constituyen atributos fundamentales en tareas básicas de NLP. Estos son esenciales en procesos de preprocesamiento de texto e ingeniería de características, tales como la segmentación, la tokenización, la normalización, el análisis morfosintáctico y el reconocimiento de entidades nombradas (NER, por su sigla en inglés), de acuerdo con lo señalado por Jurafsky y Martin (2025). La consideración sistemática de estos elementos permite desarrollar modelos computacionales más eficaces para el análisis y tratamiento automático del lenguaje, contribuyendo así tanto a la investigación como a la innovación educativa.

Además, la justificación teórica de esta investigación se fundamenta en aspectos muy importantes, como su potencial contribución a futuros estudios en cuanto a recursos para la validación o para el desarrollo de modelos de inteligencia artificial que aborden las etapas del proceso de simplificación léxica para contextos educativos nacionales relacionados, así como la necesidad identificada en la literatura actual de expandir este campo de estudio en el idioma español.

Por otro lado, la revisión de la literatura considerada en el marco de este estudio recoge importantes áreas de oportunidad que la presente investigación pretende abordar. Por ejemplo, las investigaciones (Gooding y Tragut, 2022; Lee y Yeung, 2018) han mostrado la necesidad de crear modelos de inteligencia artificial para la simplificación léxica que se adapten a las características de cada contexto determinado. Ante esta necesidad, la presente investigación se presenta como un primer acercamiento en el contexto de las instituciones educativas de educación secundaria básica regular pública de un distrito del departamento de Cajamarca.

Del mismo modo, trabajos recientes, como por ejemplo el de Shardlow et al. (2024), han indicado que existe un considerable desbalance en cuanto al número de trabajos llevados a cabo en esta área de estudio en la medida en que la mayoría de estos se encuentra en inglés lo que supone una notable escasez en cuanto a trabajos y recursos (Alarcón et al., 2021; Bott et al., 2024; Degraeuwe y Goethals, 2024; Ortiz-Zambrano et al., 2022, 2023b) en otras lenguas, siendo el español una de ellas. En esta línea, el trabajo de (Alarcón et al., 2021) hace eco de las oportunidades de mejora en cada una de las etapas del proceso de simplificación léxica, sobre todo con la llegada de tecnologías y recursos actuales con un gran potencial.

Si bien la investigación se llevará a cabo, de una forma inicial, en el marco contextual concreto de un distrito de Cajamarca, su aporte al desarrollo de modelos computacionales del procesamiento del lenguaje natural y más concretamente a la simplificación léxica, puede permitir su posterior ampliación para que, luego, sirva como base de apoyo para otra investigación, en otras zonas del país.

1.4 Justificación práctica

El desarrollo de un modelo de aprendizaje automático con finalidad predictiva sobre la complejidad léxica de las palabras para su posible adopción en instituciones públicas de educación básica regular secundaria en un distrito de Cajamarca puede generar diversos beneficios medibles en su uso práctico para sus alumnos y docentes.

Uno de los beneficios que este modelo puede brindar es la personalización de los materiales de aprendizaje, ya que puede ser usado en el proceso de adaptar dichos materiales al nivel de dominio léxico de los estudiantes (mediante la identificación de palabras complejas) y hacer el proceso de enseñanza-aprendizaje más eficaz y eficiente.

Ofrece la capacidad objetiva de evaluar y comparar el nivel de dificultad de los textos o exámenes que se imparten al alumnado de dichas instituciones utilizando métricas como índices de complejidad léxica en una escala estandarizada.

Asimismo, el modelo podría facilitar la identificación temprana de palabras que los estudiantes encuentran difíciles, permitiendo así, diseñar y facilitar la construcción de glosarios para desarrollar estrategias específicas de refuerzo y fortalecer las competencias lingüísticas que son esenciales para el desempeño académico en todas las asignaturas.

Por otro lado, al reducir las barreras de acceso al significado de los textos, se espera disminución en los índices de frustración, lo que podría reflejarse en una reducción de la deserción escolar anual del alumnado.

El impacto del modelo puede evaluarse mediante la mejora en las calificaciones de comprensión lectora, la reducción de repitencia y deserción escolar, y la cantidad de materiales adaptados generados. Otras métricas incluyen el tiempo que los estudiantes tardan en comprender un texto, la actitud hacia la lectura, y la relación entre el uso de materiales adaptados y el rendimiento académico. Para el dominio léxico, pueden medirse la disminución de palabras desconocidas, el aumento de vocabulario aprendido y la satisfacción de los estudiantes con el apoyo recibido.

Adicionalmente, el modelo podría funcionar como una herramienta educativa en sí misma, facilitando la comprensión de la complejidad léxica y su influencia en el aprendizaje. Además, en un mundo donde las habilidades de machine learning y análisis de datos son cada vez más valiosas, trabajar en este tipo de proyectos puede dar muestra a los estudiantes de una experiencia práctica valiosa en cuanto a las aplicaciones de inteligencia artificial para un problema concreto del mundo real.

Así mismo, resulta conveniente estudiar este problema porque existe la necesidad de tener más posibilidades de mejora en el desempeño de lectura de los alumnos de nivel secundario. El empleo de dicho modelo podría servir como un catalizador en el campo de la comprensión lectora que es de suma importancia para todas las sociedades.

De acuerdo con la ONU (2015), la educación es fundamental para salir de la pobreza. Además, la mayoría “de niños y adolescentes de todo el mundo no han alcanzado estándares mínimos de competencia en lectura y matemáticas”.

Según la ONU (2016), la educación es de suma importancia para poder lograr los objetivos de desarrollo sostenibles (ODS), reducir desigualdades, lograr la igualdad de género, mejorar la calidad de vida y desarrollar sociedades más pacíficas.

1.5 Objetivos de la investigación

1.5.1 Objetivo general

Determinar la efectividad de un modelo de machine learning basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica de palabras, en comparación con un baseline del CWI Shared Task 2018, en instituciones secundarias en un distrito de Cajamarca.

1.5.2 Objetivos específicos

- Identificar las variables seleccionadas basadas en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, entrenadas en el dataset CWI 2018, que son más importantes al evaluarlas en el dataset de Textos Escolares de instituciones secundarias de un distrito de Cajamarca.
- Identificar el algoritmo de machine learning evaluado y basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, que obtiene el mejor rendimiento en el dataset CWI 2018 para proponerse a instituciones secundarias de un distrito de Cajamarca.
- Determinar el rendimiento y la eficiencia en memoria de la arquitectura del modelo basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, al compararlo con un baseline del CWI 2018, considerando los requisitos de memoria RAM viables para las instituciones secundarias de un distrito de Cajamarca.

CAPÍTULO 2: MARCO TEÓRICO

En este capítulo se aborda el marco filosófico o epistemológico, los antecedentes y las bases teóricas que sirven como pilares para el desarrollo de esta investigación.

2.1 Marco filosófico o epistemológico de la investigación

En esta sección, se introduce el marco filosófico-epistemológico seguido para el desarrollo de esta investigación, proporcionando una base teórica que guía el diseño, implementación y evaluación de esta investigación.

2.1.1 Empirismo

En base a Locke (1689), este trabajo fundamenta su conocimiento en la experiencia y la observación. Así, el modelo de machine learning se entrena y evalúa a partir de datos empíricos, específicamente un corpus de texto etiquetado según la complejidad de las palabras, lo cual permite al modelo discernir entre palabras complejas y no complejas.

2.1.2 Positivismo

Según Comte (1830), el positivismo destaca la observación empírica y el método científico como fundamentos del conocimiento. Este estudio adopta este enfoque al hacer uso de datos recolectados de manera sistemática, al formular hipótesis sobre el modelo de machine learning propuesto y la predicción binaria de la complejidad léxica de las palabras, al experimentar mediante el entrenamiento del modelo y al validar su rendimiento.

2.1.3 Racionalismo

En 1641, Descartes planteó la idea de que el conocimiento puede ser obtenido a través del razonamiento lógico y deductivo. En este estudio, el racionalismo se refleja en el desarrollo del modelo de machine learning, el cual se fundamenta en principios matemáticos y estadísticos. La estructura del modelo y la interpretación de los resultados son en base a la optimización algorítmica y el razonamiento lógico.

2.1.4 Pragmatismo

Peirce (1878) sostuvo la idea de que la verdad y el conocimiento deben evaluarse en función de su utilidad práctica y sus resultados. En esta investigación, el éxito del modelo se mide por su capacidad para identificar correctamente palabras complejas. La utilidad práctica del modelo en contextos reales, como por ejemplo en el campo de la educación, en el soporte a la comprensión lectora, en el soporte a la simplificación de textos o en el soporte a otras posibles aplicaciones, puede ser algo muy importante en la práctica.

La consideración conjunta de los marcos filosófico y epistemológico viene a dar consistencia a la investigación. El empirismo y el positivismo guían la recolección y el análisis de datos, el racionalismo apoya el diseño algorítmico y, por otro lado, el pragmatismo garantiza que el modelo tenga un uso práctico. Esto permite la comprensión del proceso de desarrollo y de evaluación del modelo de machine learning para la predicción binaria de la complejidad léxica de palabras, facilitando el manejo de los aspectos teóricos y también técnicos de manera coherente y además rigurosa.

2.2 Antecedentes de la investigación

En esta sección se presenta un análisis de investigaciones que han abordado la predicción de complejidad léxica, un componente fundamental para la simplificación textual y la mejora de accesibilidad lingüística. La predicción de complejidad léxica puede implementarse mediante dos enfoques principales: clasificación binaria (simple o complejo) o evaluación continua (en una escala gradual de complejidad).

Es importante destacar que, hasta mediados de 2024, según la literatura científica disponible, existe una asimetría en los recursos para diferentes idiomas. Para el español, se han desarrollado datasets que permiten la predicción de complejidad léxica binaria, pero aún no se cuenta con recursos anotados para la predicción en escala continua. En contraste, para el idioma inglés se dispone de ambos tipos de recursos, lo que ha permitido un avance más diversificado en las metodologías y aplicaciones para este idioma.

2.2.1 Predicción de la complejidad léxica en el idioma español

A continuación, se analizan trabajos que emplearon el dataset de la tarea compartida CWI 2018, agrupándolos según las características utilizadas en sus modelos y las arquitecturas implementadas. Investigaciones recientes en predicción de complejidad léxica para español han explorado diversas características lingüísticas utilizando este dataset.

Uso de características léxicas con algoritmos de machine Learning

En la tarea de predicción de la complejidad léxica de palabras se analizan los siguientes trabajos:

Las investigaciones de Kajiwara & Komachi (2018) establecieron una base con bosques aleatorios y características léxicas básicas como longitud de palabras y frecuencias en diversos corpus, alcanzando un F1 de 0.770. Posteriormente, Finnimore et al. (2019) lograron un avance significativo al expandir el conjunto a 25 características que operaban en múltiples niveles lingüísticos, incorporando innovadoramente tipos de entidades nombradas y probabilidades de unigramas mediante regresión logística, logrando un F1 de 0.776. Este modesto incremento evidenció una tendencia importante: la necesidad de características más sofisticadas para mejorar el rendimiento, sugiriendo que los enfoques tradicionales requerían complementarse con nuevas técnicas, lo que eventualmente conduciría hacia modelos híbridos que integraran embeddings con características lingüísticas convencionales.

Uso de embeddings como características del modelo con algoritmos de machine learning

A continuación, se presentan investigaciones que incorporaron embeddings en sus modelos de machine learning y que utilizaron el dataset de la tarea compartida CWI 2018:

La evolución de la investigación en complejidad léxica muestra un claro progreso metodológico. Aroyehun et al. (2018) establecieron una base comparando árboles de decisión con redes neuronales convolucionales, alcanzando un Macro-F1 de 0.767 mediante características

lingüísticas tradicionales y embeddings Word2Vec. Alarcon et al. (2019) exploraron un enfoque similar con SVM, obteniendo un F1-score de 0.7497, mientras Sheang (2019) logró un avance significativo (F1 de 0.7970) al optimizar redes convolucionales con características tradicionales y embeddings GloVe.

Alarcon et al. (2020) reorientaron la investigación hacia aplicaciones educativas para estudiantes con dificultades de aprendizaje, combinando corpus especializados con un diccionario de lectura fácil (E2R), embeddings (Word2Vec y BERT), alcanzando un F1-score de 0.794 con SVM. Los autores sugirieron usar recursos educativos específicos para mejorar la precisión en palabras técnicas. Alarcón et al. (2021) profundizaron esta línea en dos trabajos distintos: primero combinando características como longitud de palabras con embeddings y el diccionario E2R, manteniendo un F1-score de 0.794, y posteriormente comparando SVM tradicional (F1 de 0.792) con un innovador enfoque de reconocimiento de entidades nombradas (NER) basado en BERT (F1 de 0.694). Pimienta Castillo (2021) Pimienta exploró la regresión logística con embeddings BERT, características (tradicionales y basadas en la ley de Zipf), logrando un F1-score de 0.77.

Además, se presentan investigaciones que incorporaron diversos conjuntos de datos y embeddings en sus modelos de aprendizaje automático:

Ortiz-Zambrano et al. (2022) utilizaron el corpus CLexIS2 y experimentaron con embeddings de transformadores (BERT, XLM-RoBERTa y RoBERTa-large-BNE) combinados con características lingüísticas tradicionales, logrando resultados sobresalientes con los embeddings BERT ejecutándose con BERT afinado mediante Random Forest Regressor (MAE: 0.1592, correlación de Pearson: 0.9883). Posteriormente, Ortiz-Zambrano et al. (2023b) crearon LegalEc, un corpus de textos legales ecuatorianos, donde experimentaron con embeddings de XLM-RoBERTa y RoBERTa-BNE junto con características léxicas, utilizando un clasificador de red neuronal de dos capas densas con dropout y activación tanh. El mejor rendimiento se obtuvo con los embeddings de XLM-RoBERTa-large complementado con características lingüísticas (MAE: 0.1338), aunque con baja concordancia entre anotadores (Kappa: 0.13). Esta evolución muestra un enfoque hacia dominios más específicos (legal), manteniendo en ambos casos la estrategia de combinar embeddings de transformers con características lingüísticas tradicionales para optimizar el rendimiento.

Uso de Modelos del lenguaje grandes (LLMs)

Las investigaciones recientes en predicción de complejidad léxica binaria para español han explorado tanto algoritmos de machine learning tradicionales como modelos de lenguaje avanzados:

Anonymous (2024a) exploró el uso de modelos de lenguaje grandes (LLMs) para esta tarea, específicamente Llama 2 y ChatGPT-3.5 turbo. Lograron resultados notables. Particularmente, ChatGPT-3.5-turbo superó el rendimiento previo en español alcanzando un macro F1 de 0.781. Sin embargo, los autores señalaron que estos modelos presentaban dificultades para etiquetar correctamente frases muy complicadas, sugiriendo que, aunque los LLMs son efectivos para determinar el nivel de dificultad de las palabras en español, aún existen desafíos relacionados con las alucinaciones y ejemplos adversos que afectan su rendimiento.

2.2.2 Predicción de la complejidad léxica de palabras en el idioma inglés

Para el idioma inglés, hasta al año 2024 existen más estudios en este idioma tanto de predicción de la complejidad léxica continua como binaria en comparación con los del idioma español.

Los siguientes estudios tuvieron por objetivo la predicción de la complejidad léxica para el idioma inglés:

Gooding & Kochmar (2018) emplearon algoritmos clásicos (AdaBoost, Random Forest) con características lingüísticas tradicionales en tres datasets, alcanzando métricas F1 entre 0.8115-0.8736 mediante n-gramas, etiquetas POS, métricas léxicas y datos psicolingüísticos.

Bani Yaseen et al. (2021) trabajaron con el dataset CompLex anotado con escala Likert, implementando BERT y RoBERTa con técnicas de promediado, lo que les permitió alcanzar una correlación de Pearson de 0.788 y un error absoluto medio de 0.0609. Por su parte, Pan et al. (2021) utilizaron el mismo dataset en inglés, pero enfocándose en el ajuste fino de múltiples modelos (BERT, ALBERT, RoBERTa y ERNIE) combinados con estrategias como pseudo-etiquetado y aumento de datos; su apilación sobre estos modelos logró una correlación de Pearson de 0.7882, demostrando la superioridad de sus enfoques de conjunto para predecir la complejidad léxica en una escala continua.

Rozi et al. (2021) utilizó el dataset del SemEval 2021, con anotaciones de complejidad en escala Likert. Los autores experimentaron con varios modelos, destacando un LGBMRegressor que combinó embeddings GloVe (con PCA) y características tradicionales como longitud de palabra, número de sílabas, frecuencia en Wikipedia, y etiquetas POS. Este enfoque alcanzó un rendimiento de 0.7533 en correlación de Pearson. Aunque también probaron BERT, sorprendentemente no superó a los modelos basados en árboles de decisión. Los autores sugieren que futuros trabajos podrían beneficiarse de mejores técnicas de stacking y de embeddings contextuales más avanzados para capturar mejor los matices semánticos que influyen en la complejidad léxica.

Aziz et al. (2023) propusieron el modelo ITRM-LCP, que combina embeddings de BERT, RoBERTa, XLNet y DistilBERT con una capa BiLSTM, logrando mejoras significativas en la predicción de la complejidad léxica. Evaluado en los datasets CWI-2018 y CompLex, alcanzó incrementos de hasta 10.2%. Por otro lado, Alfter (2024) empleó un modelo RoBERTa-base con el dataset SemEval 2021 y entrenó su arquitectura durante 20 épocas, logrando un MSE de 0.0078 y R^2 de 0.524. Este autor enfatiza que los avances en embeddings contextuales y modelos de lenguaje pueden capturar mejor los matices semánticos, mejorando el rendimiento en la predicción de la complejidad léxica. Ambos enfoques destacan por su eficacia y por explorar distintas estrategias basadas en transformers para abordar la complejidad léxica en inglés. En contraste, Ortiz-Zambrano et al. (2023a) implementaron un enfoque híbrido que combina embeddings de transformadores (BERT y XLM-RoBERTa) con características lingüísticas tradicionales sobre CompLex 2.0, evaluando ocho algoritmos supervisados y logrando su mejor resultado con el SVM (MAE: 0.0688, correlación: 0.8911), demostrando la eficacia de integrar representaciones contextuales con características lingüísticas para la predicción de complejidad léxica en entornos con recursos limitados.

Predicción de la complejidad léxica de textos y oraciones para otros idiomas

La investigación sobre la complejidad léxica también ha sido abordada en otros idiomas, aunque en menor medida que en el idioma inglés, abordando las necesidades específicas de diversos idiomas con características lingüísticas únicas. Estos estudios no solo contribuyen a mejorar la accesibilidad textual para hablantes nativos con diferentes niveles de alfabetización, sino también para estudiantes de segundas lenguas y para optimizar sistemas de procesamiento de lenguaje natural en idiomas menos representados en la literatura científica.

Ivanov (2022) utilizó embeddings como fastText, representaciones contextuales de BERT y árboles de dependencia sintáctica para evaluar la complejidad de oraciones en ruso utilizando un corpus de libros de texto escolares, logrando capturar características semánticas y sintácticas profundas; su modelo BERT ajustado fino alcanzó una precisión de 0.869 en clasificación binaria, mientras que el modelo GNN también mostró un buen desempeño. Por otro lado, Nigusie y Tegegne (2023) aplicaron técnicas tradicionales como TFIDF y BOW con modelado bigram para clasificar la complejidad de textos en amárico, sin recurrir a embeddings, y obtuvieron su mejor resultado con un modelo SVM basado en BOW, logrando un accuracy de 0.871. Aunque ambos estudios buscan clasificar la complejidad textual, el primero utilizó embeddings, mientras que el segundo optó por métodos más tradicionales.

2.2.3 Vacíos de la literatura revisada, justificación de la originalidad y contribución de la tesis.

A continuación, se discuten los vacíos de conocimiento, fortalezas y debilidades detectadas en los antecedentes presentados anteriormente.

La revisión de la literatura revela una clara tendencia metodológica en la predicción de complejidad léxica: los enfoques híbridos que combinan características lingüísticas tradicionales con embeddings han demostrado consistentemente mejor rendimiento que los modelos basados exclusivamente en embeddings o solo en características tradicionales. Esta tendencia se evidencia en trabajos como el de Sheang (2019), quien logró un F1-score de 0.7970 combinando GloVe con características morfológicas, y Ortiz-Zambrano et al. (2022), que alcanzó una correlación de Pearson de 0.9883 integrando BERT con características lingüísticas tradicionales. Ortiz-Zambrano et al. (2023a) señalaron que este enfoque híbrido resulta especialmente adecuado para entornos con recursos computacionales limitados, característica fundamental para su aplicación en instituciones educativas secundarias del distrito de Cajamarca, donde existen estas restricciones tecnológicas.

Por otro lado, hay un vacío significativo en la aplicación de embeddings de última generación en contextos educativos específicos con recursos limitados. Aunque Anonymous (2024a) reporta resultados prometedores utilizando ChatGPT-3.5 para la clasificación de textos en español, aún no se han explorado embeddings avanzados desarrollados por OpenAI u otras arquitecturas recientes en este ámbito. Alfter (2024) y Rozi et al. (2021) coinciden en que los embeddings contextuales de nueva generación poseen un potencial considerable para mejorar la predicción de la

complejidad léxica, ya que permiten capturar con mayor precisión los matices semánticos y contextuales del lenguaje. Sin embargo, hasta el momento, no se han realizado estudios enfocados en la predicción de complejidad léxica en el contexto peruano, lo que implica que estas recomendaciones aún no se han implementado. Esta carencia es aún más relevante si se considera la realidad de los entornos educativos regionales del Perú, donde las limitaciones computacionales exigen el desarrollo de soluciones eficientes y adaptadas a recursos tecnológicos restringidos, un aspecto que sigue pendiente de atención en la literatura actual.

La originalidad de la presente investigación radica en abordar esta brecha mediante la implementación de un modelo híbrido que combina características lingüísticas con embeddings de nueva generación de OpenAI, evaluado en el contexto educativo y teniendo en cuenta las restricciones computacionales de instituciones secundarias de un distrito de Cajamarca. A diferencia de trabajos previos como los de Pimienta Castillo (2021) y Aroyehun et al. (2018), que demostraron la efectividad de combinar embeddings con características lingüísticas en contextos generales, este estudio se enfoca en desarrollar un sistema que pueda funcionar con los limitados recursos tecnológicos disponibles en las instituciones educativas mencionadas. Esta aproximación responde directamente a la necesidad identificada por investigadores como Ortiz-Zambrano et al. (2023a) de explorar modelos avanzados adaptados a contextos con recursos limitados, aplicándola a un entorno educativo regional donde puede tener un impacto directo en la mejora de materiales didácticos y comprensión lectora, sin requerir infraestructura computacional sofisticada.

2.3 Bases teóricas

2.3.1 Machine learning

Alpaydin (2020) afirma que el aprendizaje automático empieza con un modelo que presenta una gran variedad de parámetros susceptibles de ser ajustados. De hecho, el proceso de entrenamiento en sí mismo consiste en modificar los parámetros de la forma más adecuada para que el modelo se adapte a los datos que observa. Y el resultado de este proceso de ajuste es la especialización de un modelo general en una tarea de los datos. De este modo, al final del proceso de entrenamiento, obtenemos un algoritmo que está especializado para una tarea a partir de los datos. Una de las grandes ventajas del aprendizaje automático es que es capaz de encontrar patrones en los datos, patrones que permiten no sólo la interpretación de los procesos responsables de la producción de los datos, sino que, además, posibilitan hacer predicciones. Suponiendo que el futuro se parezca al pasado, las predicciones basadas en datos como patrones desempeñarán dicha función correctamente.

2.3.1.1 Modelo de machine learning

Un algoritmo de aprendizaje automático es un conjunto de instrucciones que indican a la computadora cómo aprender patrones de los datos que son parte del entrenamiento del modelo, por otro lado, un modelo de aprendizaje automático es el producto que se obtiene después de aplicar este algoritmo a dicho conjunto de datos. Dicho de otra manera, el modelo es lo que utilizamos para predecir a partir de nuevos datos tomando como base lo que se ha aprendido del conjunto de datos de entrenamiento dado (Burkov, 2020).

2.3.1.2 Algoritmo de machine learning

De acuerdo con el paper de Ray (2019), un algoritmo de machine learning es un conjunto de instrucciones computacionales diseñadas para que un programa aprenda a realizar tareas a partir de los datos, por lo que funciona aprendiendo patrones, haciendo predicciones, mejorando su precisión con el tiempo. Estos se utilizan en diversos campos, desde el reconocimiento de patrones hasta el diagnóstico médico. Su aplicación es primordialmente en tres tipos de problemas: la clasificación, la regresión y el agrupamiento.

Además, Ray (2019) también menciona que, por ejemplo, para la regresión, se emplean la regresión lineal, las máquinas de soporte vectorial, el descenso de la gradiente, árboles de decisión y k-vecinos más cercanos. Para la clasificación, se incluyen regresión logística, máquinas de soporte vectorial, k-vecinos más cercanos, árboles de decisión, redes neuronales artificiales y naïve bayes. Para agrupamiento, se emplea el algoritmo de agrupamiento de k-medias. Además, dependiendo de la disponibilidad y tipo de datos de entrenamiento, se pueden aplicar enfoques de aprendizaje supervisado, no supervisado, semisupervisado y aprendizaje reforzado.

2.3.1.3 Variables de modelos de machine learning

Parámetros

Como sostienen Chollet y Allaire (2018), en machine learning, las variables son las características o atributos específicos de los datos que se utilizan para entrenar al modelo. El modelo aprende la relación entre estas variables y el resultado que estamos tratando de predecir, basándose en muchos ejemplos de datos. Este proceso es una forma de programación en la que la computadora “aprende” cómo realizar una tarea a partir de ejemplos, en lugar de seguir instrucciones explícitas paso a paso.

Hiperparámetros

Los hiperparámetros son importantes para alterar la forma de trabajar de los algoritmos de aprendizaje automático e influir en la eficacia del modelo para mejorar su rendimiento. No son parte de los datos de entrenamiento, de modo que, tampoco podemos llegar a determinar los mismos automáticamente a partir de los datos de entrenamiento. Algunos ejemplos de hiperparámetros son la profundidad máxima que puede alcanzar un árbol en el algoritmo de árbol de decisión, la penalidad que hemos asociado con la clasificación incorrecta en el caso de las máquinas de vectores de soporte, el número k en los k vecinos más cercanos, la dimensión objetivo en la reducción de la dimensionalidad y el método de selección con el que vamos a tratar los datos que faltan (Burkov, 2020).

2.3.1.4 Arquitectura de modelo de machine learning

Las arquitecturas de machine learning (o de aprendizaje automático) hacen referencia al enfoque global o al marco de trabajo general utilizado para hacer frente a un problema o tarea en específico, proporcionando los parámetros para trabajar en ello, tales como las estructuras y las configuraciones de las redes, los cuales pueden contener el número, el

tamaño y el tipo de capas de las redes neuronales. Dicho de otra forma, las arquitecturas en machine learning son como los esquemas o los planes de cómo se construirán y organizarán los modelos de aprendizaje automático (Johnson, 2020).

Redes neuronales y deep learning

Tal como señala aws (s. f.), las redes neuronales se asemejan a una red de puntos que se conectan y trabajan juntos con la finalidad de analizar los datos y aprender en los casos de errores que se puedan presentar. Las redes neuronales que pueden procesar los datos en un solo sentido, desde un punto de entrada hasta un punto de salida, en las cuales comúnmente solo tienen una capa oculta, son redes neuronales simples que también se las conocen como redes prealimentadas, las cuales son más sencillas y se entrenan rápidamente, pero su capacidad de análisis es limitada. Por el contrario, el deep learning o aprendizaje profundo, que consiste en una subcategoría de la inteligencia artificial, permite a las computadoras procesar los datos de una forma muy parecida a lo que lo hace el cerebro humano, reconociendo patrones para producir información precisa. Consiste en una versión más sofisticada de las redes neuronales, las cuales contienen múltiples capas ocultas. Dichos sistemas pueden ser básicamente de dos tipos: redes neuronales convolucionales y redes neuronales recurrentes, cada una con su propia estructura y usos. Estos sistemas son más complicados, necesitan más recursos para su entrenamiento y pueden manejar tareas más intrincadas. Aunque los términos a menudo se usan como sinónimos, existen diferencias notables entre el aprendizaje profundo y las redes neuronales.

Como afirma (Kelleher, 2019), el aprendizaje profundo, una rama especializada de la inteligencia artificial, destaca por tomar decisiones precisas aprovechando grandes modelos de redes neuronales, especialmente cuando se trata de conjuntos de datos grandes y complejos. Se utiliza ampliamente en empresas en línea, tecnologías de consumo de alto nivel y atención médica. Google, Baidu o Microsoft lo utilizan, por ejemplo, en la búsqueda de imágenes e incluso en la traducción automática. Es la tecnología por defecto de los teléfonos inteligentes modernos para el reconocimiento de voz y la detección de rostros en las cámaras digitales. En el ámbito sanitario, se utiliza también para procesar imágenes médicas, así como para diagnosticar condiciones de salud. Los automóviles autónomos utilizan el aprendizaje profundo también para la localización, el mapeo, la planificación del movimiento, la conducción, la percepción del entorno y la estimación del estado de los conductores.

A diferencia de los modelos de “deep learning”, donde la mayoría de los parámetros del modelo se aprenden a través de las salidas de las capas anteriores y no directamente de las características de los ejemplos de entrenamiento, los modelos superficiales (“shallow”) ajustan sus parámetros directamente a partir de los datos de entrada. Los modelos de aprendizaje automático tradicionales son generalmente superficiales, pero los modelos de deep learning, que utilizan redes neuronales con múltiples capas ocultas, desarrollan una comprensión más compleja al construir conocimiento en niveles progresivos de abstracción (Burkov, 2020).

Transformers

La arquitectura del transformador, basada en el diseño de codificador-decodificador, es ampliamente utilizada para tareas como la traducción automática. Esta arquitectura convierte una secuencia de entrada de tokens en vectores embebidos y luego utiliza estos vectores para generar una secuencia de salida de tokens. Aunque inicialmente diseñado para tareas de secuencia a secuencia, los bloques del codificador y decodificador se han adaptado como modelos independientes. Existen tres tipos principales: solo codificador (como BERT), que convierte el texto en una representación numérica rica para tareas como clasificación de texto; solo decodificador (como GPT), que autocompleta el texto basándose en un contexto dado; y codificador-decodificador, que se utiliza para mapear completamente una secuencia de texto a otra, siendo útil para tareas como la traducción automática y la generación de resúmenes (Tunstall et al., 2022).

2.3.1.5 Tipos aprendizaje en machine learning

Aprendizaje supervisado

El aprendizaje supervisado, según Géron (2022), es un tipo de machine learning que se basa en el uso de datos de entrenamiento con respuestas etiquetadas. Un ejemplo habitual de esto sería la clasificación, que puede implementarse en un filtro para reducir el correo no deseado (lo que se conoce como spam) que se entrena con correos electrónicos previamente clasificados (correo spam, correo no spam). Otro tipo de aplicación habitual es la regresión, que permite predecir cuánto vale un carro, dadas diferentes características de los carros para los cuales se quiere predecir su valor, tales como el kilometraje, la antigüedad, la marca, entre otras. Cabe recalcar que algunos modelos pueden ser utilizados tanto como para problemas de clasificación como para problemas de regresión. Un ejemplo es la regresión logística, que permite la clasificación y la regresión que puede calcular un valor que indica la probabilidad de que una observación pertenezca a una clase específica.

Aprendizaje no supervisado

Según Jo (2021), el aprendizaje no supervisado es un método de machine learning donde los datos de entrenamiento no están etiquetados. El sistema se encarga de encontrar patrones y relaciones por sí mismo. En lugar de enseñarle al sistema lo que debe buscar (como en el aprendizaje supervisado), el sistema explora los datos y encuentra conexiones. Un ejemplo común es la agrupación de datos similares en grupos, llamado clustering. El sistema decide cuántos grupos hay, asigna datos a cada grupo según su similitud y ajusta los grupos hasta que los datos dentro de cada uno son lo más similares posible entre sí. Esta técnica se ha utilizado con éxito en tareas como agrupar textos similares. Además, puede ser adaptada para usarse en situaciones de aprendizaje supervisado.

Aprendizaje semisupervisado

Zhu y Goldberg (2009) indican que el aprendizaje semisupervisado es una técnica de machine learning que combina elementos del aprendizaje supervisado y no supervisado. Incluye la clasificación semisupervisada, donde se utilizan datos etiquetados y no etiquetados para entrenar un modelo que es más eficaz que uno entrenado solo con datos etiquetados. También incluye el agrupamiento restringido, que mejora el agrupamiento no

supervisado al agregar información adicional sobre los grupos. Este método es especialmente útil cuando hay pocos datos etiquetados disponibles o cuando es difícil obtenerlos. Además, tiene un gran valor tanto práctico para mejorar los algoritmos como teórico para entender cómo aprenden las máquinas y los humanos.

Aprendizaje autosupervisado

De acuerdo con IBM (2023), el aprendizaje autosupervisado es una técnica avanzada de machine learning que aprende de grandes cantidades de datos no etiquetados, en lugar de depender de datos etiquetados. Este tipo de aprendizaje ha sido el motor de descubrimientos importantes en el procesamiento de la lengua natural y de la visión por computadora. A diferencia del aprendizaje supervisado, este tipo de aprendizaje estima los resultados en función de una verdad general que se deriva de los datos no etiquetados del entrenamiento.

Aprendizaje reforzado

El aprendizaje por refuerzo (RL, por sus siglas en inglés) es una rama del machine learning inspirada en cómo los humanos aprenden mediante la interacción con su entorno. En RL, un agente aprende a tomar decisiones en una serie de pasos para lograr un objetivo, basándose en su experiencia y teniendo en cuenta la incertidumbre del entorno. A diferencia de otros métodos de aprendizaje, en RL no se necesitan ejemplos predefinidos de buenas decisiones, el agente aprende por sí mismo qué acciones conducen a los mejores resultados a través de un proceso de prueba y error. Este conjunto de decisiones se llama política. El RL ha demostrado su eficacia en una variedad de áreas, desde juegos hasta robótica, enseñando a las máquinas a realizar tareas complejas (Palmas et al., 2020).

2.3.1.6 Métricas de evaluación de machine learning

Las distintas métricas de evaluación son herramientas fundamentales en el aprendizaje automático y hay que seleccionar la adecuada a la tarea específica. Es decir, podemos clasificar, hacer predicciones, ranking entre otras tareas. Existen métricas de evaluación como la precisión y el recall que pueden servir para diferentes tareas. Zheng (2015), además, hace una revisión de las métricas más generales relacionadas con la evaluación de machine learning y da las siguientes definiciones:

1. **Métricas de clasificación.** La clasificación consiste, en predicciones de categorías a partir de datos, del tipo binario o multiclase, las métricas que se podrían utilizar por ejemplo para evaluar el rendimiento en la clasificación son: la métrica “accuracy”, que nos informa de cuántas veces ha predicho correctamente el clasificador y la matriz de confusión, que nos da el detalle de predicciones correctas e incorrectas para cada categoría.
2. **Métricas de rendimiento.** Las métricas de clasificación/ranking, tales como “accuracy” o “recall”, son imprescindibles para tareas que requieren de clasificación o ranking de elementos, tales como búsquedas a través de internet, la recomendación personalizada, entre otras tareas, donde los elementos se ordenan por relevancia o se les asigna una puntuación. El “accuracy” determina hasta qué punto el modelo realiza predicciones correctas, lo que implica una medición de todos los ejemplos de entrada. En el caso del “recall”, la métrica refiere a los casos relevantes que el modelo

determina como tales. Para facilitar la comparación de información entre el accuracy y el recall, se hace uso del “F1 Score”, que es una métrica que combina ambas informaciones mediante una media armónica, siendo bajo el F1 Score cuando bajo el accuracy o el recall son bajos; además, también hay que destacar que en la tarea CWI Shared Task 2018 se hizo uso del accuracy junto el macro promedio de F1 para valorar las predicciones de los modelos construidos por los equipos participantes (Yimam et al., 2018). Como indica Evidently AI (s. f.) la métrica accuracy es la siguiente:

$$\text{accuracy} = \frac{\# \text{ predicciones correctas}}{\# \text{ todas las predicciones}}$$

Además, es común emplear las puntuaciones medias de F1 para evaluar problemas de 2 o más clases. Un ejemplo de ello es la métrica que se utilizó en la competencia de CWI 2018 (Yimam et al., 2018) para evaluar los modelos de los equipos, clasificándolos en un ranking de mayor a menor, siendo el mayor el primer puesto de dicha competencia. A continuación, de acuerdo con Takahashi et al. (2022), se muestran las fórmulas del F1 score dentro de cada clase (F_{1i}), así como del Macro F1.

$$F_{1i} = 2 \frac{P_i \times R_i}{P_i + R_i}$$

Donde P_i y R_i son la precisión y el recall de la clase i.

$$\text{maF}_{1i} = \frac{1}{r} \sum_{r=1}^r F_{1i}$$

3. **Métricas de regresión.** En las tareas de regresión, donde el modelo aprende a predecir valores numéricos como el precio futuro de una acción, se utiliza comúnmente el RMSE para medir el rendimiento del modelo. El RMSE representa la diferencia promedio entre las predicciones del modelo y los valores reales, dando más peso a las grandes diferencias, y puede interpretarse como la distancia promedio entre las predicciones y los resultados reales.

2.3.1.7 Procesamiento del lenguaje natural (NLP)

De acuerdo con Patel (2021), el texto define el Procesamiento del Lenguaje Natural (NLP) como un subcampo de la lingüística, la informática, la ingeniería de la información y la inteligencia artificial que se ocupa de las interacciones entre las computadoras y los lenguajes humanos (naturales). El NLP implica programar computadoras para procesar y analizar grandes cantidades de datos de lenguaje natural, lo que puede incluir texto, voz y audio.

Sowmya V. B. et al. (2020) identifica las principales tareas del procesamiento de lenguaje natural, las cuales son:

1. Modelado del lenguaje: Estimar la palabra que sigue en un texto basándose en las palabras previas. Crucial para funciones como reconocer voz o traducir entre idiomas.
2. Clasificación de texto: Categorizar escritos según su contenido, aplicable en funciones como detectar correo no deseado o evaluar emociones.
3. Extracción de información: Recopilar elementos concretos de información de textos, tales como fechas importantes o identificar personas en posts de redes sociales.
4. Recuperación de información: Encontrar textos que se ajusten a una consulta específica, como lo que hace un buscador web.
5. Agente conversacional: Implementar sistemas que puedan interactuar en diálogos con personas.
6. Respuesta a preguntas: Construir sistemas que automáticamente proporcionen respuestas a interrogantes en lenguaje humano.
7. Traducción automática: Transformar textos de un lenguaje a otro, como lo hace un servicio de traducción en línea.
8. Modelado de temas: Analizar y determinar los temas predominantes en grandes volúmenes de documentos, útil en análisis de texto en múltiples campos.

Pasos frecuentes en NLP

Ghosh y Gunning (2019) señalan que los pasos utilizados frecuentemente en NLP son:

1. Tokenización: Es como dividir una oración en palabras individuales. Por ejemplo, "Estoy leyendo un libro" se convierte en "Estoy", "leyendo", "un", "libro". A veces, se toman dos o tres palabras juntas.
2. Etiquetado PoS: Es como identificar la función de cada palabra en una oración, como si es un sustantivo, un verbo, un adjetivo, etc. Por ejemplo, en "El cielo es azul", "El" es un determinante, "cielo" es un sustantivo, "es" es un verbo y "azul" es un adjetivo.
3. Eliminación de palabras vacías: Las palabras vacías son palabras comunes que se usan para construir oraciones, pero no afectan su significado. Ejemplos son "un", "estoy", "el". Se eliminan porque aparecen con mucha frecuencia y no aportan mucho al sentido de la oración.
4. Normalización de texto: Algunas palabras se deletrean, pronuncian y representan de diferentes maneras, pero significan lo mismo, como "Mumbai" y "Bombay" o "US" y "United States". La normalización de texto convierte estas variaciones en una forma estándar. También convierte las diferentes formas de una palabra en su forma base, como "hace" y "haciendo" a "hacer".
5. Corrección de ortografía: Es una tarea importante en cualquier proyecto de procesamiento del lenguaje natural. Aunque puede ser un proceso que consume tiempo, es necesario para no perder información relevante.
6. Derivación (Stemming): Es un proceso que transforma las palabras a su forma base. Por ejemplo, "producción" y "productos" se convierten en "producto". Esto es útil porque estas palabras tienen el mismo significado básico.
7. Lematización: A veces, la derivación puede dar resultados incorrectos. Por ejemplo, "luchando" podría convertirse en "luch", que no tiene sentido. La lematización soluciona este problema al revisar el diccionario para encontrar la forma base de una palabra. Sin embargo, este proceso es más lento.
8. Reconocimiento de Entidades Nombradas (NER): Este proceso identifica entidades nombradas, como nombres propios, y las categoriza. Por ejemplo, puede identificar

nombres de personas, lugares, etc. Las entidades nombradas usualmente no están en los diccionarios, por lo que deben ser tratadas de manera diferente.

9. Desambiguación del sentido de las palabras: Algunas palabras pueden tener diferentes significados dependiendo del contexto en el que se usen. La desambiguación del sentido de las palabras es el proceso de determinar el significado correcto de una palabra en su contexto. Esto es importante para tratar palabras con múltiples significados como entidades diferentes durante el análisis.

10. Detección de límites de oraciones: Este proceso identifica dónde termina una oración y comienza otra. Esto puede ser más complicado de lo que parece, ya que no todos los puntos denotan el final de una oración (por ejemplo, en abreviaturas). Detectar los límites de las oraciones es esencial para realizar análisis a nivel de oración.

2.3.2 Complejidad léxica de palabras

En los últimos años ha crecido el interés en investigar cómo detectar palabras complejas para su uso en la simplificación automática de textos (G. Paetzold & Specia, 2016; Shardlow et al., 2022). Se entiende por palabras complejas aquellas que son difíciles de identificar, comprender o pronunciar y que pueden obstaculizar de manera significativa la comprensión de un texto. Por esta razón, la complejidad léxica percibida se refiere al grado de dificultad que una palabra determinada implica para un individuo o grupo en particular (North & Zampieri, 2023).

Por otro lado, según Almeida et al. (2021), la complejidad léxica desempeña un papel fundamental en la comprensión lectora. Predecir la complejidad léxica de las palabras dentro de oraciones y en su contexto específico permite que los sistemas puedan mejorar en tareas de procesamiento de lenguaje natural, como la simplificación de textos, beneficiando especialmente a aquellos lectores con mayores dificultades en el idioma objetivo.

A continuación, se presentan conceptos relacionados a la complejidad de palabras:

2.3.2.1 Complejidad absoluta

Bulté y Housen (2012) y Paetzold y Specia (2016) señalan que las características lingüísticas incluidas dentro de la complejidad absoluta de palabras comprenden elementos morfosintácticos, semánticos y fonológicos que contribuyen a que una palabra se perciba como complicada, avanzada o intrincada en contraste con otra palabra.

Dentro de la teoría lingüística la gramática es una parte fundamental.

La morfología examina la estructura interna de las palabras, analizando cómo se forman a partir de morfemas (unidades mínimas con significado). Este análisis revela procesos de flexión, derivación y composición; por ejemplo, en "habilidades" se identifican tres componentes: el radical "habil-" (núcleo semántico), el derivativo "-idad" (transformador de categoría) y el flexivo "-es" (marcador de número) (RAE, s. f.-a).

Por su parte, la sintaxis estudia cómo se combinan las palabras para formar grupos y oraciones, siendo una disciplina de naturaleza combinatoria que analiza las relaciones dentro de todos los

grupos sintácticos. Esta rama examina las diversas funciones y relaciones de dependencia entre elementos, incluyendo el significado de las unidades constituidas como parte integral de su estudio (RAE, s. f.-b).

En cuanto a la semántica, esta disciplina analiza el significado de las expresiones lingüísticas con una naturaleza transversal a toda la gramática. Se divide en semántica léxica, que estudia el significado de palabras individuales y sus relaciones (sinonimia, antonimia, polisemia), y semántica composicional, que examina cómo los significados se integran en unidades mayores. Actualmente, se reconoce que el significado es fundamental para el análisis gramatical, destacándose el desarrollo de la semántica formal que utiliza lenguajes formalizados para su estudio (RAE, s. f.-f).

Finalmente, la fonología investiga la organización sistemática de los sonidos del habla, complementando a la fonética que se centra en sus aspectos físicos. Trabaja con fonemas como unidades abstractas contrastivas que distinguen significados (evidenciado en pares mínimos como /"kaso/-/"baso/), y enfatiza los rasgos distintivos cuya alteración (como la sonoridad en el contraste /p/-/b/) genera diferencias significativas entre palabras (RAE, s. f.-d).

La complejidad absoluta de una palabra se refiere a las características objetivas e intrínsecas, independientemente del usuario o contexto, evaluándose mediante diversas métricas. Entre estas pueden incluir la longitud de la palabra (como "mesa" con 4 letras y 2 sílabas frente a "inconstitucionalidad" con 20 letras y 8 sílabas), la frecuencia léxica (palabras comunes como "casa" versus términos poco frecuentes como "inefable"), la complejidad morfológica (considerando palabras compuestas como "pararrayos" o derivadas como "felicidad"), la irregularidad ortográfica o fonológica que mide cuánto se desvía una palabra de los patrones regulares o predecibles de escritura y pronunciación, etc.

2.3.2.2 Complejidad relativa

Como señalan Bulté y Housen (2012), la complejidad lingüística relativa de una palabra se define por la experiencia personal o los factores psicolingüísticos de la persona. Aspectos tales como la carga cognitiva, la dificultad para aprenderla, así como cuán familiarizado está un individuo con una palabra o su tipografía, pueden influir en el grado de complejidad relativa que se le atribuye a una palabra.

Esta conceptualización de la complejidad relativa encuentra respaldo en investigaciones psicolingüísticas sustanciales. Perfetti (2007) observa que la calidad de las representaciones léxicas varía entre individuos y dentro del léxico de un mismo individuo, mientras que Ellis (2002), enfatiza que la adquisición de vocabulario está determinada por la contingencia entre forma y significado, moldeada por la atención del aprendiz y potenciada por la frecuencia de exposición.

Además, como señalan Bulté & Housen (2012), estudios psicolingüísticos han encontrado que ciertas estructuras incrustadas (por ejemplo, cláusulas relativas) y pasivas son más difíciles de procesar, o emergen más tarde en la adquisición del lenguaje, que otras estructuras (por ejemplo, estructuras coordinadas y activas) (Byrnes & Sinicropo, 2008; Diessel, 2004). La dificultad de una característica lingüística es hasta cierto punto subjetiva o dependiente del aprendiz. Una

característica lingüística que es costosa, difícil o complicada para algunos aprendices o usuarios puede ser menos costosa, menos difícil o incluso fácil para otros aprendices, dependiendo de factores de individualidad como su aptitud lingüística, capacidad de memoria, antecedentes de su lengua nativa, motivación, etcétera.

Dentro de la perspectiva de la teoría psicopedagógica se conceptualiza la complejidad relativa desde el punto de vista de quienes utilizan el idioma: un elemento o conjunto de elementos del lenguaje se percibe como complejo cuando representa un costo cognitivo significativo para estudiantes y hablantes, especialmente considerando el esfuerzo mental y los recursos cognitivos necesarios para su procesamiento o asimilación. En este sentido, la complejidad relativa (también denominada complejidad cognitiva o, más sencillamente, dificultad) hace referencia al grado de esfuerzo mental requerido para aprender, procesar o expresar determinados componentes lingüísticos durante el aprendizaje y la comunicación en un idioma (Hulstijn & Graaff, 1994).

La complejidad relativa de las palabras considera los aspectos subjetivos y contextuales del léxico, variando según el usuario y la situación específica, mediante diversas métricas y metodologías. Estas incluyen medidas de dificultad percibidas y obtenidas a través de cuestionarios donde los aprendices evalúan la complejidad de las palabras; evaluaciones de carga cognitiva mediante experimentos que miden tiempos de reacción en tareas de reconocimiento o memorización; índices de adquisición que documentan la edad o secuencia en que las palabras son típicamente aprendidas por niños y adultos; y medidas psicométricas que analizan la frecuencia de errores al usar determinadas palabras o el tiempo requerido para procesarlas en contextos experimentales controlados.

2.3.2.3 Tipos de predicción de la complejidad

De acuerdo con la literatura revisada, los tipos de predicción enfocados en la tarea de identificar la complejidad de palabras son:

Complejidad binaria

Bani Yaseen et al. (2021) indican que la predicción de la complejidad léxica de palabras en machine learning es binaria, si se la clasifica como compleja o no, lo que además North et al. (2023) señala que esto se denomina también Identificación de Palabras Complejas (CWI) y se refiere a la tarea de asignar a una palabra específica un valor binario de complejidad: 1 si la palabra es considerada compleja, o 0 si se considera no compleja (CWI binario).

North et al. (2023), por otro lado, menciona que además existe la identificación probabilística de palabras complejas, que es una extensión del enfoque binario, añadiendo un nivel de probabilidad a la clasificación binaria (CWI probabilístico), es decir, se utiliza un modelo de regresión para asignar valores continuos que indican la probabilidad de que una palabra sea compleja o no.

Complejidad continua

Shardlow et al. (2020) señalan que, al representar la complejidad en una escala continua, la predicción de Complejidad Lexical (LCP) ofrece una descripción más precisa de la complejidad

de una palabra en particular, ya que permite predecir más de dos grados de dificultad (por ejemplo, usando escala de Likert para la anotación).

Complejidad personalizada

Conforme a lo expresado por Lee y Yeung (2018), la mayor parte de los trabajos desarrollados han puesto en entredicho la estrategia de "talla única" usada en la predicción de la complejidad lexical, y hacen hincapié en que no representa de forma adecuada las diferencias en la forma como se percibe la complejidad léxica. En respuesta a ello, ha habido estudios de identificación personalizada de palabras complejas. Dicho método presenta dichos modelos personalizados que se adecúan mejor al usuario individual o a un grupo demográfico determinado, e integra o aprende características demográficas del usuario para hacer predicciones a medida. Estos sistemas, que toman en cuenta dichos aspectos han demostrado ser más efectivos que los sistemas genéricos al predecir las complejidades léxicas de un grupo demográfico específico. Además, se ha comprobado que los sistemas personalizados son más eficientes que sus equivalentes no personalizados. Por lo tanto, la complejidad personalizada emerge como un área de investigación con gran potencial en la predicción de complejidad, superando a los métodos más generalizados.

2.3.2.4 Predicción binaria de la complejidad léxica de palabras (Identificación de palabras complejas)

La predicción binaria de la complejidad léxica, también conocida como Identificación de Palabras Complejas (CWI) (Bani Yaseen et al., 2021), se centra en decidir qué términos en un texto dado pueden ser complicados para el lector de dicho texto. Este proceso puede formar parte de un proceso más amplio que es conocido como simplificación léxica, cuyo propósito es el de sustituir palabras y expresiones que son complicadas por alternativas que resulten más sencillas para el lector (Paetzold y Specia, 2016).

2.3.3 Simplificación léxica

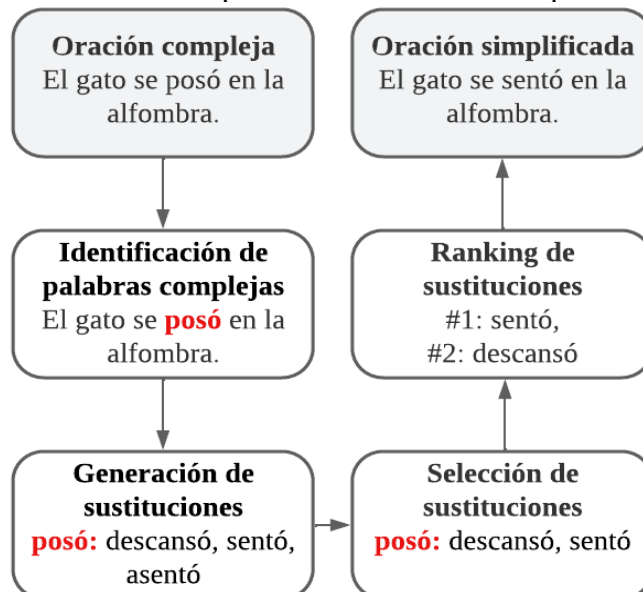
Paetzold y Specia (2017) describen la simplificación léxica como un proceso secuencial compuesto por 4 subprocesos interrelacionados:

1. Identificación de Palabras Complejas: Se identifican los términos que podrían ser difíciles para el lector objetivo.
2. Generación de Sustituciones: Consiste en encontrar alternativas más sencillas.
3. Selección de Sustituciones: Se eligen las palabras más adecuadas según el contexto.
4. Clasificación de Sustituciones: Se organizan las alternativas de acuerdo con su nivel de dificultad, priorizando las más simples.

En la figura 6 se muestra el proceso de simplificación léxica de 4 etapas.

Figura 6

Proceso de simplificación léxica de 4 etapas



Nota. Basado en “A Survey on Lexical Simplification”, Paetzold y Specia (2017).

2.3.4 Comprensión lectora

Para Snow (2010), la comprensión lectora se define como el proceso mediante el cual se construye y se extrae el significado de los textos. Por tanto, la comprensión lectora implica construir una representación mental de las ideas presentes en el texto escrito, la realización de inferencias a partir del texto y de la realidad en el mundo, así como llevar a cabo procesos más elaborados para una comprensión integral.

Oakhill et al. (2019) argumentan que la comprensión lectora se ve sometida a la influencia de múltiples factores como la habilidad para decodificar, la habilidad para reconocer las palabras escritas, el vocabulario, la habilidad para hacer inferencias, los conocimientos previos, el monitoreo de la comprensión y el conocimiento del tipo de estructura que presenta un texto. Todos estos factores interactúan con el objetivo de influir en la comprensión general de un texto. Van además más allá al indicar que existe una teoría que recibe el nombre de “Simple Vista de la Lectura” la cual fue creada para poder explicar cómo se produce la lectura y cómo enfrentarnos a las dificultades que puedan aparecer en la comprensión lectora.

De acuerdo con Kim (2023), la comprensión lectora implica percibir visualmente letras y palabras, convertirlas en sonidos del habla (decodificación), analizar información lingüística (la gramática, la sintaxis, el vocabulario y estructuras lingüísticas), desarrollar una comprensión inicial de las proposiciones en la lectura e integrarlas para una representación mental precisa y coherente.

Connor y Al’Otaiba (2008) también mencionan que el vocabulario es uno de los elementos más importantes de la comprensión lectora, ya que se podría tener las habilidades de

decodificación necesarias para leer una palabra, pero si la palabra no fuera parte de su léxico, es posible que no puedan asignar significado al texto decodificado. Con el propósito de poder llegar a comprender dicha significación del lenguaje escrito mediante los textos, el conocimiento del vocabulario es de suma importancia y es un previo requisito muy importante.

La Organización para la Cooperación y el Desarrollo Económico (OCDE) cuya definición de la comprensión lectora nos dice que es la habilidad de poder entender, utilizar, reflexionar e involucrarse con textos escritos para conseguir y alcanzar los objetivos individuales, para poder progresar en el conocimiento personal y para poder explotar el potencial de cada uno para ser partícipes de las actividades de la sociedad.

2.3.4.1 Dimensiones de la comprensión lectora

Según Monica Melby-Lervåg y Ame Lervåg, 2014, la comprensión lectora se sustenta en tres partes que son fundamentales: la decodificación, el vocabulario y la comprensión gramatical, fundamentales para entender textos en cualquier idioma. De forma similar, Hoover y Gough (1990), sostienen que la comprensión lectora se compone de dos elementos clave: la decodificación y la comprensión lingüística. Además, Oakhill et al. (2019), hicieron el concepto más amplio en cuanto a sus dimensiones, al realizar la identificación de un conjunto de habilidades y de estrategias que ayudan a la contribución de la comprensión de la lectura, que incluye la comprensión literal e inferencial, la metacompreensión y las habilidades de razonamiento y pensamiento crítico.

La comprensión lingüística

Según Rogde et al. (2019), la comprensión lingüística es el proceso de realizar la interpretación de la información de las palabras, de las oraciones y de los discursos. Por lo que se trata de la habilidad para entender el lenguaje hablado, que se mide comúnmente con la realización de pruebas de vocabulario o de comprensión auditiva. Por otro lado, se considera que además del vocabulario, la comprensión lingüística también está compuesta por múltiples habilidades que tienen que ver con la sintaxis del lenguaje, que viene dado por la capacidad para el entendimiento y la formulación de las oraciones, y a la morfología de las palabras, que tiene que ver con el proceso de creación de las palabras de un lenguaje dado.

CAPÍTULO 3: METODOLOGÍA

Este capítulo presenta los aspectos metodológicos centrales de la investigación, incluyendo: el tipo y diseño de estudio, la unidad de análisis, la población, la muestra y las técnicas de recolección de datos empleadas para el desarrollo del presente trabajo.

3.1 Tipo y diseño de la investigación

Esta investigación es de tipo cuantitativo con un diseño experimental.

La elección del enfoque cuantitativo se justifica por múltiples razones:

- **Medición objetiva del fenómeno:** La complejidad léxica requiere ser evaluada mediante indicadores precisos y medibles que permitan clasificaciones binarias (complejo/no complejo), lo que demanda un tratamiento numérico de los datos, tal como lo señalan Saunders et al. (2019).
- **Necesidad de validación estadística:** Para determinar la efectividad predictiva de los modelos de machine learning propuestos, la importancia de las variables evaluadas, así como la eficiencia en el uso de recursos de la memoria del modelo, es indispensable aplicar pruebas estadísticas rigurosas que cuantifiquen su precisión, sensibilidad y especificidad en la identificación de palabras complejas.
- **Comparabilidad de resultados:** El enfoque cuantitativo permite establecer comparaciones objetivas entre el rendimiento del modelo y los baselines establecidos, mediante métricas estandarizadas como Macro F1 (F1 macro promedio), precisión y recall.

Específicamente, se empleó un diseño experimental tal como lo sugiere Hernández Sampieri et al. (2014), donde se manipuló sistemáticamente las características de entrada (features) de los modelos de machine learning para evaluar su impacto en la precisión de la predicción binaria de la complejidad léxica del modelo. Este diseño permite establecer relaciones causales entre las características seleccionadas y la efectividad predictiva, controlando y seleccionando variables mediante técnicas de validación cruzada y conjuntos de prueba independientes.

3.2 Unidad de análisis

Es cada palabra en español que puede ser parte de una frase u oración a la que se le puede determinar un nivel binario de complejidad léxica.

3.3 Población de estudio

Las palabras en español, un conjunto amplio y dinámico que refleja la riqueza y complejidad del idioma. Según la última versión del Diccionario de la Real Academia Española RAE (s. f.-e), existen 93000 lemas, lo que indica que el número real de palabras supera esta cifra. Esta amplitud léxica se evidencia mejor en recursos como el Corpus del Español del Siglo XXI (CORPES XXI) de la RAE, que, en su versión más reciente, la versión 1.1, contiene más de 410 millones de formas ortográficas (RAE, 2024a).

La población está compuesta de todas las palabras en español que pueden ser parte de frases u oraciones a las que se les puede determinar un nivel de complejidad léxica binaria.

3.4 Tamaño de la muestra

Respecto al tamaño de la muestra, en la Tabla 1, se especifica el número de instancias, las fuentes de los conjuntos de datos y su uso en el desarrollo y evaluación del modelo de machine learning de este estudio.

Tabla 1

Conjuntos de datos y propósitos en el desarrollo del modelo de este estudio

# Instancias	Fuente del conjunto de datos	Propósito
13750	CWI Shared Task 2018 (train)	Entrenamiento
500	Textos escolares	Prueba en el contexto de la institución secundaria
Total: 14250 Instancias		

Nota. Se han seleccionado las 500 instancias como muestra representativa, siguiendo estudios previos sobre simplificación léxica que utilizaron la misma cantidad para pruebas (Alarcón et al., 2021).

En la sección 3.5 siguiente se presenta información más detallada de los datos presentados en la Tabla 1.

3.5 Selección de la muestra

La presente investigación se centra en el desarrollo de un modelo para la predicción binaria de la complejidad léxica de palabras en español, en el contexto de instituciones educativas secundarias de educación básica regular públicas en un distrito de Cajamarca.

Para ello, se siguió el enfoque de Yimam et al. (2018) para la selección de muestra: en un track de la competencia, los modelos se entrenaron utilizando datos de entrenamiento de un conjunto más amplio. Sin embargo, no se dispuso de datos de entrenamiento para el dominio específico, aunque sí se contó con datos de prueba para este dominio.

Siguiendo este enfoque, se utilizó el conjunto de datos en español del CWI Shared Task 2018 para entrenar el modelo de la presente investigación. Posteriormente, se evaluó en el contexto educativo específico utilizando el conjunto de prueba recopilado de los Textos Escolares de la institución educativa.

La selección de los datasets CWI Shared Task 2018 y Textos Escolares responde a criterios metodológicos específicos que garantizan la robustez del entrenamiento y evaluación del modelo:

CWI Shared Task 2018

- Estándar internacional: Constituye el dataset más reconocido para la tarea de la predicción binaria de la complejidad léxica de palabras en español (identificación de palabras complejas).
- Anotaciones de calidad: Contiene 13,750 instancias en español anotadas por 10-20 anotadores nativos y no nativos por palabra, lo que proporciona una medida de complejidad léxica con alta fiabilidad inter-anotador.

- Diversidad contextual: Incluye textos de Wikipedia y noticias, abarcando distintos registros lingüísticos y dominios temáticos que reflejan la variabilidad del lenguaje real.

Textos Escolares

- Relevancia educativa: Extraído directamente del material curricular utilizado en las instituciones públicas secundarias del distrito de Cajamarca, garantizando la pertinencia para el público objetivo final (los estudiantes).
- Gradación por niveles: Contiene textos de diferentes grados educativos (del primero al quinto de secundaria), lo que permite evaluar la capacidad del modelo para medir la complejidad en función de todos los grados de las instituciones secundarias mencionadas.
- Representatividad lingüística: Incorpora el español educativo peruano en la enseñanza pública de nivel secundario, asegurando que el modelo se evalúe en dicho contexto.

La combinación de ambos datasets permite un entrenamiento equilibrado: CWI Shared Task 2018 aporta rigor metodológico y comparabilidad internacional, mientras que Textos Escolares proporciona especificidad contextual y aplicabilidad directa al entorno educativo específico mencionado.

Esta selección se basa en una consideración importante: la escasez de datasets de palabras en español anotados con su complejidad léxica (Alarcón et al., 2021; Bott et al., 2024; Degraeuwe y Goethals, 2024; Ortiz-Zambrano et al., 2022, 2023b), especialmente en el contexto educativo peruano.

A continuación, se presenta más detalle acerca de esta selección.

3.5.1 Conjunto de entrenamiento del Dataset CWI Shared Task 2018

El dataset CWI Shared Task 2018 fue recolectado y etiquetado para ser usado en una competencia que se centró en la identificación de palabras complejas. Este sirvió como base en dicha tarea compartida que desafió a investigadores y programadores a desarrollar modelos de inteligencia artificial eficaces para identificar palabras complejas (*CWI Shared Task 2018 - Datasets*, s. f.).

Este dataset fue anotado por hablantes nativos de español, lo que proporciona una evaluación auténtica de la complejidad léxica. Su pertenencia a un dominio general permite una aplicabilidad más amplia que puede servir como punto de partida para futuras adaptaciones en diferentes contextos; además, es un recurso reconocido internacionalmente en el campo de la identificación de palabras complejas.

Dentro del ANEXO B, en la Tabla B-1 se presenta la distribución de entradas para los cuatro idiomas del dataset CWI Shared Task 2018, en la Tabla B-2 se muestra la estructura general de dicho dataset y en la Tabla B-3 se incluyen ejemplos de entradas de su subconjunto de entrenamiento en el idioma español.

3.5.2 Conjunto de prueba basado en textos escolares

Para evaluar las hipótesis y así validar la efectividad del modelo de machine learning desarrollado en este estudio de la predicción binaria de la complejidad léxica de palabras en instituciones secundarias de educación básica regular del distrito de Lajas, provincia de Chota, en el departamento de Cajamarca, se recolectó una muestra de 500 elementos, que incluye tanto palabras complejas como no complejas con su respectiva oración de contexto. Esta muestra se extrajo de los libros de textos escolares de Comunicación, desde primero hasta quinto grado de secundaria, que estuvieron disponibles en la biblioteca de la institución educativa “Anaximandro Vega” - Lajas, los cuales fueron: Comunicación 1: Secundaria (2016), Comunicación 2: Secundaria (2016), Comunicación 3: Educación Secundaria (2016), Comunicación 4: Educación Secundaria (2016) y Comunicación 5: Educación Secundaria (2016).

El tamaño correspondiente a 500 palabras es metodológicamente válido para evaluar modelos de predicción de complejidad léxica binaria, respaldado por estos corpus de referencia en orden cronológico:

- LSeval (De Belder & Moens, 2012): 430 instancias de clasificaciones de simplicidad en inglés.
- CW Corpus (Shardlow, 2013b): 731 oraciones de Wikipedia Simple English, cada una con una palabra simplificada. Que comúnmente son usados para el entrenamiento y la evaluación de sistemas de predicción de la complejidad léxica binaria.
- LexMTurk (Horn et al., 2014): 500 oraciones con una palabra objetivo simplificada. Comúnmente utilizado para el entrenamiento y evaluación de sistemas de Simplificación Léxica. Demostró que este tamaño muestral permite obtener datos confiables.
- NNSeval (G. H. Paetzold & Specia, 2016b): 239 instancias orientadas a hablantes no nativos de inglés.
- EASIER-500 (Alarcón et al., 2021): 500 instancias para español. Siendo el primer dataset de simplificación léxica para español publicado.
- ALEXSIS (Ferrés & Saggion, 2022): 381 instancias en español.

La consistencia de estos estudios en torno a muestras de 239-731 instancias confirma que 500 palabras constituyen una base representativa para esta tarea. Sin embargo, la cantidad podría mejorarse en futuros trabajos para obtener resultados aún más robustos y generalizables.

3.6 Técnicas de recolección de datos

En esta investigación, se empleó una combinación de métodos de recolección de datos secundarios y recolección documental. Se reutilizó el dataset CWI Shared Task 2018, siguiendo el método de recolección de datos secundarios descrito por Hernández Sampieri et al. (2014). Adicionalmente, se utilizó el método de recolección de datos de documentos, analizado por Creswell y Creswell (2018), mediante el análisis de los Textos Escolares de Comunicación, clasificados como documentos públicos. Este enfoque permitió identificar y clasificar palabras complejas y no

complejas en su contexto adaptándolo a las necesidades específicas de la investigación y recopilando la información en un dataset de validación del modelo de machine learning.

Criterio para etiquetado de palabras en el dataset de Textos Escolares

Para establecer un criterio objetivo de la complejidad léxica, se sigue un enfoque similar al propuesto por Mukherjee et al. (2016), basado en la presencia o ausencia de las palabras en los glosarios de los Textos Escolares de Comunicación mencionados.

- Siguiendo este criterio, se considera una palabra como compleja si se encuentra incluida en dichos glosarios. Este método parte de la premisa de que las palabras que requieren una explicación adicional o que pertenecen a un campo de conocimiento específico tienden a ser más complejas.
- Para el conjunto de palabras no complejas se seleccionan aleatoriamente palabras del texto que no estén en los glosarios y que no sean: artículos, preposiciones, conjunciones, pronombres personales, demostrativos y posesivos, adverbios comunes, verbos auxiliares en sus formas más frecuentes, números cardinales y ordinales básicos, e interjecciones comunes y monosílabos. Esto con la finalidad de poder incluir palabras no triviales dentro de esta selección.

Bajo este enfoque se seleccionaron 250 palabras complejas (50 por cada grado) y 250 palabras no complejas (50 por cada grado).

Los libros de Textos Escolares de Comunicación del 2016 del Ministerio de Educación constituyen un recurso valioso para la predicción binaria de la complejidad léxica de palabras. Estos materiales fueron desarrollados siguiendo el Currículo Nacional de la Educación Básica Regular Secundaria peruano y contienen glosarios elaborados por equipos editoriales y pedagógicos especializados (expertos) en el contexto educativo nacional.

Los términos incluidos en estos glosarios refleja un proceso deliberado de identificación de vocabulario que los expertos consideraron potencialmente desafiante para los estudiantes peruanos del nivel de educación secundaria. Esta selección incorpora implícitamente criterios pedagógicos adaptados a la realidad del Perú, considerando tanto zonas urbanas como rurales.

El uso de glosarios como método para identificar y etiquetar palabras complejas encuentra un sólido respaldo en las tendencias actuales de investigación como la Simplificación Automática de Textos. Particularmente relevante es la observación de que "el desarrollo y utilización de recursos lingüísticos, como glosarios, memorias de traducción y bases de datos terminológicas, son fundamentales para lograr este objetivo" (ACL, 2024) de simplificación de terminología compleja.

Los glosarios representan una herramienta metodológica privilegiada por varias razones:

- Validación experta: Los términos incluidos en glosarios han pasado por un proceso de selección deliberado por especialistas que identifican vocabulario potencialmente desafiante.

- Contextualización: Los glosarios escolares, en particular, están diseñados específicamente para abordar las necesidades léxicas de una población estudiantil peruana, considerando su nivel educativo y contexto sociocultural.
- Base para sistemas automáticos: Como recursos lingüísticos estructurados, los glosarios pueden integrarse eficientemente en sistemas de procesamiento de lenguaje natural para la identificación automática de términos complejos.

Esta aproximación basada en glosarios se alinea perfectamente con el objetivo de crear "herramientas más efectivas y recursos" para la simplificación textual, estableciendo un puente entre el conocimiento pedagógico especializado y las aplicaciones tecnológicas para la evaluación y simplificación de la complejidad léxica.

En la Tabla B – 4 del ANEXO B, se muestra una selección de palabras como parte de la muestra para el dataset de validación basado en Textos Escolares, el cual contiene sus respectivas oraciones y etiquetas binarias. La tabla está organizada de la siguiente manera: PALABRA, ORACIÓN y ETIQUETA BINARIA. Dicha tabla incluye 20 instancias de muestra, cada una con su respectiva oración y etiqueta binaria de complejidad.

CAPÍTULO 4: RESULTADOS Y DISCUSIÓN

Este capítulo aborda el desarrollo del modelo de machine learning para la predicción binaria de la complejidad léxica de palabras (identificación de palabras complejas) en instituciones educativas públicas de educación básica regular del nivel secundario en un distrito de Cajamarca, la presentación de resultados, pruebas de hipótesis y la discusión de resultados de este estudio.

4.1 Desarrollo del modelo de machine learning

Para desarrollar el modelo de machine learning de este estudio se usó Python 3. Asimismo, se empleó el valor 42 como semilla aleatoria en todas las etapas para favorecer su reproducibilidad.

4.1.1. Preprocesamiento de Datos e Ingeniería de Características

El preprocesamiento de datos es una etapa fundamental en el desarrollo de sistemas de procesamiento de lenguaje natural (PLN). Por otro lado, la ingeniería de características consiste en la extracción y transformación de información relevante para alimentar los modelos. A continuación, se describen las técnicas implementadas para cada recurso:

4.1.1.1 Embeddings de nueva generación

OPENAI (2024) ha presentado un nuevo tipo de embeddings que ofrecen representaciones vectoriales de conceptos en los textos. Estos nuevos modelos, text-embedding-3-small y text-embedding-3-large, proporcionan mejores resultados que sus versiones predecesoras y además, proporcionan una técnica que ayuda a optimizar el rendimiento con el costo de dichos embeddings de forma aceptable, es decir, disminuyendo el tamaño de los embeddings sin comprometer su capacidad de representación de estos.

En este estudio se utilizó la variante de los embeddings text-embedding-3-large empleando diversas dimensiones para obtener los embeddings sobre las palabras objetivo y sobre las frases que las contienen. Para este análisis, las consultas se efectuaron durante los meses de septiembre y octubre del 2024.

Preprocesamiento de datos

No se realizó ningún preprocesamiento para esta característica, ya que la API del modelo text-embedding-3-large no requería ninguna transformación previa ni para la palabra objetivo (target_word) ni para la oración (sentence) en la que se encuentra.

Ingeniería de características

1. Cálculo de embeddings:

Se utilizó la biblioteca cliente de OpenAI, versión 1.35.13 para Python, para realizar consultas a su API REST y generar embeddings de texto. Se empleó el modelo text-embedding-3-large, configurando distintas dimensiones (300, 400 y 500) tanto para la palabra objetivo (target_word) como para la oración que la contiene (sentence). Los embeddings generados son vectores numéricos que capturan la representación semántica del texto de entrada.

2. Procesamiento paralelo:

Se utiliza ThreadPoolExecutor para calcular los embeddings de manera eficiente en paralelo para cada fila del conjunto de datos.

4.1.1.2 Corpus CORPES

De acuerdo con la Real Academia Española RAE (2024a), el Corpus del Español del Siglo XXI (CORPES XXI) lo constituye una inmensa base de datos lingüística que se compone de textos escritos y hablados que están en el idioma castellano contemporáneo de diversas naciones, y que es usada como un recurso muy importante para muchos estudios lingüísticos. Una de las principales mejoras aplicadas en la versión 1.1, lanzada en abril de 2024, es un nuevo diccionario de frecuencias léxicas con 116707 lemas y 387429 formas RAE (2024b). También el corpus ha ido más allá de los 410 millones de formas, un aumento muy importante con respecto a los 395 millones de la versión 1.0 de mayo del 2023.

Preprocesamiento de datos

1. Carga y limpieza de datos:

Los datos se cargaron desde el archivo `diccionario_frecuencias_corpes_alfa.tsv` utilizando la biblioteca `pandas`, lo que permitió gestionar y manipular eficientemente el diccionario de frecuencias léxicas.

2. Normalización:

Para obtener las correspondencias para la extracción de la frecuencia de CORPES, se aplicaron los siguientes pasos tanto a las palabras objetivo (`target_word`) como a la palabra correspondiente en CORPES:

- Se eliminaron espacios en blanco al inicio y al final.
- Se convirtieron las palabras a minúsculas.

3. Mapeo de etiquetas gramaticales:

El diccionario `_POS_TO_CATEGORY` se utiliza para convertir las etiquetas gramaticales (POS) generadas por el modelo `spaCy` (`es_core_news_lg`, optimizado para CPU) en las categorías específicas que requiere el diccionario de frecuencias. Primero, se obtuvieron la etiqueta POS a partir de la palabra objetivo y de la oración en la que aparece. Luego, estas etiquetas se ajustaron usando `_POS_TO_CATEGORY` para que coincidan con las categorías gramaticales definidas en el diccionario de frecuencias.

```

_POS_TO_CATEGORY = {
    'ADJ': 'A', # Adjetivo
    'CONJ': 'C', # Conjunción
    'INTJ': 'I', # Interjección
    'PRON': 'L', # Pronombre personal
    'NUM': 'M', # Numeral
    'NOUN': 'N', # Sustantivo (común)
    'ADV': 'R', # Adverbio
    'VERB': 'V', # Verbo,
    'AUX': 'V', # Verbo auxiliar,
    'PROP': 'N', # Nombre propio
}

```

Ingeniería de características

1. Cálculo de frecuencias absolutas:

Con la palabra objetivo (`target_word`) y la categoría que se obtuvo en el preprocesamiento, se busca la frecuencia en el diccionario CORPES; si no se encuentra en la categoría específica, se intenta con una máscara que permite una búsqueda más general. Para la casuística de las frases, se consideró la frecuencia más baja de todas las palabras que las integran.

2. Procesamiento paralelo:

Para optimizar el rendimiento, se utiliza la clase `ThreadPoolExecutor` de la biblioteca `concurrent.futures`. Esto permite calcular las frecuencias absolutas de manera paralela para grandes volúmenes de datos.

4.1.1.3 Corpus CREA

El Corpus de Referencia del Español Actual (CREA) viene dado por un conjunto de textos en formato digital que presenta los usos del español contemporáneo y que cuenta con múltiples listas de frecuencias, de las cuales la más amplia de ellas en la actualidad la constituyen 737779 entradas con sus respectivas frecuencias absolutas y normalizadas (Anonymous, 2024b; Iandola et al., 2016; Lê et al., 2023; RAE, s. f.-c).

Preprocesamiento de datos

1. Carga y limpieza de datos:

- Los datos se cargan desde archivos en formato `CREA_total.csv` utilizando la biblioteca `pandas`.
- Se eliminan celdas vacías y se limpian los valores de las celdas eliminando espacios en blanco no deseados.
- Las frecuencias absolutas en CREA, originalmente almacenadas como texto, se convierten a valores numéricos (`float`).

2. Normalización:

Para obtener las correspondencias en la extracción de la frecuencia CREA:

- Se eliminaron las tildes de las palabras tanto en `target_word` y en CREA.
- Se convirtieron las palabras a minúsculas.

3. Tokenización:

Las frases contenidas en `target_word` se dividieron en palabras individuales utilizando expresiones regulares.

4. Lematización:

Se utiliza TreeTagger a través de la biblioteca python `treetaggerwrapper` versión 2.3 para obtener el lema de la palabra (`target_word`).

Ingeniería de características

1. Cálculo de frecuencias absolutas:

Se buscan las frecuencias absolutas de palabras (con el lema obtenido en el preprocesamiento) en el corpus CREA para palabras simples como frases (en el `target_word`), para el caso de las frases se toma la menor frecuencia de sus palabras.

2. Procesamiento paralelo:

Para optimizar el rendimiento, se utiliza la clase `ThreadPoolExecutor` de la biblioteca `concurrent.futures`. Esto permite calcular las frecuencias absolutas de manera paralela para grandes volúmenes de datos.

4.1.1.4 Wordfreq

En el año 2022, Speer desarrolló Wordfreq, un programa de Python que facilita consultar la frecuencia que alcanza el uso de palabras en más de 40 idiomas. Esta biblioteca se apoya en diversas fuentes de datos y los cálculos se ejecutan a partir de una escala logarítmica de Zipf.

Preprocesamiento de datos

1. Tokenización:

Las frases (en `target_word`) se descompusieron en sus palabras utilizando expresiones regulares.

2. Normalización de texto:

Las palabras se convirtieron a minúsculas y se eliminaron espacios en blanco al inicio y al final.

Ingeniería de características

1. Cálculo de frecuencias:

Se utilizó la biblioteca Wordfreq en su versión 3.1.1 para obtener las frecuencias de las palabras proporcionadas por esta herramienta. En el caso de las frases objetivo, se tomó como referencia el menor valor entre las frecuencias de todas las palabras que las componen.

4.1.1.5 Spanish billion word corpus and embeddings

El Spanish Billion Word Corpus and Embeddings (SBW) es un recurso en español que contiene un conjunto de aproximadamente 1.5 mil millones de palabras en este idioma y que está sin anotaciones. Este conjunto fue creado a partir de varias fuentes en internet e incluye vectores de palabras creados con el algoritmo word2vec, mediante el uso del modelo skip-gram. Los vectores de palabras fueron evaluados con una serie de tests que exploran las relaciones entre palabras (Cardellino, 2016).

Preprocesamiento de datos

1. Normalización de texto:

Las palabras objetivo (target_word) se convierten a minúsculas y se eliminan espacios en blanco al inicio y al final.

Ingeniería de características

1. Cálculo de frecuencias:

Para obtener la frecuencia del recurso, se utilizó el modelo preentrenado SBW-vectors-300-min5.bin en español, utilizando el método count para calcular las frecuencias del vocabulario y considerando las posibles variaciones en la composición del modelo. Se empleó una función basada en Word2Vec (implementado con gensim.models.KeyedVectors) para extraer la frecuencia de las palabras en el vocabulario. En los casos donde las palabras o frases no estaban presentes en el vocabulario del modelo, se les asignó una frecuencia de 0.

4.1.1.6 Propiedades léxicas

Preprocesamiento de datos

1. Tokenización:

- Las frases contenidas en target_word se dividieron en palabras individuales utilizando expresiones regulares.

2. Normalización de texto:

- Se eliminan espacios en blanco al inicio y al final.

Ingeniería de características

1. En las características léxicas de palabras, se midió su extensión mediante el número de letras que componen la palabra o frase objetivo.
2. Para determinar la cantidad de palabras, se realiza un conteo en la frase objetivo, asignando el valor de 1 si se trata de una sola palabra.
3. Se codificó la presencia de mayúsculas:
 - 1 indica que la palabra o frase contiene al menos una letra mayúscula.
 - 0 indica que no contiene ninguna letra mayúscula.

4.1.2 Selección de la combinación óptima de características

En la selección de la combinación óptima de características para el modelo se utilizó XGBClassifier, ya que de acuerdo con Alsahaf et al. (2022), XGBoost supera a modelos tradicionales, ofreciendo mejor interpretación de interacciones complejas y reducción de redundancia.

Siguiendo parte del método de evaluación realizado por Alarcon et al. (2021), la Tabla 2 presenta los resultados de algunas combinaciones de características evaluadas mediante validación cruzada de 5 pliegues con el algoritmo XGBClassifier en el conjunto de entrenamiento.

Tabla 2

Resultados de la validación cruzada 5 pliegues para las combinaciones de características en el conjunto de entrenamiento del dataset CWI Shared Task 2018

Categoría	Características	Accuracy	Macro Precision	Macro Recall	Macro F1
Embeddings de palabras y oraciones	p300	0.7852	0.7781	0.7679	0.7717
	p400	0.7878	0.7809	0.7708	0.7746
	p500	0.7900	0.7832	0.7732	0.7770
	o500	0.5748	0.5371	0.5314	0.5262
Frecuencias	co (CORPES XXI)	0.6751	0.6594	0.6396	0.6420
	cr (CREA)	0.6441	0.6311	0.6336	0.6320
	wf (wordfreq)	0.6842	0.6710	0.6726	0.6717
	sbw (Spanish Billion Word Corpus and Embeddings)	0.7204	0.7097	0.6927	0.6971
Propiedades léxicas	l (longitud)	0.7385	0.7482	0.6974	0.7030
	np (número de palabras)	0.7354	0.8476	0.6664	0.6597
	m (contiene mayúsculas)	0.6033	0.3016	0.5000	0.3763
Combinaciones	p500 + o500	0.7882	0.7827	0.7692	0.7739
	l + np + m	0.7520	0.7833	0.7037	0.7098
	p500 + l + np + m	0.8035	0.8005	0.7834	0.7893
	p500 + co + cr + wf + sbw	0.8017	0.7954	0.7867	0.7901
	p500 + co + cr + l + np + m	0.8060	0.8019	0.7881	0.7931
	p500 + co + wf + l + np + m	0.8081	0.8033	0.7915	0.7959
	p500 + co + sbw + l + np + m	0.8083	0.8041	0.7907	0.7956
	p500 + cr + wf + l + np + m	0.8049	0.7998	0.7882	0.7926
	p500 + cr + sbw + l + np + m	0.8045	0.8000	0.7866	0.7915
	p500 + wf + sbw + l + np + m	0.8037	0.7982	0.7872	0.7914
	p500 + co + cr + wf + l + np + m	0.8065	0.8015	0.7897	0.7941
	p500 + co + cr + sbw + l + np + m	0.8078	0.8036	0.7902	0.7951
	p500 + co + wf + sbw + l + np + m	0.8078	0.8028	0.7912	0.7956
	p500 + cr + wf + sbw + l + np + m	0.8072	0.8021	0.7906	0.7950
	p500 + co + cr + wf + sbw + l + np + m	0.8104	0.8059	0.7938	0.7983

La Tabla 2 muestra que la combinación de características "p500 + co + cr + wf + sbw + l + np + m" proporciona el mejor desempeño en todas las métricas evaluadas. Esta combinación integra representaciones semánticas avanzadas (embeddings de palabras de 500 dimensiones), medidas de frecuencia léxica provenientes de distintos corpus, y propiedades formales como la longitud de la palabra, el número de palabras y la presencia de mayúsculas. La mejora en los resultados al combinar estos distintos tipos de características refuerza la idea de que la complejidad léxica es un fenómeno multidimensional, que no puede ser capturado adecuadamente por una sola fuente de información.

Este enfoque integrado aprovecha tanto los avances recientes en embeddings de nueva generación, capaces de capturar matices contextuales y semánticos, como información léxica más tradicional, permitiendo así modelar de manera más precisa la complejidad léxica, en línea con los desafíos actuales del procesamiento de lenguaje natural.

En particular, los resultados evidencian que los embeddings de nueva generación, específicamente los de 500 dimensiones para la palabra objetivo, logran el mayor Macro F1 entre todas las configuraciones evaluadas. No obstante, la inclusión de embeddings de la oración completa en la que aparece la palabra objetivo no mejora el rendimiento del modelo, por lo que esta característica fue descartada en la selección final.

Estos hallazgos también reflejan la evolución del NLP hacia enfoques que integran representaciones distribuidas con propiedades lingüísticas tradicionales. El éxito de este enfoque híbrido demuestra cómo los avances en representaciones semánticas densas pueden complementarse con conocimiento lingüístico estructurado para modelar fenómenos complejos del lenguaje.

Las variables entrenadas con el algoritmo XGBClassifier obtuvieron las siguientes importancias: las representaciones semánticas (embeddings p500 con Macro F1 de 0.7770), seguidas por propiedades formales como la longitud de palabra (l: 0.7030) y métricas de frecuencia léxica de diversos corpus (sbw: 0.6971), que son lingüísticamente significativas porque capturan diferentes dimensiones cognitivas del procesamiento léxico. Los embeddings modelan relaciones conceptuales y matices semánticos que influyen en la comprensión, mientras que la longitud refleja la complejidad morfológica y visual que afecta el reconocimiento de palabras. Las frecuencias léxicas, por su parte, representan el grado de exposición y familiaridad que tienen los hablantes con determinados términos, factor determinante en la velocidad de acceso léxico y comprensión.

4.1.3 Algoritmo de Machine Learning

Para la selección de un posible mejor algoritmo para la tarea y basado en los experimentos de Gooding y Kochmar (2018), quienes sugirieron que entre los mejores algoritmos para la tarea de identificación de palabras complejas se encuentran el uso de técnicas basadas en conjunto, se realizó la validación cruzada de 5 pliegues en el conjunto de entrenamiento con los algoritmos de clasificación RandomForestClassifier, XGBClassifier y LGBMClassifier, utilizando sus

hiperparámetros por defecto y las características seleccionadas anteriormente. Tras esta evaluación, el LGBMClassifier obtuvo el mejor Macro F1.

En la Tabla 3 se muestran dichos resultados.

Tabla 3

Resultados de la validación cruzada de 5 de pliegues en el conjunto de entrenamiento para los algoritmos con la selección óptima de características

Algoritmo	Accuracy	Macro Precision	Macro Recall	Macro F1
RandomForestClassifier	0.8058	0.8009	0.7891	0.7935
XGBClassifier	0.8104	0.8059	0.7938	0.7983
LGBMClassifier	0.8131	0.8107	0.7942	0.8000

La selección de LGBMClassifier como el algoritmo óptimo para esta tarea se fundamenta en los resultados obtenidos durante la validación cruzada de 5 pliegues cuyos resultados se muestran en la Tabla 3, donde este algoritmo alcanzó el valor más alto de Macro F1 (0.8000) en comparación con RandomForestClassifier y XGBClassifier. El Macro F1 es una métrica especialmente relevante en problemas de clasificación con clases desbalanceadas, ya que considera tanto la precisión como la exhaustividad de manera equilibrada para todas las clases. Además, LGBMClassifier también obtuvo los mejores resultados en Accuracy (0.8131) y Macro Precision (0.8107), lo que refuerza su desempeño superior en el conjunto de características evaluado. Por lo tanto, se eligió el algoritmo LGBMClassifier y se descartaron los otros algoritmos evaluados.

Los tres algoritmos evaluados presentan características distintivas que pueden explicar sus diferencias de rendimiento en la tarea de identificación de palabras complejas.

Por un lado, el RandomForestClassifier construye múltiples árboles de decisión que funcionan de forma independiente (sin influirse entre sí) y luego promedia sus predicciones. A diferencia de los algoritmos de boosting, estos árboles no aprenden secuencialmente de los errores de los árboles anteriores, lo que reduce su capacidad para detectar patrones lingüísticos sutiles o complejos en la identificación de palabras difíciles (Breiman, 2001).

Por otro lado, el XGBClassifier implementa el algoritmo XGBoost (eXtreme Gradient Boosting), una técnica avanzada de gradient boosting que a diferencia del RandomForest, XGBoost construye árboles secuencialmente donde cada nuevo árbol se entrena específicamente para corregir los errores cometidos por el conjunto de árboles anteriores (T. Chen & Guestrin, 2016).

Finalmente, el LGBMClassifier alcanzó el mejor rendimiento gracias a su innovador algoritmo de crecimiento leaf-wise, que expande el árbol desde las hojas con mayor ganancia de información, permitiendo explorar más eficientemente características lingüísticas discriminantes. Su técnica GOSS (Gradient-based One-Side Sampling) permite gestionar más efectivamente el desbalance entre clases al mantener todas las instancias con gradientes altos mientras muestreando aleatoriamente aquellas con gradientes bajos. Adicionalmente, su capacidad para manejar su eficiencia computacional superior le permiten capturar mejor los patrones lingüísticos que determinan la complejidad léxica, resultando particularmente efectivo para esta tarea de clasificación. Su arquitectura optimizada para consumo de memoria reducido permite entrenar

modelos más complejos sin comprometer los recursos computacionales, lo que resulta muy importante para dispositivos con recursos limitados (Ke et al., 2017).

4.1.4 Hiperparámetros

Para el desarrollo de este estudio, se optó por emplear los valores predeterminados de los hiperparámetros establecidos por la biblioteca utilizada, sin realizar modificaciones adicionales, considerando que estos parámetros han sido definidos para garantizar un desempeño óptimo en la mayoría de los casos de uso. No obstante, investigaciones futuras podrían explorar el ajuste fino de estos hiperparámetros con el fin de optimizar aún más los resultados en contextos específicos.

4.1.5 Arquitectura del modelo

La arquitectura del modelo fue planteada en función del tipo de aprendizaje que se utilizó, el cual, en este caso, fue un aprendizaje supervisado de clasificación binaria y principalmente está conformada por las variables de entrada, el algoritmo de machine learning y la variable de salida.

A continuación, se presentan estos componentes principales:

1. La combinación óptima de características seleccionadas (variables de entrada para el modelo) que se obtienen a partir de la palabra o frase objetivo, la cual viene integrada por embeddings de 500 dimensiones provistos por el modelo text-embedding-3-large, frecuencias basadas en los recursos (CORPES XXI, CREA, Wordfreq y Spanish Billion Word Corpus and Embeddings) y propiedades léxicas (longitud, número de palabras e indicador de mayúsculas).
2. El algoritmo LGBMClassifier de la biblioteca Python sklearn con sus hiperparámetros por defecto.
3. Un valor binario de salida del modelo (1 ó 0) si la palabra es compleja o no compleja respectivamente.

En la Figura C-1 del Anexo C se muestra la arquitectura del modelo planteado dentro de su pipeline completo de machine learning, en el cual se muestran los detalles más importantes de esta.

4.2 Análisis, interpretación y discusión de resultados

En una etapa final, el algoritmo que se seleccionó fue el clasificador LGBMClassifier, el cual fue entrenado en el conjunto de entrenamiento del dataset CWI Shared Task 2018 en español con la combinación óptima de características encontrada y luego fue evaluado tanto en el dataset de prueba en español del CWI Shared Task 2018 como en el conjunto de prueba basado en Textos Escolares utilizando varias métricas para evaluar su rendimiento.

A continuación, se presentan y analizan los resultados de cada objetivo específico.

4.2.1 Objetivo Específico 1: Identificar las variables seleccionadas basadas en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, entrenadas en el dataset CWI 2018, que son más importantes al evaluarlas en el dataset de Textos Escolares de instituciones secundarias de un distrito de Cajamarca.

a) Análisis de resultados

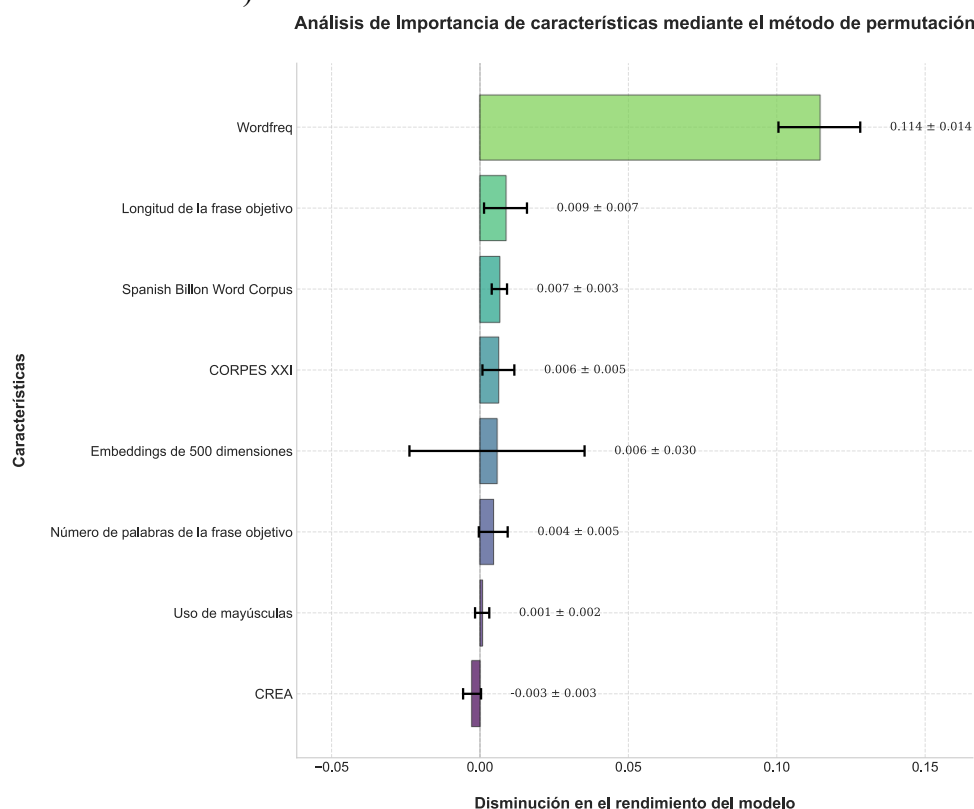
Tabla 4

Características seleccionadas durante el entrenamiento para el modelo de machine learning

Categoría	Característica	Abreviación	Descripción
Embeddings	Embeddings de palabras	p500	Embeddings de 500 dimensiones de la palabra o frase objetivo mediante text-embedding-3-large
	CORPES XXI	co	Basado en diccionario frecuencias corpes_alfa.tsv
	CREA	cr	Basado en CREA_total.csv
Frecuencias	Wordfreq	wf	Basado librería wordfreq 3.1.1
	Spanish Billion Word Corpus and Embeddings	sbw	Basado en SBW-vectors-300-min5.bin
Propiedades léxicas	Longitud	l	Longitud de la palabra o frase objetivo
	Número de palabras	np	Número de palabras en la frase o palabra objetivo
	Contiene mayúsculas	m	1 si contiene mayúsculas, 0 si no contiene

Figura 7

Importancia de características mediante el método de permutación (en el dataset de prueba basado en Textos Escolares)



Nota. Nota. La importancia por permutación de características mide cuánto disminuye el rendimiento del modelo cuando los valores de dicha característica se aleatorizan (permutan), rompiendo así su relación con la variable objetivo. Este proceso se ejecutó mediante 30 repeticiones utilizando accuracy como métrica de rendimiento, y la importancia final reportada corresponde al promedio de estas 30 repeticiones. Para evaluar la importancia global de los 500 embeddings (0.006), cada dimensión fue permutada de forma independiente, midiendo su contribución individual a la disminución del accuracy; estas 500 contribuciones individuales se sumaron para obtener la importancia total de la característica embeddings. La incertidumbre asociada (± 0.030) se calculó aplicando la fórmula de propagación de errores para variables independientes, tomando la raíz cuadrada de la suma de las varianzas individuales de las 500 dimensiones, reflejando la incertidumbre acumulada del proceso de permutación múltiple.

b) Interpretación de resultados

En la Tabla 4 se muestra la combinación óptima de variables de entrada (características) seleccionada para el desarrollo del modelo y basada en el conjunto de entrenamiento. En la Figura 7 se muestra la importancia de características mediante el método de permutación, tomando como referencia el dataset de prueba basado en Textos Escolares. A continuación, se analizan las variables de mayor relevancia predictiva identificadas:

Las variables más importantes (importancia por permutación ($a \pm b \geq 0$)) incluyen Wordfreq como predictora dominante (0.114 ± 0.014), longitud de frase objetivo (0.009 ± 0.007), Spanish Billion Word Corpus (0.007 ± 0.003) y CORPES XXI (0.006 ± 0.005). Las variables menos importantes comprenden los Embeddings de 500 dimensiones (0.006 ± 0.030), el número de palabras de la frase objetivo (0.004 ± 0.005), el uso de mayúsculas (0.001 ± 0.002) y CREA con contribución negativa (-0.003 ± 0.003).

A continuación, se interpreta cada una de las importancias de las variables:

1. Wordfreq

La frecuencia Wordfreq alcanzó el primer lugar en importancia de características por permutación en el dataset basado en Textos Escolares (en Figura 7). Este análisis revela que esta característica tiene un impacto predominante en el modelo predictivo, reduciendo su rendimiento en un 11.4% (± 0.014) cuando se aleatoriza, consolidándola como la variable más influyente en este contexto educativo específico, respaldando estudios psicolingüísticos previos que evidencian una relación inversa entre la frecuencia de uso y la dificultad de procesamiento (Ellis, 2002).

Esta prominencia de la frecuencia léxica se alinea con múltiples marcos teóricos. Nation (2001) señala que las palabras frecuentes son más fáciles de reconocer y procesar, lo que puede reflejar una menor percepción de complejidad léxica. Los resultados también apoyan la Teoría de la Carga Cognitiva de Sweller (1994), donde las palabras menos frecuentes (potencialmente más complejas) pueden demandar mayor esfuerzo de procesamiento. Además, la disponibilidad según frecuencia se alinea con la Zona de Desarrollo Próximo de Vygotsky (1978), donde la frecuencia léxica actúa como puente entre el conocimiento actual y el potencial dominio léxico.

Desde la perspectiva del machine learning, los trabajos de Chen & Meurers (2016) y Kauchak (2016) respaldan la robustez de las frecuencias como predictores. Esta convergencia teórica y empírica coincide con las observaciones de Perfetti (2007) sobre la variabilidad en la calidad de representaciones léxicas y con Ellis (2002) respecto a cómo la frecuencia de exposición potencia la adquisición y disponibilidad del vocabulario. En conjunto, estos hallazgos establecen que la frecuencia léxica constituye un factor fundamental para la predicción binaria de la complejidad léxica en modelos de machine learning aplicados a contextos educativos.

2. Longitud de la frase objetivo

La longitud de la frase objetivo alcanza el segundo lugar en importancia para el dataset de Textos Escolares (0.009 ± 0.007), lo que refleja los principios establecidos por la investigación en carga cognitiva.

Los hallazgos se alinean con los postulados de Baddeley (1992) y Sweller (1994) sobre memoria de trabajo y carga cognitiva. En contextos educativos, donde los estudiantes tienen capacidades de procesamiento en desarrollo, las frases más largas representan una mayor demanda cognitiva, explicando su relevancia como predictor. Los estudiantes deben procesar más información simultáneamente, lo que incrementa la dificultad percibida del texto.

Desde la perspectiva computacional, Jurafsky y Martin (2025) mencionan que rasgos como la longitud de las oraciones han sido indicadores importantes de la dificultad de los textos. Esto es porque pueden influir directamente en la carga de procesamiento tanto para los sistemas automáticos como para las personas. Así, la longitud textual no constituye solo una característica superficial, sino que está estrechamente relacionada con los procesos cognitivos involucrados en la comprensión lectora. Esto resulta especialmente relevante en aplicaciones educativas orientadas a ajustar los materiales a las capacidades individuales del estudiante, favoreciendo una mejor adaptación y accesibilidad del contenido.

3. Spanish Billion Word Corpus

La frecuencia basada en el Spanish Billion Word Corpus ocupa el tercer lugar en importancia por permutación para el dataset de Textos Escolares (0.007 ± 0.003). Esto demuestra que esta característica se alinea específicamente con el vocabulario académico utilizado en los textos escolares de las instituciones secundarias de este estudio. Capturando patrones léxicos relevantes para predecir complejidad. Esta alineación confirma que Spanish Billion Word Corpus no es redundante, sino que aporta valor específico para este dominio en particular, complementando estratégicamente otros recursos según el contexto textual analizado.

Su contribución no redundante confirma que diferentes corpus aportan valor según el dominio textual. Esta convergencia empírica respalda estudios previos (Chen y Meurers, 2016; Kauchak, 2016) y establece las frecuencias como predictores robustos donde la exposición frecuencial (Ellis, 2002) Ellis (2002) y variabilidad léxica Perfetti (2007) determinan la accesibilidad del vocabulario en contextos educativos.

4. CORPES XXI

CORPES XXI, presenta una contribución modesta en las métricas de importancia por permutación (0.006 ± 0.005 , con el cuarto lugar en términos de importancia por permutación), sí aporta valor al rendimiento global del modelo de complejidad léxica. Esta contribución del 1%, si bien parece reducida en comparación con el dominio de Wordfreq (0.114 ± 0.014 en permutación), representa información complementaria valiosa que enriquece las predicciones del modelo. La baja contribución relativa sugiere que su rendimiento podría estar disminuido por redundancia informativa con otros recursos, especialmente con Wordfreq, ya que ambos capturan patrones de frecuencia léxica, lo que hace que el modelo privilegie Wordfreq y relegue parcialmente las contribuciones de CORPES XXI. Sin embargo, a pesar de esta redundancia, CORPES XXI proporciona matices distintivos del español contemporáneo que los otros corpus no reflejan completamente. La variabilidad observada en su contribución sugiere que CORPES XXI captura patrones específicos que, aunque potencial y parcialmente solapados, ofrecen información única en ciertos contextos. Incluso contribuciones aparentemente pequeñas como la de CORPES XXI pueden ser decisivas para mejorar la precisión en casos límite, ya que la combinación sinérgica de múltiples fuentes es lo que permite al modelo alcanzar su rendimiento óptimo en la predicción binaria de complejidad léxica.

5. Embeddings text-embedding-3-large de 500 dimensiones

En el contexto educativo de las instituciones de secundaria de esta investigación, los embeddings de 500 dimensiones muestran una importancia por permutación modesta (0.006 ± 0.030), ubicándose en el quinto lugar en términos de relevancia al analizar la complejidad léxica de textos escolares.

Sin embargo, en muchos otros estudios, los embeddings alcanzan niveles de importancia significativamente mayor, subrayando su capacidad para capturar representaciones semánticas avanzadas, respaldando la hipótesis de Perfetti (2007), según la cual las palabras complejas están asociadas con representaciones semánticas menos desarrolladas. Este contraste ilustra cómo el rendimiento de los embeddings puede variar drásticamente dependiendo del dominio y el contexto en el que se utilicen.

Los embeddings, al representar información semántica en espacios vectoriales, pueden capturar eficazmente la distancia conceptual entre palabras fáciles y difíciles. Desde una perspectiva educativa, esto explica por qué los estudiantes enfrentan dificultades no solo con palabras raras, sino con aquellas que están "semánticamente distantes" de su vocabulario conocido. Sin embargo, esta sofisticación presenta ciertas limitaciones cuando se aplica a este contexto en específico.

Esta capacidad de modelar distancias semánticas conecta directamente con la teoría de Vygotsky (1978) sobre la zona de desarrollo próximo: los embeddings revelan qué palabras están al alcance del estudiante y cuáles necesitan más apoyo para ser comprendidas. Complementariamente, la teoría de Sweller (1994) sobre la carga cognitiva refuerza esta perspectiva: las palabras que aparecen alejadas del vocabulario común en el espacio vectorial requieren más esfuerzo mental para procesarse.

Esta correspondencia entre distancia vectorial y esfuerzo cognitivo constituye un fundamento de por qué los embeddings pueden ser buenos predictores para la tarea de predicción binaria de la complejidad léxica de palabras.

Mientras que su sofisticación semántica los convierte en predictores destacados en muchos estudios de procesamiento del lenguaje natural, esta misma complejidad puede reducir su transferibilidad directa a contextos específicos como el de este estudio.

6. Número de palabras de la frase objetivo

El número de palabras de la frase objetivo muestra un comportamiento se ubica en la sexta posición con una importancia de 0.004 ± 0.005 .

La influencia positiva de esta variable se alinea con los postulados de la teoría de la carga cognitiva de Chandler y Sweller (1991), quienes sostienen que la cantidad de elementos informativos simultáneos afecta la capacidad de procesamiento. Aunque en menor medida en comparación con varias características empleadas, conserva una influencia positiva en el rendimiento general de este estudio, lo cual evidencia su relevancia como predictor complementario en la identificación de palabras complejas.

En textos educativos, donde prevalecen estructuras sintácticas simplificadas, el número de palabras opera como un indicador secundario de complejidad. En contraste, en textos generales con mayor variabilidad estructural, esta característica adquiere mayor poder predictivo al reflejar la complejidad composicional del mensaje.

7. Uso de Mayúsculas

El uso de mayúsculas constituye la característica con menor relevancia predictiva, ocupando la séptima posición en el dataset de Textos Escolares con importancia de 0.001 ± 0.002 . Esta consistencia en su limitada contribución sugiere que las convenciones ortográficas no constituyen un indicador primario de complejidad léxica.

Desde el marco de la automaticidad lectora propuesto por LaBerge y Samuels (1974), el uso de mayúsculas no interfiere significativamente con los procesos automáticos de decodificación una vez alcanzada la competencia lectora básica. Esta característica opera como una señal superficial que no modifica sustancialmente la demanda cognitiva asociada al acceso léxico-semántico.

Esto refuerza la hipótesis de que las convenciones ortográficas constituyen elementos periféricos en la determinación de complejidad léxica. Esto se alinea con los hallazgos de Perfetti (2007) respecto a que la familiaridad léxica y la frecuencia ejercen influencia predominante sobre las características tipográficas.

Desde la perspectiva computacional, estos resultados validan la jerarquización de características en modelos de identificación automática, sugiriendo que los recursos de procesamiento deben priorizarse hacia variables léxico-semánticas y frecuenciales antes que hacia a estos indicadores.

8. CREA

El análisis de permutación de características presentado en la Figura 7 revela algo interesante: el corpus CREA muestra un impacto muy reducido, incluso ligeramente negativo (-0.003 ± 0.003) en el dataset basado en Textos Escolares. Esta importancia mostrada en la Figura 7 puede explicarse mediante dos factores complementarios.

1. En primer lugar, podríamos estar ante un fenómeno de redundancia informativa, ampliamente documentado en la literatura sobre selección de características (Guyon y Elisseeff, 2003). En el contexto del presente estudio, la información lingüística capturada por el corpus CREA parece estar ya representada a través de otras características del sistema. Wordfreq, Spanish Billion Word Corpus y CORPES XXI proporciona frecuencias léxicas actualizadas y potencialmente más alineadas con el contexto educativo de las instituciones secundarias de este estudio. Esta superposición informativa explicaría por qué la permutación de los valores de CREA no deteriora el rendimiento del modelo, e incluso podría mejorarlo marginalmente al eliminar ruido redundante del sistema.
2. En segundo lugar, la discrepancia entre dominios lingüísticos constituye otro factor relevante. El modelo fue entrenado con datos del CWI Shared Task 2018, que comprende un corpus de dominio general con textos diversos, mientras que la evaluación se realizó sobre textos escolares, que representan un registro lingüístico considerablemente más específico. El corpus CREA, al ser una colección de textos de naturaleza amplia y heterogénea, podría no capturar adecuadamente las particularidades del lenguaje académico simplificado presente en los materiales educativos. Esta desalineación entre el dominio de entrenamiento (general) y el de evaluación (educativo) sugiere que las frecuencias y patrones extraídos de CREA podrían ser menos relevantes para predecir la complejidad en el contexto específico de textos escolares.

c) Discusión de resultados

Se identificaron y examinaron variables para construir modelos de machine learning que puedan predecir binariamente la complejidad léxica de palabras en instituciones secundarias en un distrito en Cajamarca. Dicha selección se realizó en el conjunto de entrenamiento del CWI Shared Task 2018 en español. Además, se examinó la importancia de cada una de ellas en dicho modelo, evaluándolo en el conjunto de prueba basado en Textos Escolares. Se logra determinar que todas las características, excepto una, tienen una influencia positiva en la evaluación de dicho dataset, la característica que tiene una influencia nula o negativa es la frecuencia basada en el corpus CREA.

Este proceso se fundamentó en las dimensiones de complejidad absoluta señaladas por Bulté y Housen (2012) así como por Paetzold y Specia (2016). Se incorporaron variables como la longitud de palabra, la presencia de mayúsculas y el número de palabras de la frase objetivo que representan aspectos morfosintácticos intrínsecos que pueden aumentar la complejidad estructural del término; además, se incluyen variables basadas en corpus y embeddings: los embeddings de palabras proporcionan una representación del contexto semántico; las frecuencias del CORPES XXI, CREA, Wordfreq y Spanish Billion Word Corpus and Embeddings que ofrecen frecuencias del uso general del lenguaje español en un contexto general.

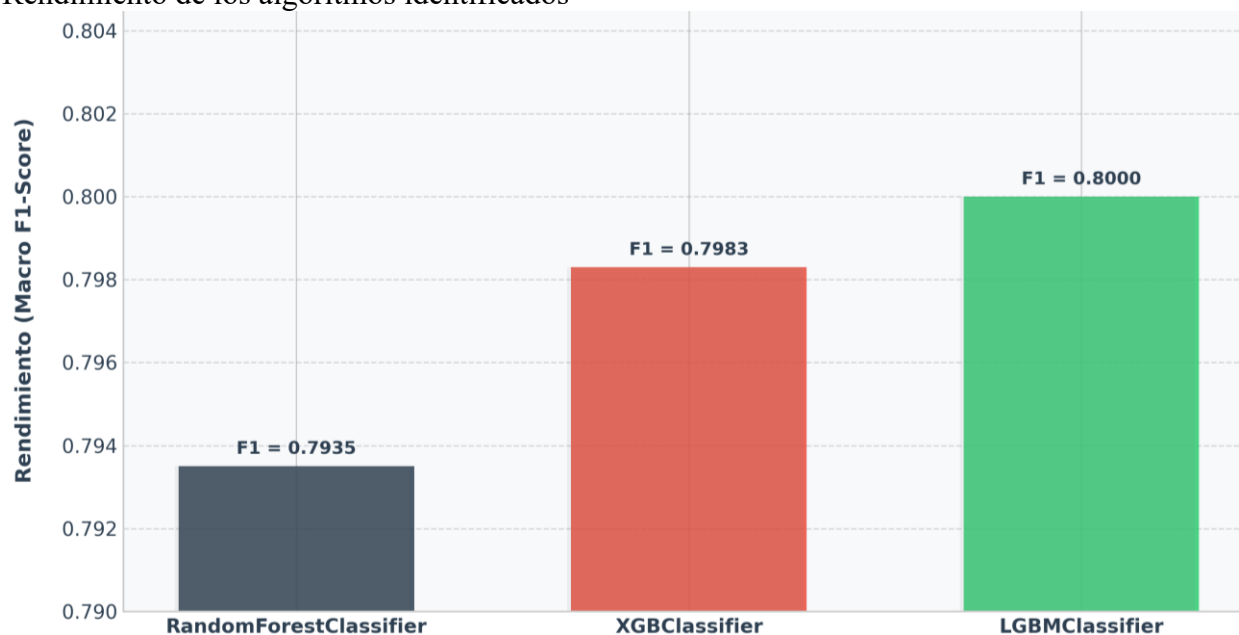
Respecto a los antecedentes revisados, en este estudio se utilizaron variables distintas a modelos anteriores revisados en la literatura, como por ejemplo, los embeddings basados en el modelo de nueva generación text-embedding-3-large presentados a principios del año 2024 por OpenAI, que permiten una representación más sofisticada del contexto semántico y se utilizaron las frecuencias basadas en recursos lingüísticos actuales como el diccionario de frecuencias de la versión 1.1 de CORPES XXI lanzada en abril de 2024, proporcionando así datos más contemporáneos sobre el uso del español. Por otro lado, los resultados que se obtienen a partir del análisis de las características mediante validación cruzada son compatibles con los experimentos de Shardlow et al. (2022), acerca de que existe correlación entre la longitud de las palabras y su complejidad.

4.2.2 Objetivo Específico 2: Identificar el algoritmo de machine learning evaluado y basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, que obtiene el mejor rendimiento en el dataset CWI 2018 para proponerse a instituciones secundarias de un distrito de Cajamarca.

a) Análisis de resultados

Figura 8

Rendimiento de los algoritmos identificados



b) Interpretación de resultados

De acuerdo con la Figura 8, los resultados muestran que, de los algoritmos identificados (RandomForestClassifier, XGBClassifier y LGBMClassifier), el algoritmo LGBMClassifier consiguió el mejor rendimiento en cuanto a Macro F1 de los tres algoritmos empleados para ser propuesto para la predicción binaria de la complejidad léxica en un contexto educativo en Cajamarca.

c) Discusión de resultados

Se identificó e implementó algoritmos de machine learning para diseñar el modelo que permita la predicción binaria de la complejidad léxica de palabras en instituciones secundarias de un distrito en Cajamarca. Para dicho proceso de diseño se realizó:

- La selección de la combinación óptima de características para el modelo utilizando XGBClassifier.
- El proceso de selección de un algoritmo de clasificación binaria para el modelo tuvo como candidatos considerados los algoritmos RandomForestClassifier, XGBClassifier y LGBMClassifier. Estos fueron comparados entre sí mediante validación cruzada sobre el conjunto de entrenamiento.
- El clasificador que se seleccionó fue el LGBMClassifier, que logró un rendimiento ligeramente mejor que los otros dos en cuanto a Macro F1, como se muestra en la Figura 8.

Este proceso se basó en lo señalado por Alsahaf et al. (2022) de que XGBoost supera a modelos tradicionales en el proceso de selección de características, ofreciendo mejor interpretación de interacciones complejas y reducción de redundancia, y por Gooding y Kochmar (2018), sobre la eficacia de los métodos ensemble en problemas como la predicción binaria de la complejidad léxica de palabras (identificación de palabras complejas), así como los principios de aprendizaje automático al refrendar la necesidad de la selección algorítmica mediante el rendimiento empírico, según su desempeño.

En el ámbito de la predicción binaria de la complejidad léxica, los algoritmos de entrenamiento han evolucionado significativamente, transitando desde enfoques tradicionales como Support Vector Machines (Alarcon et al., 2019, 2020, 2021) y Random Forest (Gooding y Kochmar, 2018; Kajiwar y Komachi, 2018), hacia modelos más sofisticados basados en redes neuronales Sheang (2019) y transformers (Ortiz-Zambrano et al., 2022, 2023a). Los algoritmos más recientes priorizan la integración de modelos de lenguaje grandes (Anonymous, 2024a), técnicas de ajuste fino, y mecanismos que mejoran la generalización y reducen sesgos, especialmente en tareas multilingües y multidominio. En el enfoque del trabajo actual, el algoritmo LGBMClassifier, que ha mostrado un rendimiento prometedor destacando por su eficiencia en el manejo de datos de alta dimensionalidad, menor tiempo de entrenamiento y mejor capacidad de generalización, aunque su uso en estudios de predicción de la complejidad léxica binaria (CWI) aún es limitado, lo que sugiere la necesidad de más investigaciones que exploren su potencial en diferentes idiomas y dominios.

4.2.3 Objetivo Específico 3: Determinar el rendimiento y la eficiencia en memoria de la arquitectura del modelo basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, al compararlo con un baseline del CWI 2018, considerando los requisitos de memoria RAM viables para las instituciones secundarias de un distrito de Cajamarca.

a) Presentación de resultados

Tabla 5

Resultados de la evaluación en el dataset de prueba vs el baseline de la tarea compartida del CWI Shared Task 2018

Métrica	Datasets de prueba	
	Baseline CWI Shared Task 2018	Textos escolares
Macro F1 (F1 macro promedio)	0.6344	0.6736

Figura 9

Matriz de confusión de la evaluación en el dataset basado en textos escolares para el modelo desarrollado

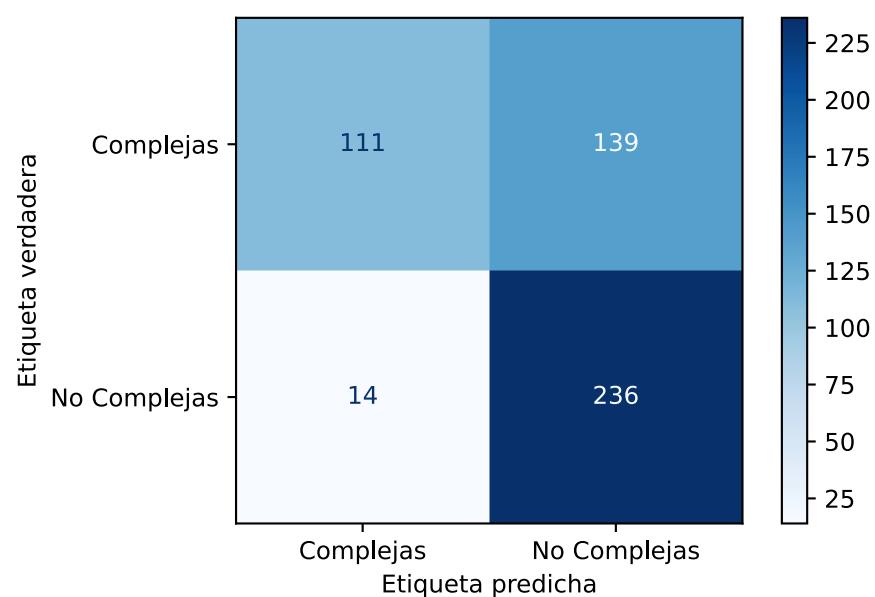
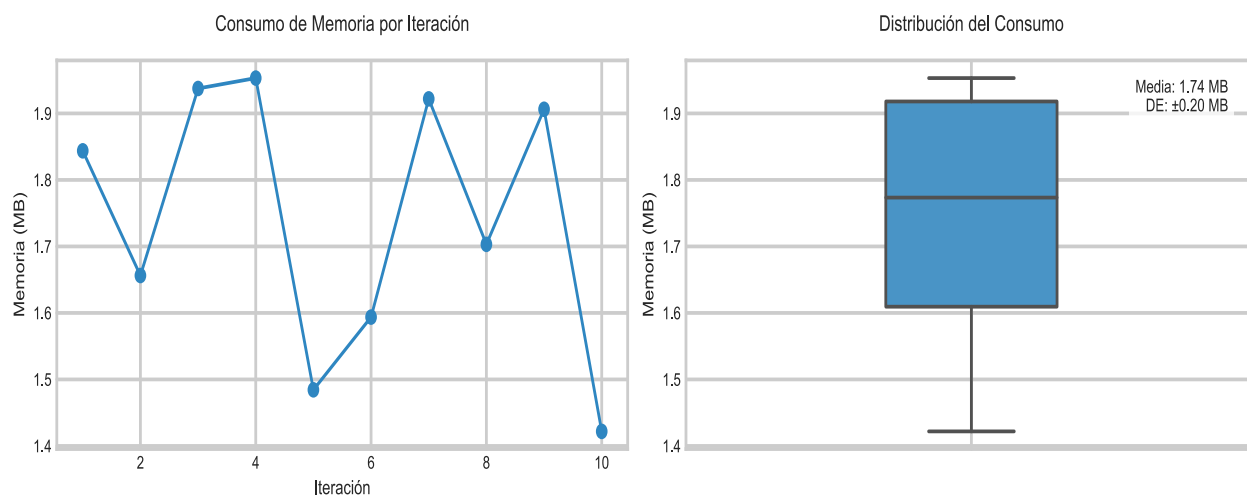


Figura 10

Medidas del consumo de memoria en MB para carga del modelo en 10 iteraciones



b) Interpretación de resultados

El modelo muestra un desempeño superior en el dataset de Textos Escolares que en el baseline en español para la predicción binaria de la complejidad léxica establecida en el CWI Shared Task 2018 (Tabla 5):

1. Los resultados muestran que el modelo funciona mejor en textos escolares (Macro F1 = 0.6736) que en el baseline en español de CWI 2018 (Macro F1 = 0.6344), superándolo con una mejora del 6.2%. Este rendimiento superior es notable considerando que se trata de un estudio pionero en el campo de la predicción de la complejidad léxica en español para contextos educativos en el Perú, y la escasez de recursos disponibles para este tipo de investigación en el contexto de las instituciones secundarias en el distrito de Cajamarca y en el Perú. El macro F1 de 0.6736 sugiere un balance moderado entre precisión y recall para las clases compleja y no compleja, estableciendo una base sólida que podría beneficiarse de refinamientos adicionales en investigaciones futuras.
2. Además, en la Figura 9 se muestra la matriz de confusión de la evaluación del modelo en el dataset de prueba basado en Textos Escolares. La matriz de confusión muestra el resultado de las predicciones del modelo en el dataset de prueba entre las clases "Complejas" y "No Complejas", donde se observa que el modelo acertó en 111 casos de "Complejas" (verdaderos positivos) y 236 casos de "No Complejas" (verdaderos negativos), mientras que cometió errores al clasificar 139 casos como "No Complejas" cuando eran "Complejas" (falsos negativos) y solo 14 casos como "Complejas" cuando eran "No Complejas" (falsos positivos), lo que indica que el modelo tiene una baja tasa de falsos positivos, aunque tiene un poco de dificultad para identificar correctamente los casos de "Complejas".

La matriz de confusión muestra que el modelo funciona bien en general (acierta el 69.2% de las veces). Sin embargo, el modelo tiende a ser cauteloso al marcar palabras como difíciles: de 250 palabras realmente difíciles, solo detecta 111, mientras que las otras 139 las considera fáciles por error. En contraste, raramente comete el error opuesto (solo marca 14 palabras fáciles como difíciles).

Este comportamiento puede explicarse por la diferencia entre los datos de entrenamiento (CWI 2018) y los datos de prueba (textos escolares). El modelo aprendió a identificar complejidad basándose en un conjunto de datos general, pero al aplicarlo específicamente al contexto educativo de las instituciones secundarias del estudio, encuentra palabras que pueden ser complejas para estudiantes pero que no fueron marcadas como tales en el entrenamiento original. Esta discrepancia entre dominios sugiere que el vocabulario basado en textos escolares tiene características particulares de complejidad que requieren un ajuste específico del modelo para el contexto educativo.

3. Por otro lado, en la Figura 10 se muestra un histograma, así como un diagrama de bigotes en los cuales se representa el consumo de memoria en la carga del modelo de machine learning para 10 iteraciones, demostrando que la arquitectura propuesta posee una

importante eficiencia a nivel computacional con 1.74 MB de consumo de memoria en promedio para ser cargado.

Los resultados muestran que el modelo presenta un consumo promedio de 1.74 MB con una desviación estándar de ± 0.20 MB durante la carga, manteniéndose consistentemente por debajo de los 2 MB en todas las iteraciones. Este consumo representa menos del 0.05% de la memoria disponible en las computadoras típicas del colegio (4-8 GB RAM), garantizando una implementación sin conflictos con otros procesos del sistema y permitiendo que funcione simultáneamente con aplicaciones educativas, navegadores web y software de oficina sin comprometer el rendimiento.

Considerando que muchas instituciones educativas secundarias nacionales en Cajamarca operan con hardware modesto y presupuestos limitados, la baja demanda de memoria del modelo (muy por debajo del umbral establecido de 100 MB) lo convierte en una solución tecnológicamente viable. Esta característica es crucial para asegurar que la herramienta de análisis de complejidad léxica pueda implementarse ampliamente sin requerir actualizaciones costosas de hardware, democratizando el acceso a tecnologías educativas avanzadas en contextos con recursos limitados.

La baja variabilidad del consumo de memoria entre iteraciones (desviación estándar de ± 0.20 MB) sugiere un comportamiento predecible y estable del modelo, facilitando su integración en sistemas educativos que requieren confiabilidad operacional. Esta estabilidad es fundamental para el despliegue en entornos educativos donde la continuidad del servicio y la predictibilidad del rendimiento son esenciales para el proceso de enseñanza-aprendizaje.

c) Discusión de resultados

Se planteó una arquitectura de un modelo de machine learning para la predicción binaria de la complejidad léxica de palabras en instituciones secundarias en un distrito de Cajamarca.

Siguiendo lo que indicó Ray (2019), la arquitectura propuesta se basa en la naturaleza del problema, que en este caso es de clasificación binaria. Se ha tenido en cuenta la ingeniería de características y se han aplicado técnicas de validación cruzada para optimizar tanto la selección de características como la elección del algoritmo de aprendizaje automático para el modelo. También se han considerado las limitaciones de recursos computacionales en las computadoras de dicho colegio como la capacidad de memoria, debido a que se debe tener como un pilar de esta cuando se tienen recursos modestos o limitados (Anonymous, 2024b; Iandola et al., 2016; Lê et al., 2023).

La arquitectura del modelo planteada en este estudio consta principalmente de variables como embeddings de nueva generación, frecuencias basadas en recursos lingüísticos actuales y algunas propiedades lingüísticas, así como el algoritmo basado en conjunto LGBMClassifier. Esta arquitectura logró un mejor que el baseline para predicción binaria de la complejidad léxica en español del CWI Shared Task 2018 obteniendo un macro F1 de 0.6736. Esto sugiere que, aunque el modelo es efectivo, podría haber espacio para mejorar su rendimiento en este dominio

específico. Es un modelo viable que podría tener futuras mejoras y es comparable en rendimiento a estudios listados en el reporte de Yimam et al. (2018) para la misma tarea, en el contexto del idioma español. Por otro lado, dicha arquitectura alcanzó una utilización de recursos de memoria eficiente de 1.74 MB lo cual es importante para ser utilizada en recursos modestos o limitados. Esto estuvo en la misma línea que el estudio realizado por Iandola et al. (2016), que logró arquitecturas que permiten su despliegue en recursos con memoria limitada, alcanzando 4.8 MB sin compresión.

4.3. Pruebas de hipótesis

4.3.1 Contrastación de hipótesis específica H1

- Hipótesis Nula (H_0): Menos de la mitad de las variables seleccionadas basadas en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, entrenadas en el dataset CWI 2018, son más importantes al evaluarlas en el dataset de Textos Escolares de instituciones secundarias de un distrito de Cajamarca (H_0 : $<50\%$ de las variables evaluadas tienen importancia $a \pm b > 0$).
- Hipótesis Alternativa (H_a): Al menos la mitad de las variables seleccionadas basadas en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, entrenadas en el dataset CWI 2018, son más importantes al evaluarlas en el dataset de Textos Escolares de instituciones secundarias de un distrito de Cajamarca (H_a : $\geq 50\%$ de las variables evaluadas tienen importancia $a \pm b > 0$).

De los resultados presentados en la sección 4.2.1, con 4 de 8 variables mostrando importancia positiva ($50\% \geq 50\%$), se rechaza H_0 y acepta H_a . Al menos la mitad de las variables basadas en embeddings, frecuencias y propiedades léxicas mantienen relevancia predictiva al transferirse del dataset CWI 2018 al contexto educativo local. Esta convergencia empírica respalda estudios previos (Chen y Meurers, 2016; Kauchak, 2016) y establece las frecuencias como predictores robustos, donde la exposición frecuente, mencionada por Ellis (2002), y la variabilidad léxica, mencionada por Perfetti (2007), determinan la accesibilidad del vocabulario en contextos educativos.

4.3.2 Contrastación de hipótesis específica H2

- Hipótesis Nula (H_0): No existe al menos un algoritmo de machine learning evaluado y basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica que obtiene un rendimiento superior en el dataset CWI 2018, para proponerse a instituciones secundarias de un distrito de Cajamarca (H_0 : No existe un algoritmo con rendimiento superior al evaluarse en el dataset CWI 2018).
- Hipótesis Alternativa (H_a): Sí existe al menos un algoritmo de machine learning evaluado y basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica que obtiene un rendimiento superior en el dataset CWI 2018, para proponerse a instituciones secundarias de un distrito de Cajamarca (H_a : Sí existe un algoritmo con rendimiento superior al evaluarse en el dataset CWI 2018).

De los resultados presentados en la sección 4.2.2, Se rechaza la hipótesis nula (H_0) y se acepta la hipótesis alternativa (H_a) basándose en la evidencia empírica que demuestra que el algoritmo LGBMClassifier cumple con los criterios especificados: es un algoritmo de machine learning que basado en embeddings modernos, frecuencias y propiedades léxicas para predicción binaria de complejidad léxica, fue evaluado exitosamente en el dataset CWI 2018 y obtuvo el mejor rendimiento en términos de Macro F1 comparado con RandomForestClassifier y XGBClassifier mediante validación cruzada, confirmando así que sí existe al menos un algoritmo con rendimiento superior para ser propuesto a instituciones secundarias del distrito de Cajamarca.

4.3.3 Contratación de hipótesis específica H3

Tomando como referencia uno de los baselines de la competencia CWI Shared Task 2018 en uno de sus tracks abordados, el cual fue un macro F1 de 0.6344 (Yimam et al., 2018), para este estudio se decidió establecer este umbral mínimo para evaluar esta hipótesis específica. Esto, debido a que la competencia presenta una condición similar a la de este estudio en ese track: la evaluación de modelos en un contexto para el cual se tiene data de entrenamiento para un contexto más amplio, pero no para el contexto específico, sino únicamente se dispuso de datos de prueba para su evaluación (en el track multilingüe). Esta analogía en las condiciones experimentales justifica la adopción de este umbral como punto de referencia.

Además, considerando las limitaciones de memoria en las computadoras del colegio (la mayoría fluctúa entre 4 a 8 GB de memoria RAM) y basándose en investigaciones previas, como Anonymous (2024b) que experimentó con pipelines de machine learning en entornos modestos con memoria limitada de 8 GB, y Lê et al. (2023) que estableció rangos de memoria para diferentes plataformas (cloud >100 MB, Mobile >1 MB, MCU <1 MB), se establece un umbral máximo de 100 MB de memoria para evaluar la idoneidad de la arquitectura del modelo de predicción binaria de complejidad léxica, lo cual podría permitir su despliegue efectivo en dicho hardware con recursos limitados.

- Hipótesis Nula (H_0): La arquitectura del modelo basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica no logra un rendimiento superior a un baseline del CWI 2018 manteniendo los requisitos de memoria RAM viables para las instituciones secundarias de un distrito de Cajamarca. (macro F1 ≤ 0.6344 ó consumo promedio de memoria en la carga del modelo > 100 MB)
- Hipótesis Alternativa (H_a): La arquitectura del modelo basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica sí logra un rendimiento superior a un baseline del CWI 2018 manteniendo los requisitos de memoria RAM viables para las instituciones secundarias de un distrito de Cajamarca. (macro F1 > 0.6344 y consumo promedio de memoria en la carga del modelo ≤ 100 MB)

De acuerdo con los resultados presentados en la sección 4.2.3, como el macro F1 obtenido en la evaluación del modelo en el dataset basado en Textos Escolares es $0.6736 > 0.6344$ y el consumo medio en memoria que ocupa el modelo es de $1.74 \text{ MB} \leq 100 \text{ MB}$, se rechaza

la hipótesis nula H_0 (macro $F1 \leq 0.6344$ ó consumo promedio de memoria en la carga del modelo > 100 MB). Por ende, se acepta la hipótesis alternativa, H_a : (macro $F1 > 0.6344$ y consumo promedio de memoria en la carga del modelo ≤ 100 MB).

4.3.4 Contrastación de hipótesis general

Basado en la contrastación de las tres hipótesis específicas antes contrastadas y dada la hipótesis general:

H: El modelo de machine learning basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica de palabras, es efectivo al lograr un rendimiento que supera al baseline del CWI 2018, en instituciones en secundarias en un distrito de Cajamarca.

Se acepta dicha hipótesis general, sustentado en la aceptación de dichas hipótesis específicas:

1. Al menos la mitad de las variables seleccionadas basadas en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, entrenadas en el dataset CWI 2018, son más importantes al evaluarlas en el dataset de Textos Escolares de instituciones secundarias de un distrito de Cajamarca (H1).
2. Sí existe al menos un algoritmo de machine learning evaluado y basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica que obtiene un rendimiento superior en el dataset CWI 2018, para proponerse a instituciones secundarias de un distrito de Cajamarca (H2).
3. La arquitectura del modelo basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica sí logra un rendimiento superior a un baseline del CWI 2018 manteniendo los requisitos de memoria RAM viables para las instituciones secundarias de un distrito de Cajamarca (H3).

Por lo tanto, al confirmarse las tres hipótesis específicas, se acepta la hipótesis general, sobre el modelo de machine learning es efectivo al lograr un rendimiento que supera al baseline del CWI 2018, en instituciones en secundarias en un distrito de Cajamarca.

4.4 Presentación de resultados

Objetivo general: Determinar la efectividad de un modelo de machine learning basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica de palabras, en comparación con un baseline del CWI Shared Task 2018, en instituciones secundarias en un distrito de Cajamarca.

- Se determinó la efectividad de un modelo de machine learning (macro $F1=0.6736$) basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica de palabras, en comparación con un baseline del CWI Shared Task 2018, en instituciones secundarias en un distrito de Cajamarca, mediante los resultados

obtenidos para los objetivos específicos antes mencionados, así como la contrastación de hipótesis:

1. Se identificaron las variables seleccionadas basadas en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, entrenadas en el dataset CWI 2018, que son más importantes al evaluarlas en el dataset de Textos Escolares de instituciones secundarias de un distrito de Cajamarca.

Las cuales incluyen frecuencia Wordfreq como predictora dominante, longitud de frase objetivo, frecuencia de Spanish Billion Word Corpus y frecuencia CORPES XXI (0.006 ± 0.005).

2. Se identificó el algoritmo de machine learning evaluado y basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, que obtiene el mejor rendimiento en el dataset CWI 2018 para proponerse a instituciones secundarias de un distrito de Cajamarca.

Dicho algoritmo escogido fue el LGBMClassifier por su mejor rendimiento en la validación cruzada, con un macro F1 de 0.8000.

3. Se determinó el rendimiento y la eficiencia en memoria de la arquitectura del modelo basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, al compararlo con un baseline del CWI 2018, considerando los requisitos de memoria RAM viables para las instituciones secundarias de un distrito de Cajamarca.

Dicha arquitectura obtuvo un macro F1 score de 0.6736 en el dataset basado en Textos Escolares y un promedio de uso de memoria en la carga del modelo de 1.74 MB.

Basado en dichos resultados se determinó la efectividad ya que superó a los baseline que se establecieron para el estudio en cuanto a su rendimiento predictivo y en cuanto a su uso de memoria para el modelo, lo cual representa una contribución muy importante, especialmente cuando se considera la falta de recursos para el desarrollo y para la evaluación de dichos modelos de inteligencia artificial en el idioma español y más específicamente en el contexto educativo de las instituciones públicas de educación básica regular del país; además, esto propone una base para su aplicación práctica en el entorno educativo local.

Teniendo en cuenta los antecedentes revisados en este estudio, esta investigación podría ser pionera en la tarea de predecir binariamente la complejidad léxica de palabras en instituciones secundarias de un distrito de Cajamarca mediante machine learning. A nivel nacional, este estudio podría servir como un punto de partida para abordar este tipo de tareas, que están relacionadas a la simplificación léxica, para el dominio específico de las instituciones secundarias en nuestro país. Lo cual podría contribuir al desarrollo de sistemas de simplificación léxica personalizados, que se adecuen mejor al contexto específico de los estudiantes, tal como lo indican (Gooding y Tragut, 2022; Lee y Yeung, 2018).

CONCLUSIONES

Conclusión general:

El modelo de machine learning basado en embeddings modernos, frecuencias y propiedades léxicas alcanzó un Macro F1 de 0.6736 que superó el baseline del CWI 2018 de 0.6344 (mejora del 6.18%), mientras mantiene eficiencia computacional con solo 1.74 MB de consumo de memoria, representando únicamente el 1.74% del límite establecido de 100 MB (98.26% por debajo del umbral máximo). La efectividad se sustenta en que el 50% de las variables (4 de 8) basadas en embeddings, frecuencias y propiedades léxicas mantuvieron relevancia predictiva al transferirse del contexto internacional al educativo local, con Wordfreq como predictora dominante (0.114 ± 0.014), seguida por longitud de frase objetivo (0.009 ± 0.007), Spanish Billion Word Corpus (0.007 ± 0.003) y CORPES XXI (0.006 ± 0.005), mientras que el algoritmo LGBMClassifier demostró rendimiento superior comparado con RandomForestClassifier y XGBClassifier mediante validación cruzada rigurosa, confirmando así que el modelo no solo supera el baseline internacional sino que presenta viabilidad práctica para su implementación en las instituciones secundarias de un distrito de Cajamarca.

Las implicaciones teóricas de esta investigación pionera en el contexto educativo confirman la importancia de modelos personalizados (Gooding y Tragut, 2022; Lee y Yeung, 2018). Al incorporar embeddings contextualizados con características lingüísticas específicas, este enfoque aborda la escasez de recursos en español (Alarcón et al., 2021; Bott et al., 2024; Degraeuwe y Goethals, 2024; Ortiz-Zambrano et al., 2022, 2023a) y permite mejor caracterización de la complejidad léxica siguiendo recomendaciones actuales (Alfter, 2024; Rozi et al., 2021). Al superar la línea base del CWI Shared Task 2018, establece un precedente prometedor comparable a estudios reportados por Yimam et al. (2018) para español, con uso eficiente de recursos apropiado para entornos con limitaciones computacionales como las de las instituciones secundarias del distrito de Cajamarca de este estudio (Anonymous, 2024b; Iandola et al., 2016; Lê et al., 2023).

La implicación práctica de los resultados de este estudio es la mostrada viabilidad del modelo para predecir binariamente la complejidad léxica en el contexto educativo tratado, abriendo la posibilidad de integrar el modelo en sistemas de simplificación léxica más completos y personalizados, que se ajusten al entorno educativo de las instituciones secundarias del distrito de Cajamarca y sirviendo como base para el desarrollo de herramientas educativas más sofisticadas y personalizadas que se adapten a las necesidades específicas ayudando a facilitar la comprensión lectora y el aprendizaje del alumnado, y ofreciendo soluciones tecnológicas que pueden mejorar directamente la experiencia educativa y el proceso de aprendizaje en el aula.

Una limitación de este estudio fue la carencia de un conjunto de datos para realizar el entrenamiento específico en el contexto educativo peruano de instituciones secundarias públicas de educación básica regular, por lo que se decidió utilizar solamente datos de prueba para la evaluación en dicho dominio. Lo cual fue impulsado por las restricciones prácticas relacionadas con la recolección y anotación de un dataset completo de entrenamiento, las cuales incluyen los altos costos en términos de tiempo, los recursos humanos y los recursos económicos necesarios para realizarlas. Sin embargo, el desarrollo del modelo fue llevado a cabo mediante el uso de un

dataset público para el entrenamiento y mediante la recolección y anotación de datos de prueba para la validación de dicho modelo en el dominio específico, dicha práctica es respaldada por investigaciones previas que han demostrado la viabilidad de entrenar modelos con datos de un contexto más amplio y evaluarlos en dominios específicos, permitiendo así validar su efectividad en el contexto educativo objetivo.

Conclusiones específicas:

1. Los resultados revelan que las variables más importantes para predecir la complejidad léxica en textos escolares son Wordfreq como predictora dominante, seguida por longitud de frase objetivo, Spanish Billion Word Corpus y CORPES XXI. Estos hallazgos confirman que la combinación de frecuencias léxicas, y propiedades léxicas proporciona caracterización efectiva de la complejidad léxica en este contexto educativo, estableciendo un marco para la adaptación de modelos CWI a entornos educativos específicos.

Por otro lado, el dataset basado en Textos Escolares recolectado es pionero en este contexto. Este podría ser usado para validar y/o ser parte de datasets más amplios en el desarrollo de modelos de machine learning que estén relacionados con una de las tareas del proceso de simplificación léxica para textos que se imparten en las instituciones secundarias públicas de educación básica regular del Perú.

2. El algoritmo LGBMClassifier (Light Gradient Boosting Machine) demostró ser la mejor opción para la predicción binaria de complejidad léxica en el dataset CWI 2018, alcanzando un Macro F1-Score de 0.8000 que supera a XGBClassifier (0.7983) y RandomForestClassifier (0.7935). Esta superioridad se fundamenta en su capacidad para integrar efectivamente embeddings modernos, frecuencias léxicas y propiedades lingüísticas en un modelo ensemble robusto.

La implementación de este algoritmo conlleva ventajas prácticas cruciales para su despliegue en instituciones secundarias de Cajamarca, incluyendo eficiencia computacional superior, menor consumo de memoria y mayor rapidez en el procesamiento, características esenciales que garantizan la viabilidad operativa del modelo considerando las limitaciones tecnológicas típicas del contexto educativo.

Además, esto representa una implicación teórica al ser uno de los primeros estudios en utilizar el algoritmo LGBMClassifier hasta la fecha en la misma tarea tratada, estableciendo un precedente para esta tarea en el contexto educativo de este estudio.

3. La arquitectura del modelo basado en embeddings modernos, frecuencias y propiedades léxicas demuestra un rendimiento superior al baseline del CWI 2018 con una mejora del 6.2% (Macro F1 = 0.6736 vs 0.6344) para la predicción binaria de complejidad léxica en textos escolares, mientras que su eficiencia en memoria con un consumo promedio de 1.74 MB (± 0.20 MB), representando menos del 0.05% de la memoria RAM disponible en equipos de instituciones secundarias del distrito de Cajamarca de este estudio.

Esta combinación de rendimiento mejorado y demanda de recursos computacionales confirma que la arquitectura propuesta es tanto efectiva y viable para su implementación en las instituciones educativas mencionadas con recursos tecnológicos limitados.

RECOMENDACIONES

Recomendación general

Se recomienda a las instituciones secundarias de educación básica regular implementar el modelo de predicción binaria de complejidad léxica, con el propósito de facilitar la identificación automática de palabras complejas y apoyar a mejorar la comprensión lectora. Asimismo, se sugiere instalar el modelo en las aulas de innovación pedagógica, capacitar a los docentes en su manejo y realizar pruebas piloto. Para su correcta implementación, se requiere que los equipos cuenten preferentemente con 8GB de memoria RAM, que se tenga acceso a textos curriculares y la participación del personal responsable del aula de innovación pedagógica. Para superar posibles desafíos, como la resistencia al cambio o dificultades técnicas, se recomienda organizar talleres de formación y elaborar materiales de apoyo.

Se recomienda al Ministerio de Educación (MINEDU) implementar el modelo desarrollado en este estudio con el objetivo de mejorar los materiales educativos que provee a las instituciones secundarias públicas de educación básica regular de todo el país, permitiendo identificar y ajustar adecuadamente los términos complejos de forma más eficiente, construyendo e incorporando glosarios más efectivos que faciliten y apoyen en el proceso de adquisición de términos más avanzados para enriquecer el léxico de los estudiantes y para prepararlos para una comprensión lectora más profunda de textos con mayor complejidad léxica. Dado que el modelo actualmente no dispone de una interfaz de usuario, es importante que el MINEDU lo integre en sistemas existentes o desarrollar una interfaz específica para optimizar su uso, en las instituciones educativas.

Recomendaciones específicas

1. Dado que el modelo obtuvo un rendimiento que tiene un margen de mejora, se recomienda a los investigadores futuros realizar procesos de recolección y etiquetado de palabras complejas y no complejas para el entrenamiento del modelo, para el cual se puede usar recursos como glosarios (para proponer las palabras complejas) así como recursos de lectura fácil (para las palabras no complejas) en el cual participen los alumnos en el etiquetado para permitir llevar a cabo la identificación y examinación de variables desde un enfoque más contextualizado, de esa forma potenciando la precisión de los resultados para la predicción binaria de la complejidad léxica de palabras en este entorno educativo específico. Además, dicho etiquetado podría incluir una escala de likert para tener una apreciación de la complejidad de palabras más exacta.
2. Se sugiere a los investigadores futuros que analicen el efecto de modelos que incluyan combinaciones de diversos clasificadores mediante distintos métodos de ensamblaje (por ejemplo, Stacking, Voting, Bagging o Blending) que incorporen al clasificador LGBMClassifier, lo que podría permitir aumentar la robustez del modelo. Este enfoque podría permitir sacar provecho de las fortalezas de cada clasificador y reducir las debilidades de cada uno de ellos.
3. Debido a que, en este estudio inicial, se construyó el modelo con los hiperparámetros por defecto, se recomienda a los investigadores educativos y especialistas la optimización de

la arquitectura del modelo realizando el ajuste fino de hiperparámetros mediante técnicas de optimización como GridSearchCV, RandomizedSearchCV, etc.

REFERENCIAS BIBLIOGRÁFICAS

- ACL. (2024, mayo). *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024 - ACL Anthology*.
<https://aclanthology.org/volumes/2024.determin-1/>
- Alarcon, R., Moreno, L., & Martínez, P. (2020). *Hulat-AlexS CWI Task-CWI for Language and Learning Disabilities Applied to University Educational Texts*. www.spacy.io/
- Alarcon, R., Moreno, L., & Martínez, P. (2021). Lexical Simplification System to Improve Web Accessibility. *IEEE Access*, 9, 58755–58767.
<https://doi.org/10.1109/ACCESS.2021.3072697>
- Alarcón, R., Moreno, L., & Martínez, P. (2021, septiembre). Exploration of Spanish Word Embeddings for Lexical Simplification. *Proceedings of the First Workshop on Current Trends in Text Simplification (CTTS 2021)*.
- Alarcon, R., Moreno, L., Segura-Bedmar, I., & Martínez, P. (2019). Lexical simplification approach using easy-to-read resources. *Procesamiento del Lenguaje Natural*, 63, 95–102.
<https://doi.org/10.26342/2019-63-10>
- Alfter, D. (2024). Complexity and Indecision: A Proof-of-Concept Exploration of Lexical Complexity and Lexical Semantic Change. *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, 137–143.
<https://doi.org/10.18653/v1/2024.lchange-1.14>
- Almeida, R., Tissot, H., & Fabro, M. D. Del. (2021). C3SL at SemEval-2021 Task 1: Predicting Lexical Complexity of Words in Specific Contexts with Sentence Embeddings. En A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, & X. Zhu (Eds.), *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* (pp. 683–687). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.semeval-1.88>
- Alpaydin, E. (2020). *Introduction to machine learning* (The MIT Press, Ed.; 4ª ed.).
- Alsahaf, A., Petkov, N., Shenoy, V., & Azzopardi, G. (2022). A framework for feature selection through boosting. *Expert Systems with Applications*, 187, 115895.
<https://doi.org/https://doi.org/10.1016/j.eswa.2021.115895>
- Anonymous. (2024a). Investigating Large Language Models for Complex Word Identification in Multilingual and Multidomain Setups. *Submitted to ACL Rolling Review - April 2024*.
<https://openreview.net/forum?id=GnGXbekH7M>
- Anonymous. (2024b). Training and Cross-Validating Machine Learning Pipelines with Limited Memory. *AutoML 2024 Methods Track*. <https://openreview.net/forum?id=4LkaPSHUQQ>
- Aroyehun, S. T., Angel, J., Pérez Alvarez, D. A., & Gelbukh, A. (2018). Complex Word Identification: Convolutional Neural Network vs. Feature Engineering. En J. Tetreault, J. Burstein, E. Kochmar, C. Leacock, & H. Yannakoudakis (Eds.), *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 322–327). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-0538>
- aws. (s. f.). *Redes neuronales y aprendizaje profundo: diferencia entre campos de la inteligencia artificial - AWS*. Recuperado 13 de noviembre de 2023, de
<https://aws.amazon.com/es/compare/the-difference-between-deep-learning-and-neural-networks/>
- Aziz, A., Hossain, Md. A., Chy, A. N., Ullah, Md. Z., & Aono, M. (2023). Leveraging contextual representations with BiLSTM-based regressor for lexical complexity prediction.

- Natural Language Processing Journal*, 5, 100039.
<https://doi.org/https://doi.org/10.1016/j.nlp.2023.100039>
- Baddeley, A. (1992). Working Memory. *Science*, 255(5044), 556–559.
<https://doi.org/10.1126/science.1736359>
- Bani Yaseen, T., Ismail, Q., Al-Omari, S., Al-Sobh, E., & Abdullah, M. (2021). JUST-BLUE at SemEval-2021 Task 1: Predicting Lexical Complexity using BERT and RoBERTa Pre-trained Language Models. En A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, & X. Zhu (Eds.), *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* (pp. 661–666). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2021.semeval-1.85>
- Bott, S., Saggion, H., Rojas, N. P., Salazar, M. S., & Ramirez, S. C. (2024). *MultiLS-SP/CA: Lexical Complexity Prediction and Lexical Simplification Resources for Catalan and Spanish*. <https://arxiv.org/abs/2404.07814v1>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
<https://doi.org/10.1023/A:1010933404324>
- Bulté, B., & Housen, A. (2012). *Defining and operationalising L2 complexity*. 21–46.
<https://doi.org/10.1075/LLLT.32.02BUL>
- Burkov, A. (2020). *Machine learning engineering*. 282.
- Byrnes, H., & Sinicropo, C. (2008). Advancedness and the development of relativization in L2 German: A curriculum-based longitudinal study. En L. Ortega & H. Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* (pp. 109–138). Routledge.
- Care. (2023, febrero 7). *5 cifras alarmantes de la educación en el Perú – CARE Perú*.
<https://care.org.pe/5-cifras-alarmantes-de-la-educacion-en-el-peru/>
- Chandler, P., & Sweller, J. (1991). Cognitive Load Theory and the Format of Instruction. *Cognition and Instruction*, 8(4), 293–332. https://doi.org/10.1207/s1532690xci0804_2
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Chen, X., & Meurers, D. (2016). Characterizing Text Difficulty with Word Frequencies. En J. Tetreault, J. Burstein, C. Leacock, & H. Yannakoudakis (Eds.), *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 84–94). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W16-0509>
- Chollet, F., & Allaire, J. J. (2018). Deep Learning in R. En *R-bloggers* (Vol. 1, Número 7080). Manning Publications Co. <https://www.manning.com/books/deep-learning-with-r>
- Comte, A. (1830). *Curso de filosofía positiva*. Bachelier.
- Comunicación 1: Secundaria* (1ª ed.). (2016). Santillana S.A.
- Comunicación 2: Secundaria* (1ª ed.). (2016). Santillana S.A.
- Comunicación 3: Educación Secundaria* (1ª ed.). (2016). Ediciones SM S.A.C.
- Comunicación 4: Educación Secundaria* (1ª ed.). (2016). Ediciones SM S.A.C.
- Comunicación 5: Educación Secundaria* (1ª ed.). (2016). Ediciones SM S.A.C.
- Connor, C. M., & Al'Otaiba, S. (2008). Literacy. En M. M. Haith & J. B. Benson (Eds.), *Encyclopedia of Infant and Early Childhood Development* (pp. 235–247). Academic Press.
<https://doi.org/https://doi.org/10.1016/B978-012370877-9.00094-3>
- Creswell, J. W., & Creswell, J. D. (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (5th ed.). SAGE Publications.

- CWI Shared Task 2018 - Datasets. (s. f.). Recuperado 1 de agosto de 2024, de <https://sites.google.com/view/cwisharedtask2018/datasets>
- De Belder, J., & Moens, M. F. (2012). A Dataset for the Evaluation of Lexical Simplification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7182 LNCS(PART 2), 426–437. https://doi.org/10.1007/978-3-642-28601-8_36
- Degraeuwe, J., & Goethals, P. (2024). LexComSpaL2: A Lexical Complexity Corpus for Spanish as a Foreign Language. En N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 10432–10447). ELRA and ICCL. <https://aclanthology.org/2024.lrec-main.912>
- Descartes, R. (1641). *Meditationes de Prima Philosophia*. Michael Soly.
- Diessel, H. (2004). *The acquisition of complex sentences*. Cambridge University Press.
- Educación secundaria en Lajas, Chota. (s. f.). Recuperado 6 de agosto de 2024, de <https://guia-cajamarca.portaldeeducacion.pe/secundaria/lajas-chota-cajamarca/index.htm>
- Ellis, N. C. (2002). FREQUENCY EFFECTS IN LANGUAGE PROCESSING: A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *Studies in Second Language Acquisition*, 24(2), 143–188. <https://doi.org/DOI: 10.1017/S0272263102002024>
- Evidently AI. (s. f.). *Accuracy vs. precision vs. recall in machine learning: what's the difference?* Recuperado 28 de octubre de 2024, de <https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall>
- Ferrés, D., & Saggion, H. (2022). ALEXSIS: A Dataset for Lexical Simplification in Spanish. En N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 3582–3594). European Language Resources Association. <https://aclanthology.org/2022.lrec-1.383/>
- Finnimore, P., Fritzsche, E., King, D., Sneyd, A., Ur Rehman, A., Alva-Manchego, F., & Vlachos, A. (2019). Strong Baselines for Complex Word Identification across Multiple Languages. En J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 970–977). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1102>
- Géron, A. (2022). *HANDS-ON MACHINE LEARNING WITH SCIKIT-LEARN, KERAS, AND TENSORFLOW concepts, tools, and techniques to build intelligent systems* (3^a ed.). O'Reilly Media, Inc.
- Ghosh, S., & Gunning, D. (2019). *Natural language processing fundamentals*.
- Gooding, S., & Kochmar, E. (2018). CAMB at CWI Shared Task 2018: Complex Word Identification with Ensemble-Based Voting. En J. Tetreault, J. Burstein, E. Kochmar, C. Leacock, & H. Yannakoudakis (Eds.), *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 184–194). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-0520>
- Gooding, S., & Tragut, M. (2022). One Size Does Not Fit All: The Case for Personalised Word Complexity Models. En M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2022* (pp. 353–365). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.findings-naacl.27>

- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3(null), 1157–1182.
- Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). *Metodología de la investigación* (6^a ed.). McGraw-Hill.
- Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, 2(2), 127–160. <https://doi.org/10.1007/BF00401799>
- Horn, C., Manduca, C., & Kauchak, D. (2014). Learning a Lexical Simplifier Using Wikipedia. En K. Toutanova & H. Wu (Eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 458–463). Association for Computational Linguistics. <https://doi.org/10.3115/v1/P14-2075>
- Hulstijn, J., & Graaff, R. (1994). Under what conditions does explicit knowledge of a second language facilitate the acquisition of implicit knowledge? A research proposal. *AILA Review*, 11.
- Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). *SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size*. <https://arxiv.org/abs/1602.07360v4>
- IBM. (2023, diciembre 5). *What Is Self-Supervised Learning?* | IBM. <https://www.ibm.com/topics/self-supervised-learning>
- IPE. (2022). *Evolución de la pobreza regional 2004-2022* | *Tablero interactivo*. <https://www.ipe.org.pe/portal/evolucion-de-la-pobreza-regional-2004-2021/>
- Ivanov, V. V. (2022). Sentence-level complexity in Russian: An evaluation of BERT and graph neural networks. *Frontiers in Artificial Intelligence*, 5, 1008411. <https://doi.org/10.3389/frai.2022.1008411>
- Jo, T. (2021). Machine learning foundations: Supervised, unsupervised, and advanced learning. *Machine Learning Foundations: Supervised, Unsupervised, and Advanced Learning*, 1–391. <https://doi.org/10.1007/978-3-030-65900-4>
- Johnson, J. (2020). *What's a Deep Neural Network? Deep Nets Explained – BMC Software | Blogs*. <https://www.bmc.com/blogs/deep-neural-network/>
- Junyent, A. (2017). La comprensión lectora en los niños peruanos. *Economía y Sociedad*, 32–37.
- Jurafsky, D., & Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (3rd ed.). <https://web.stanford.edu/~jurafsky/slp3/>
- Kajiwar, T., & Komachi, M. (2018). Complex Word Identification Based on Frequency in a Learner Corpus. En J. Tetreault, J. Burstein, E. Kochmar, C. Leacock, & H. Yannakoudakis (Eds.), *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 195–199). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-0521>
- Kauchak, D. (2016). Pomona at SemEval-2016 Task 11: Predicting Word Complexity Based on Corpus Frequency. En S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, & T. Zesch (Eds.), *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 1047–1051). Association for Computational Linguistics. <https://doi.org/10.18653/v1/S16-1164>
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: a highly efficient gradient boosting decision tree. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 3149–3157.
- Kelleher, J. D. (2019). *The Deep Learning*. En MIT Press.

- Kim, Y.-S. G. (2023). Learning to read and write. En R. J. Tierney, F. Rizvi, & K. Ercikan (Eds.), *International Encyclopedia of Education (Fourth Edition)* (Fourth Edition, pp. 335–343). Elsevier. <https://doi.org/https://doi.org/10.1016/B978-0-12-818630-5.14045-X>
- Kincaid, P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. <https://api.semanticscholar.org/CorpusID:61131325>
- LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6(2), 293–323. [https://doi.org/https://doi.org/10.1016/0010-0285\(74\)90015-2](https://doi.org/https://doi.org/10.1016/0010-0285(74)90015-2)
- Laufer, B. (2013). Lexical Thresholds for Reading Comprehension: What They Are and How They Can Be Used for Teaching Purposes. *TESOL Quarterly*, 47(4), 867–872. <https://doi.org/10.1002/TESQ.140>
- Lê, M. T., Wolinski, P., & Arbel, J. (2023). *Efficient Neural Networks for Tiny Machine Learning: A Comprehensive Review*. <https://arxiv.org/abs/2311.11883v1>
- Lee, J. S. Y., & Yeung, C. Y. (2018). *Personalizing Lexical Simplification* (pp. 224–232). <https://aclanthology.org/C18-1019>
- Locke, J. (1689). *An Essay Concerning Humane Understanding*. Thomas Basset.
- MINEDU. (2013). *Evaluación Censal de estudiantes ¿Cómo mejorar la Comprensión Lectora de nuestros estudiantes? Informe para el Docente*.
- MINEDU. (2018). *Resultados PISA 2018*. <http://umc.minedu.gob.pe/resultadospisa2018/>
- MINEDU. (2022). *Centros de Educación Básica Alternativa a nivel nacional 2022 - Informes y publicaciones - Ministerio de Educación - Plataforma del Estado Peruano*. <https://www.gob.pe/institucion/minedu/informes-publicaciones/2897223-centros-de-educacion-basica-alternativa-a-nivel-nacional-2022>
- MINEDU. (2023). *Resultados EM 2022 | UMC | Oficina de Medición de la Calidad de los Aprendizajes*. <http://umc.minedu.gob.pe/resultadossem2022/>
- Monica Melby-Lervåg, & Ame Lervåg. (2014). Reading Comprehension and Its Underlying Components in Second-Language Learners: A Meta-Analysis of Studies Comparing First- and Second-Language Learners. *Psychological Bulletin*. <https://doi.org/10.1037/a0033890.supp>
- Mosquera, A. (2021). Alejandro Mosquera at SemEval-2021 Task 1: Exploring Sentence and Word Features for Lexical Complexity Prediction. En A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, & X. Zhu (Eds.), *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* (pp. 554–559). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.semeval-1.68>
- Mukherjee, N., Patra, B. G., Das, D., & Bandyopadhyay, S. (2016). JU_NLP at SemEval-2016 Task 11: Identifying Complex Words in a Sentence. En S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, & T. Zesch (Eds.), *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 986–990). Association for Computational Linguistics. <https://doi.org/10.18653/v1/S16-1152>
- Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge University Press.
- North, K., & Zampieri, M. (2023). Features of lexical complexity: insights from L1 and L2 speakers. *Frontiers in Artificial Intelligence*, 6, 1236963. <https://doi.org/10.3389/FRAI.2023.1236963/BIBTEX>
- North, K., Zampieri, M., & Shardlow, M. (2023). Lexical Complexity Prediction: An Overview. *ACM Computing Surveys*, 55(9), 40. <https://doi.org/10.1145/3557885>

- Oakhill, J., Cain, K., & Elbro, C. (2019). Reading comprehension and reading comprehension difficulties. En *Reading Development and Difficulties: Bridging the Gap Between Research and Practice* (pp. 83–115). Springer International Publishing. https://doi.org/10.1007/978-3-030-26550-2_5
- OECD. (2019). *Reading performance and spending on education*. <https://doi.org/10.1787/7aa618c1-en>
- ONU. (2015, abril 22). *Educación - Desarrollo Sostenible*. <https://www.un.org/sustainabledevelopment/es/education/>
- ONU. (2016). *Educación de calidad: por qué es importante*. <http://www.un.org/>
- OPENAI. (2024). *New embedding models and API updates* | OpenAI. <https://openai.com/index/new-embedding-models-and-api-updates/>
- Ortiz-Zambrano, J. A., Espin-Riofrio, C., & Montejo-Ráez, A. (2022). Transformers for Lexical Complexity Prediction in Spanish Language [Transformers para la Predicción de la Complejidad Léxica en Lengua Española]. *Procesamiento del Lenguaje Natural*, 69, 177–188. <https://doi.org/10.26342/2022-69-15>
- Ortiz-Zambrano, J. A., Espin-Riofrio, C., & Montejo-Ráez, A. (2023a). Combining Transformer Embeddings with Linguistic Features for Complex Word Identification. *Electronics (Switzerland)*, 12(1), 120. <https://doi.org/10.3390/electronics12010120>
- Ortiz-Zambrano, J. A., Espin-Riofrio, C., & Montejo-Ráez, A. (2023b). LegalEc: A New Corpus for Complex Word Identification Research in Law Studies in Ecuatorian Spanish; [LegalEc: Un nuevo corpus para la investigación de la identificación de palabras complejas en los estudios de Derecho en español ecuatoriano]. *Procesamiento del Lenguaje Natural*, 71, 247 – 259. <https://doi.org/10.26342/2023-71-19>
- Ortiz-Zambrano, J. A., & Montejo-Ráez, A. (2021). *SINAI at SemEval-2021 Task 1: Complex word identification using Word-level features*. <https://pypi>.
- Paetzold, G. H., & Specia, L. (2016a). SemEval 2016 task 11: Complex word identification. *SemEval 2016 - 10th International Workshop on Semantic Evaluation, Proceedings*, 560–569. <https://doi.org/10.18653/V1/S16-1085>
- Paetzold, G. H., & Specia, L. (2016b). Unsupervised Lexical Simplification for Non-Native Speakers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), 3761–3767. <https://doi.org/10.1609/AAAI.V30I1.9885>
- Paetzold, G. H., & Specia, L. (2017). A Survey on Lexical Simplification. *Journal of Artificial Intelligence Research*, 60, 549–593.
- Paetzold, G., & Specia, L. (2016). SemEval 2016 Task 11: Complex Word Identification. En S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, & T. Zesch (Eds.), *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 560–569). Association for Computational Linguistics. <https://doi.org/10.18653/v1/S16-1085>
- Palmas, Alessandro., Ghelfi, Emanuele., Petre, A. Galina., Kulkarni, Mayur., N.S., Anand., Nguyen, Quan., Sen, Aritra., So, A. (Data scientist), & Basak, Saikat. (2020). *The Reinforcement Learning Workshop Learn How to Apply Cutting-Edge Reinforcement Learning Algorithms to a Wide Range of Control Problems*.
- Pan, C., Song, B., Wang, S., & Luo, Z. (2021). DeepBlueAI at SemEval-2021 Task 1: Lexical Complexity Prediction with A Deep Ensemble Approach. En A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, & X. Zhu (Eds.), *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* (pp. 578–584). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.semeval-1.72>

- Patel, A. A. (2021). *Applied Natural Language Processing in the Enterprise teaching machines to read, write,... and understand*. O'REILLY MEDIA.
- Peirce, C. S. (1878). How to Make Our Ideas Clear. *Popular Science Monthly*, 12, 286–302.
- Perfetti, C. (2007). Reading Ability: Lexical Quality to Comprehension. *Scientific Studies of Reading*, 11(4), 357–383. <https://doi.org/10.1080/10888430701530730>
- Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, 18(1), 22–37.
- Pimienta Castillo, J. (2021). *Multilingual lexical simplification*. <http://repositori.upf.edu/handle/10230/49224>
- RAE. (s. f.-a). 1.2.1 unidades de la morfología | *Nueva gramática básica de la lengua española* | RAE - ASALE. Recuperado 15 de mayo de 2025, de <https://www.rae.es/gram%C3%A1tica-b%C3%A1sica/la-gram%C3%A1tica/la-morfolog%C3%ADa/unidades-de-la-morfolog%C3%ADa>
- RAE. (s. f.-b). 1.3 La sintaxis | *Nueva gramática básica de la lengua española* | RAE - ASALE. Recuperado 15 de mayo de 2025, de <https://www.rae.es/gram%C3%A1tica-b%C3%A1sica/la-gram%C3%A1tica/la-sintaxis>
- RAE. (s. f.-c). CREA | *Real Academia Española*. Recuperado 8 de septiembre de 2024, de <https://www.rae.es/banco-de-datos/crea>
- RAE. (s. f.-d). fonología | *Glosario de términos gramaticales* | RAE - ASALE. Recuperado 15 de mayo de 2025, de <https://www.rae.es/gtg/fonolog%C3%ADa>
- RAE. (s. f.-e). Presentación del «Diccionario de la lengua española» y sus ediciones | *Real Academia Española*. Recuperado 1 de agosto de 2024, de <https://www.rae.es/obras-academicas/diccionarios/presentacion-del-diccionario-de-la-lengua-espanola-y-sus-ediciones>
- RAE. (s. f.-f). semántica | *Glosario de términos gramaticales* | RAE - ASALE. Recuperado 15 de mayo de 2025, de <https://www.rae.es/gtg/sem%C3%A1tica>
- RAE. (2024a). CORPES XXI | *Real Academia Española*. <https://www.rae.es/banco-de-datos/corpes-xxi>
- RAE. (2024b, abril). CORPES. <https://www.rae.es/corpes/>
- Ray, S. (2019). A Quick Review of Machine Learning Algorithms. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 35–39. <https://doi.org/10.1109/COMITCon.2019.8862451>
- Rogde, K., Hagen, Å. M., Melby-Lervåg, M., & Lervåg, A. (2019). The effect of linguistic comprehension instruction on generalized language and reading comprehension skills: A systematic review. En *Campbell Systematic Reviews* (Vol. 15, Número 4). Wiley Blackwell. <https://doi.org/10.1002/cl2.1059>
- Rozi, E., Iyer, N., Chi, G., Choe, E., Lee, K. J., Liu, K., Liu, P., Lack, Z., Tang, J., & Chi, E. A. (2021). Stanford MLab at SemEval-2021 Task 1: Tree-Based Modelling of Lexical Complexity using Word Embeddings. En A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, & X. Zhu (Eds.), *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* (pp. 688–693). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.semeval-1.89>
- Saunders, M., Lewis, P., & Thornhill, A. (2019). *RESEARCH METHODS FOR BUSINESS STUDENTS* (8ª ed.). www.pearson.com/uk
- Shardlow, M. (2013a). A Comparison of Techniques to Automatically Identify Complex Words. En A. Dey, S. Krause, I. Nikolova, E. Vecchi, S. Bethard, P. I. Nakov, & F. Xu (Eds.), *51st*

- Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop* (pp. 103–109). Association for Computational Linguistics. <https://aclanthology.org/P13-3015/>
- Shardlow, M. (2013b). The CW Corpus: A New Resource for Evaluating the Identification of Complex Words. En S. Williams, A. Siddharthan, & A. Nenkova (Eds.), *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations* (pp. 69–77). Association for Computational Linguistics. <https://aclanthology.org/W13-2908/>
- Shardlow, M., Cooper, M., & Zampieri, M. (2020). *CompLex — A New Corpus for Lexical Complexity Prediction from Likert Scale Data* (pp. 57–62). <https://aclanthology.org/2020.readi-1.9>
- Shardlow, M., Evans, R., & Zampieri, M. (2022). Predicting lexical complexity in English texts: the Complex 2.0 dataset. *Language Resources and Evaluation*, 56(4), 1153–1194. <https://doi.org/10.1007/s10579-022-09588-2>
- Shardlow, M., North, K., & Zampieri, M. (2024). Multilingual Resources for Lexical Complexity Prediction: A Review. *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, 51–59. <https://aclanthology.org/2024.determin-1.5>
- Sheang, K. C. (2019). Multilingual Complex Word Identification: Convolutional Neural Networks with Morphological and Linguistic Features. En V. Kovatchev, I. Temnikova, B. Šandrih, & I. Nikolova (Eds.), *Proceedings of the Student Research Workshop Associated with RANLP 2019* (pp. 83–89). INCOMA Ltd. https://doi.org/10.26615/issn.2603-2821.2019_013
- Smartick. (2022). *Informe Reto de lectura*. <https://www.smartick.com/informe-reto-lectura>
- Snow, C. E. (2010). Reading Comprehension: Reading for Learning. En P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education (Third Edition)* (Third Edition, pp. 413–418). Elsevier. <https://doi.org/10.1016/B978-0-08-044894-7.00511-X>
- Sowmya V. B., Majumder, B., Gupta, A. (Computer scientist), & Surana, H. (2020). *Practical Natural Language Processing: A Pragmatic Approach to Processing and Analyzing Language Data*. O'Reilly.
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4), 295–312. [https://doi.org/10.1016/0959-4752\(94\)90003-5](https://doi.org/10.1016/0959-4752(94)90003-5)
- Takahashi, K., Yamamoto, K., Kuchiba, A., & Koyama, T. (2022). Confidence interval for micro-averaged F1 and macro-averaged F1 scores. *Applied Intelligence*, 52(5), 4961–4972. <https://doi.org/10.1007/s10489-021-02635-5>
- Tunstall, L., Werra, L. von, & Wolf, T. (Of H. (2022). *Natural language processing with transformers : building language applications with Hugging Face*. O'REILLY.
- Universidad de Guanajuato. (2022, abril 27). *Clase digital 5. Factores que influyen en el aprendizaje y la enseñanza - Maestría en Docencia Universitaria para la Educación Digital*. <https://blogs.ugto.mx/mdued/clase-digital-5-factores-que-influyen-en-el-aprendizaje-y-la-ensenanza/>
- Vygotsky, L. S. (1978). *Development of Higher Psychological Processes* (M. Cole, V. Jolm-Steiner, S. Scribner, & E. Soubberman, Eds.). Harvard University Press. <https://doi.org/10.2307/j.ctvjf9vz4>

- Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., & Zampieri, M. (2018). A Report on the Complex Word Identification Shared Task 2018. En J. Tetreault, J. Burstein, E. Kochmar, C. Leacock, & H. Yannakoudakis (Eds.), *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 66–78). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-0507>
- Zheng, A. (2015). Evaluating Machine Learning Algorithms. En O'Reilly Media (Ed.), *Springer*.
- Zhu, X., & Goldberg, A. B. (2009). Introduction to Semi-Supervised Learning. *Introduction to Semi-Supervised Learning*. <https://doi.org/10.1007/978-3-031-01548-9>

ANEXO A – CUESTIONES METODOLÓGICAS

	Problema	Objetivo	Hipótesis	Independiente	Dependiente
G	¿Cuál es la efectividad de un modelo de machine learning basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica de palabras, en comparación con un baseline del CWI Shared Task 2018, en instituciones secundarias en un distrito de Cajamarca?	Determinar la efectividad de un modelo de machine learning basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica de palabras, en comparación con un baseline del CWI Shared Task 2018, en instituciones secundarias en un distrito de Cajamarca.	El modelo de machine learning basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica de palabras, es efectivo al lograr un rendimiento que supera al baseline del CWI 2018, en instituciones en secundarias en un distrito de Cajamarca.	Modelo de machine learning	Predicción binaria de la complejidad léxica de palabras en instituciones secundarias en un distrito de Cajamarca
E1	¿Cuáles de las variables seleccionadas, basadas en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, entrenadas en el dataset CWI 2018, son más importantes al evaluarlas en el dataset de Textos Escolares de instituciones secundarias de un distrito de Cajamarca?	Identificar las variables seleccionadas basadas en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, entrenadas en el dataset CWI 2018, que son más importantes al evaluarlas en el dataset de Textos Escolares de instituciones secundarias de un distrito de Cajamarca.	Al menos la mitad de las variables seleccionadas basadas en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, entrenadas en el dataset CWI 2018, son más importantes al evaluarlas en el dataset de Textos Escolares de instituciones secundarias de un distrito de Cajamarca.	Variables de entrada del modelo	Variable de salida del modelo
E2	¿Qué algoritmo de machine learning evaluado y basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, obtiene el mejor rendimiento en el dataset CWI 2018 para proponerse a instituciones secundarias de un distrito de Cajamarca?	Identificar el algoritmo de machine learning evaluado y basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, que obtiene el mejor rendimiento en el dataset CWI 2018 para proponerse a instituciones secundarias de un distrito de Cajamarca.	Existe al menos un algoritmo de machine learning evaluado y basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica que obtiene un rendimiento superior en el dataset CWI 2018, para proponerse a instituciones secundarias de un distrito de Cajamarca.	Algoritmo de machine learning	Rendimiento de los algoritmos en esta tarea
E3	¿Cuál es el rendimiento y la eficiencia en memoria de la arquitectura del modelo basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, al compararlo con un baseline del CWI 2018, considerando los requisitos de memoria RAM viables para las instituciones secundarias de un distrito de Cajamarca?	Determinar el rendimiento y la eficiencia en memoria de la arquitectura del modelo basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica, al compararlo con un baseline del CWI 2018, considerando los requisitos de memoria RAM viables para las instituciones secundarias de un distrito de Cajamarca.	La arquitectura del modelo basado en embeddings modernos, frecuencias y propiedades léxicas para la predicción binaria de la complejidad léxica logra un rendimiento superior a un baseline del CWI 2018 manteniendo los requisitos de memoria RAM viables para las instituciones secundarias de un distrito de Cajamarca.	Arquitectura de modelo de machine learning	Precisión del modelo

Nota. Matriz de consistencia. Elaboración propia.

Matriz de operacionalización

	Dimensión	Indicadores	Definición operacional
Modelo de machine learning	Variables de entrada del modelo	F1 (macro)	Promedio aritmético del F1-score calculado independientemente para cada clase utilizando python
		Importancia de las características	Medir importancia normalizada de las características del modelo entrenado utilizando python
		Importancia de características mediante el método de permutación	Medir importancia de las características del modelo mediante el método de permutación utilizando python
	Algoritmo de machine learning	Existencia de estudios que respalden la efectividad de algoritmos	Revisión de literatura con mejores resultados para experimentos con algoritmos utilizados para el problema de la predicción de la complejidad léxica en forma binaria
		F1 (macro)	Calcular el promedio aritmético del F1-score calculado independientemente para cada clase utilizando python
	Arquitectura del modelo de machine learning	F1 (macro)	Promedio aritmético del F1-score calculado independientemente para cada clase utilizando python
		Memoria promedio en la carga del modelo	Promediar el consumo de memoria en MB para carga del modelo en 10 iteraciones utilizando python

	Dimensión	Indicadores	Definición operacional
Predicción de complejidad léxica de las palabras en instituciones secundarias en un distrito de Cajamarca	Variable de salida del modelo	Identificación de palabra como compleja o no (1 ó 0)	Producto de la predicción del modelo que se construye
	Precisión del modelo	Accuracy	Proporción de predicciones correctas realizadas por el modelo sobre el total de predicciones
		Confusion matrix	Tabla que permite visualizar el rendimiento del algoritmo mostrando verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos
		Mean absolute error	Promedio de los errores absolutos entre las predicciones del modelo y los valores reales
	Precisión del algoritmo	Accuracy	Proporción de predicciones correctas realizadas por el modelo sobre el total de predicciones
		Confusion matrix	Tabla que permite visualizar el rendimiento del algoritmo mostrando verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos
		Mean absolute error	Promedio de los errores absolutos entre las predicciones del modelo y los valores reales

Nota. Matriz de operacionalización. Elaboración propia.

ANEXO B – ACERCA DE LOS CONJUNTOS DE DATOS

Tabla B - 1

Numero de instancias del dataset CWI Shared Task 2018

Idioma	Entrenamiento	Desarrollo	Prueba
Inglés	27,299	3,328	4,252
Español	13,750	1,622	2,233
Alemán	6,151	795	959
Francés	-	-	2,251

Nota. Número de instancias para cada conjunto de entrenamiento, desarrollo y prueba. Esta tabla ha sido adaptada de “Data Format” en CWI Shared Task 2018 - Datasets, 2018.

Tabla B - 2

Formato del dataset CWI Shared Task 2018

Columna	Descripción
HIT ID	Oraciones con el mismo pertenecen al mismo HIT
Oración	Oración original con palabra compleja o no
Inicio de Palabra	Índice de inicio de la palabra objetivo
Fin de Palabra	Índice final de la palabra objetivo
Palabra Objetivo	La palabra que puede ser etiquetada como compleja o no
Total de Anotadores Nativos	Número de anotadores nativos que vieron la oración
Total de Anotadores No Nativos	Número de anotadores no nativos que vieron la oración
Nativos que marcaron como difícil	Nativos que marcaron la palabra como difícil
No nativos que marcaron como difícil	No nativos que marcaron la palabra como difícil
Etiqueta Binaria	0 para palabra simple, 1 para palabra compleja
Etiqueta Probabilística	Probabilidad de que la palabra sea compleja (0.0 a 1.0)

Nota. Campos presentes en el dataset de la tarea compartida CWI Shared Task 2018. Tabla adaptada de “Data Format” en CWI Shared Task 2018 - Datasets, 2018.

Tabla B - 3

Ejemplos de instancias provistas por el dataset CWI Shared Task 2018

ID	Oración	Inicio de palabra	Fin de palabra	Palabra objetivo	Anotadores nativos	Anotadores no nativos	Nativos marcaron como difícil	No nativos marcaron como difícil	Etiqueta binaria	Etiqueta probabilística
3CZH926 SICFIJS D3QQZQ 6U7J1FR E41	El jugador puede obtener puntos por derrotar a varios enemigos uno tras otro, y luego podrá participar en una ronda de bonos que le da la oportunidad de reunir más puntos.	119	124	bonos	10	0	1	0	1	0.1
3CZH926 SICFIJS D3QQZQ 6U7J1FR E41	Otro factor que contribuyó fue el escenario del juego: se trataba de una gran red de tuberías gigantes, por lo que se consideró que un cambio de ocupación era necesario.	16	26	contribuyó	10	0	1	0	1	0.1
3KQC8J MJGCT9 L6L1GD 2DHSCQ FVNH3Z	Normalmente hay entre seis y ocho eventos por competición, rara vez llegan a ser nueve eventos.	87	94	eventos	10	0	0	0	0	0.0

Nota. Ejemplos de instancias anotadas como compleja (1) y no compleja (0). Esta tabla ha sido adaptada de CWI Shared Task 2018 – Datasets.

Tabla B - 4

Ejemplos de instancias recopiladas para el dataset de validación basado en Textos Escolares

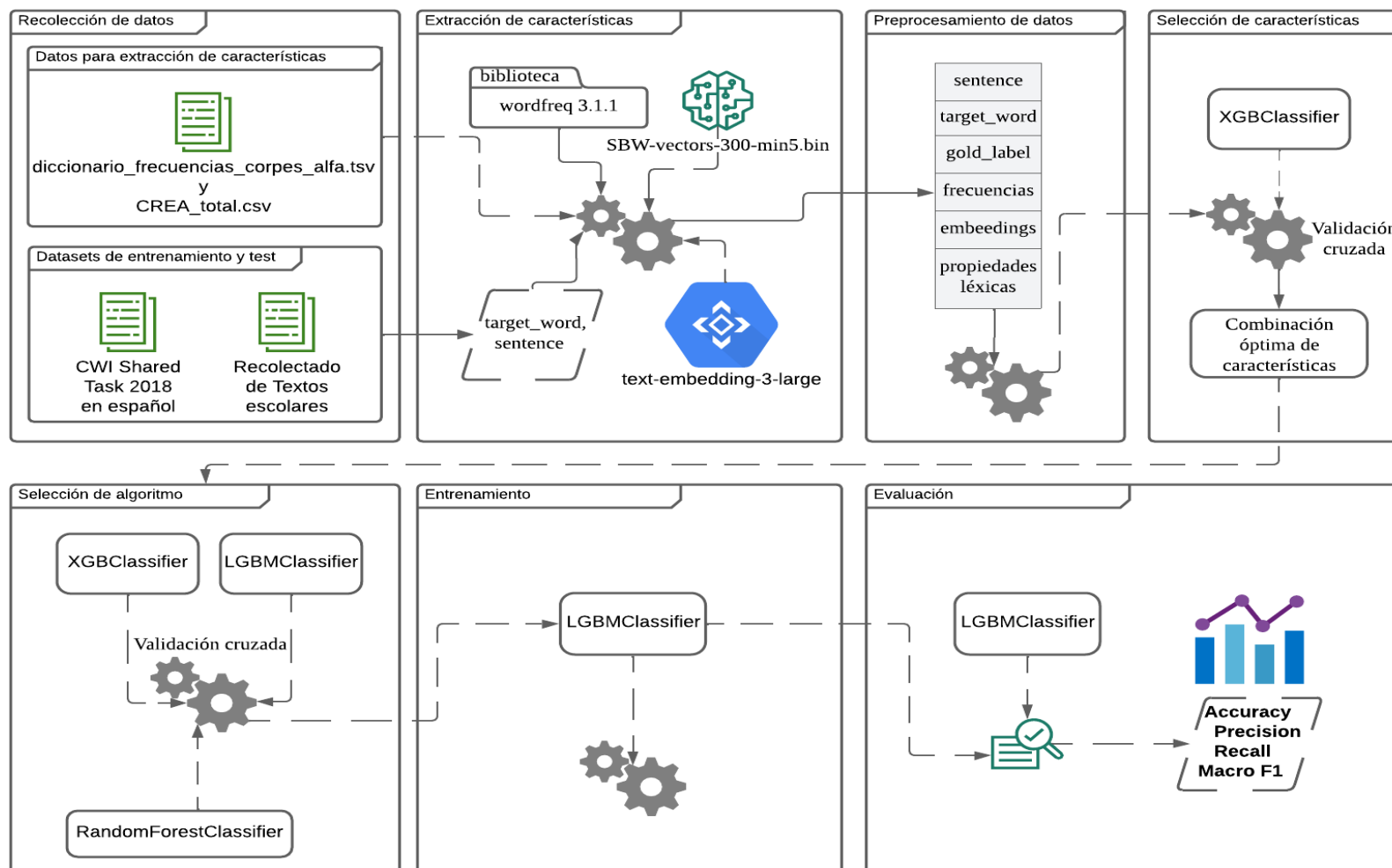
PALABRA	ORACIÓN	COMPLEJIDAD
alquitrán	El ladrillo reemplazó la piedra y el alquitrán les sirvió de mezcla	1
atinaron	Otros solo atinaron a mirar la inmensa masa de agua de lejos cautelosos	1
reminiscencia	Incluso les mostraría algunas blancas conchitas y arena que guardó en sus bolsillos como una infantil reminiscencia	1
introspección	Hoy la atención y la introspección están disminuyendo.	1
suburbio	“Como que era algo perezosa”, admite la joven de los suburbios de Chicago	1
aversión	¿Existe alguna aversión a hablar por el teléfono?	1
remoto	Solamente salió del pueblo para acompañar a su esposo a remotas montañas a las que llegaban tras muchos días de llegar a pie.	1
ancestrales	A una hora en auto desde Tarapoto Ishichihui es una de las tantas comunidades nativas quechua de San Martín que luchan por conservar las tradiciones ancestrales	1
galaxias	Había una vez la tierra y había una vez marte Estaban muy distantes el uno del otro en medio del cielo y alrededor había millones de planetas y de galaxias	1
satélites	Primero lanzaron satélites que giraban dos días alrededor de la tierra y luego regresaban	1
llanura	Pero al emigrar los hombres desde oriente encontraron una llanura en la región de Sinear	0
hermosa	Tal vez no lo asociamos fácilmente debido a todo el calor y frondosa vegetación de esta hermosa región	0
tiestos	Teje desde niña pretinas poleras chumbis y hace tiestos de barro	0
preferían	En aquellos años en las familias quechuas los padres preferían enviar a los hijos a la escuela antes que a las hijas	0
maña	Incluso ahora cuando las personas para hablar en lengua nativa escasean se da maña para expresarse en su idioma	0
apuesta	Carlos apuesta por su propia pertinencia al asumir que Pedro ha visto al Pastor y sus cabras.	0
astronauta	Como los tres eran muy buenos astronautas llegaron a marte casi en el mismo momento y encontraron allí un paisaje maravilloso e inquietante	0
lenteja	Eso te pasa por lenteja	0
luca	¡Qué lenteja! Sino tenía ni una luca	0
tacañazo	Lo que pasa es que tú eres un tacañazo	0

Nota. Ejemplos de instancias anotadas como compleja (1) y no compleja (0) para el dataset de validación. Esta tabla ha sido adaptada de Comunicación 1: Secundaria (2016), Comunicación 2: Secundaria, (2016), Comunicación 3: Educación Secundaria (2016), Comunicación 4: Educación Secundaria (2016), Comunicación 5: Educación Secundaria (2016).

ANEXO C – PIPELINE DE MACHINE LEARNING

Figura C - 1

Pipeline de machine learning utilizado para el proceso de desarrollo y operación del modelo de machine learning para predicción binaria de la complejidad léxica de palabras en instituciones secundarias en un distrito de Cajamarca



ANEXO D – SOLICITUD DE PRÉSTAMO DE MATERIALES E INFORMACIÓN PARA LA INVESTIGACIÓN

"AÑO DEL BICENTENARIO, DE LA CONSOLIDACION DE NUESTRA INDEPENDENCIA, Y DE LA
CONMEMORACION DE LAS HEROICAS BATALLAS DE JUNIN Y AYACUCHO"

SEÑOR DIRECTOR DE LA INSTITUCION EDUCATIVA SECUNDARIA (EBR.) "ANAXIMANDRO VEGA"-
LAJAS.

**SOLICITA: Apoyo en calidad de préstamo de Materiales para investigación
de tesis en Maestría de Ingeniería de Sistemas.**

Yo, Mera Dávila Luis Iván, con DNI. 46396150, con el debido respeto expongo lo siguiente:

1ro. A la fecha estoy siguiendo estudios de posgrado de una Maestría en Ingeniería de Sistemas en la Universidad Nacional Mayor de San Marcos.

2do. Mi solicitud la realizo en el marco de la investigación para mi tesis de Maestría en Ingeniería de Sistemas e Informática con mención en Gestión de Tecnología de Información y Comunicaciones, titulada: "MODELO DE MACHINE LEARNING PARA LA PREDICCIÓN DE LA COMPLEJIDAD LÉXICA DE PALABRAS EN INSTITUCIONES SECUNDARIAS EN UN DISTRITO DE CAJAMARCA".

Y para estos estudios me es necesario contar con los siguientes materiales:

- TEXTOS ESCOLARES DE COMUNICACIÓN DEL 1RO AL 5TO GRADO.
- ADICIONALMENTE, ME SERÍA DE GRAN UTILIDAD OBTENER INFORMACIÓN SOBRE LAS ESPECIFICACIONES TÉCNICAS DE LOS EQUIPOS DE CÓMPUTO CON LOS QUE CUENTA EL COLEGIO.
- Y OTROS QUE PODRIA NECESITAR EN TORNO A MI INVESTIGACION.

POR LO TANTO:

Suplico a Ud. Señor Director atender a mi solicitud.

Lajas, 04 de marzo del 2024.


Luis Iván Mera Dávila.



2/04/03/2024
H. 8 AM.

Lic. C. Rosa Dávila González
Eso. Lengua y Literatura




DIRECTOR
DNI 27289925
R: 25/03/2024