Revista Peruana de Computación y Sistemas 2024 6(2):55-64
https://doi.org/10.15381/rpcs.v6i2.29211

ISSN Online: 2617-2003
Facultad de Ingeniería de Sistemas e Informática
Universidad Nacional Mayor de San Marcos

Artículo de Contribución

Integración de embeddings de nueva generación y recursos lingüísticos actuales para identificar palabras complejas en español con machine learning

Integration of new generation embeddings and current linguistic resources to identify complex words in Spanish with machine learning

Luis Iván Mera Dávila[1,a]

[1] Universidad Nacional Mayor de San Marcos, Lima, Perú

[a] Autor de correspondencia: limd1990@gmail.com, https://orcid.org/0009-0002-7765-2691

Resumen

La complejidad de las palabras puede suponer una limitación para la accesibilidad de la información, lo que podría afectar a millones de personas hispanohablantes. El objetivo de este estudio es desarrollar un modelo de machine learning para la tarea binaria de identificación de palabras complejas en español, usando embeddings de nueva generación, recursos lingüísticos actuales y propiedades léxicas. Para ello se empleó el conjunto de datos en español de la tarea compartida CWI Shared Task 2018, obteniendo embeddings generados por el modelo text-embedding-3-large y frecuencias de palabras extraídas de recursos como el Corpus del Español del Siglo XXI, el Corpus de Referencia del Español Actual, el Spanish Billion Word Corpus and Embeddings y Wordfreq. Para seleccionar características y encontrar su mejor combinación se usó una validación cruzada de 5 pliegues utilizando XGBClassifier. Una vez comparados varios algoritmos de machine learning, el modelo final, basado en LGBMClassifier, obtuvo el macro F1 de 0.7993, logrando superar al mejor equipo de dicha competencia, a estudios más recientes que utilizaron redes neuronales y a algunos modelos de lenguaje grandes. Esto muestra el potencial de estos recursos que constantemente están actualizándose y que pueden contribuir a mejorar la precisión de esta tarea.

Palabras clave: Identificación de palabras complejas, Embeddings, Simplificación Léxica, Español.

Abstract

The complexity of words can pose a limitation to the accessibility of information, which could affect millions of Spanish-speaking people. The objective of this study is to develop a machine learning model for the binary task of identifying complex words in Spanish, using next-generation embeddings, current linguistic resources, and lexical properties. To this end, the Spanish dataset from the CWI Shared Task 2018 was used, obtaining embeddings generated by the text-embedding-3-large model and word frequencies extracted from resources such as the Corpus del Español del Siglo XXI, the Corpus de Referencia del Español Actual, the Spanish Billion Word Corpus and Embeddings, and Wordfreq. To select features and find their best combination, a 5-fold cross-validation using XGBClassifier was employed. After comparing several machine learning algorithms, the final model, based on LGBMClassifier, achieved a macro F1 score of 0.7993, surpassing the best team from that competition, more recent studies that used neural networks, and some large language models. This demonstrates the potential of these resources that are constantly being updated and that can contribute to improving the accuracy of this task.

Keywords: Complex word identification, Embeddings, Lexical Simplification, Spanish.

Recibido: 15-10-2024 - Aceptado: 12-12-2024 - Publicado: 30-12-2024

Citar como:
Luis Iván Mera Dávila (2024). Integración de embeddings de nueva generación y recursos lingüísticos actuales para identificar palabras complejas en español con machine learning. Revista Peruana de Computación y Sistemas, 6(2):55-64. https://doi.org/10.15381/rpcs.6i2.29211

1. Introducción

En la actualidad, en la era digital, la falta de acceso a la información y la comprensión del lenguaje escrito pueden suponer la afectación de la vida de millones de personas. El lenguaje, por su complejidad, puede convertirse en una barrera difícil de superar para el acceso a la educación, la inclusión social, la accesibilidad, la obtención de servicios básicos, etc. En este sentido, la simplificación léxica (LS) en español, una de las lenguas más habladas del mundo, es un ámbito de investigación importante.

El procesamiento del lenguaje natural (NLP) es esencial para enfrentar estos retos, dado que mejora la comunicación entre los humanos y las máquinas por medio de la utilización del lenguaje que las personas usan para interactuar [1]. Por su parte, la identificación de palabras complejas (CWI), es un aspecto básico del proceso de simplificación léxica [2]. Es muy importante estudiarla en el idioma español, ya que, según [3], este es hablado por aproximadamente 600 millones de personas, representando el 7.5% de la población mundial y es el segundo más hablado como lengua materna después del chino mandarín.

En el campo de la accesibilidad e inclusión, la simplificación léxica tiene un efecto importante. De acuerdo con [4], esta técnica facilita la comprensión lectora al reducir la complejidad del vocabulario y hacer los textos más accesibles, beneficiando a personas con dislexia, afasia, problemas de audición, trastorno del espectro autista, discapacidad intelectual, aprendices de un segundo idioma y niños, quienes tienen dificultades con palabras largas, infrecuentes, homófonas, similares, nuevas o inexistentes, así como con la gramática compleja y el lenguaje figurado.

En [5], se subraya que la simplificación léxica facilita el acceso a la información para individuos con niveles rudimentarios y básicos de alfabetización, al hacer los textos más accesibles y comprensibles. En el contexto educativo, una herramienta que resulta de gran importancia es la identificación automática de palabras complejas, dado que puede formar parte de sistemas de tutoría inteligente, así como de plataformas de instrucción por internet, personalizando así materiales de aprendizaje ante las necesidades particulares del alumnado [6], [7].

Por otra parte, en [8] se menciona que se espera que el contenido digital mundial crezca desmesurada y exponencialmente desde los 33 zettabytes (ZB) en 2018 a 175 ZB en el año 2025. Esto sugiere que gran parte de este contenido será texto, lo que subraya la necesidad de herramientas automáticas capaces de analizarlo y simplificarlo para mejorar su accesibilidad y comprensión.

En cuanto al estado actual del conocimiento sobre la identificación de palabras complejas, se han realizado estudios que van desde enfoques como: considerar palabras con más de tres sílabas [9], verificar si la palabra estaba en una lista específica [10], modelos basados en transformadores [11], hasta grandes modelos de lenguaje [12]. Sin embargo, la mayoría de estos estudios se han centrado en el idioma inglés; en otros como el español, han recibido menos atención en términos de investigaciones [13] y recursos [11], [14], [15], [16], [17].

Los embeddings han evolucionado significativamente, mejorando la representación del lenguaje en diversas tareas de NLP. Sin embargo, en la identificación de palabras complejas se han encontrado limitaciones importantes. En la investigación de [18], se demostró que ciertos modelos preentrenados no capturaron adecuadamente la complejidad de una palabra o expresión en diferentes contextos o dominios, mientras que en [19] se encontró que algunos embeddings contextualizados no mostraron un rendimiento significativamente superior. Por otro lado, las nuevas versiones de recursos como CORPES XXI, ofrecen un corpus actualizado con muchas incorporaciones importantes respecto a sus versiones anteriores [20]. Esto representa una oportunidad de mejora en la precisión de modelos de identificación de palabras complejas.

El objetivo de este estudio es desarrollar un modelo de machine learning que combine embeddings de nueva generación con recursos lingüísticos del español actual. Esto se realiza considerando las sugerencias de [14] sobre los márgenes de mejora que existen en todas las etapas de la simplificación léxica, de [21] para aprovechar los embeddings contextuales de punta en la mejora de la precisión de la predicción de la complejidad léxica, y de [22] de explorar modelos de embeddings más recientes. La finalidad de este modelo es examinar el potencial de dicha combinación en la identificación de palabras complejas en español, con miras a contribuir al avance del conocimiento en el campo y facilitar el desarrollo de herramientas más precisas para la simplificación léxica en este idioma.

1.1. Simplificación Léxica

En [2], se conceptualiza la simplificación léxica como un proceso secuencial de cuatro etapas interconectadas: (1) Identificación de Palabras Complejas, donde se detectan términos potencialmente difíciles para el lector objetivo; (2) Generación de Sustituciones, que implica la búsqueda de alternativas más simples; (3) Selección de Sustituciones, en la que se eligen las palabras más apropiadas considerando el contexto; y (4) Clasificación de Sustituciones, en el cual las alternativas de sustitución son tomadas en cuenta en función de su dificultad, considerando primeramente las más sencillas.

1.2. Complejidad de Palabras

El concepto de "complejidad" tiene un significado muy diverso, dependiendo de lo que se quiera estudiar, tal como plantean [23]. Según [24], la complejidad absoluta y la complejidad relativa son dos de las aproximaciones más utilizadas dentro de este dominio.

La complejidad absoluta, según [25], viene dada por las propiedades lingüísticas objetivas de las palabras, en las cuales se encuentran varios niveles: morfológico (estructura interna y formas flexivas), sintáctico (organización de palabras y frases, así como estructuras jerárquicas), léxico (tamaño, frecuencia y diversidad del vocabulario), fonológico (estructura sonora y organización silábica), y semántico (significado y polisemia).

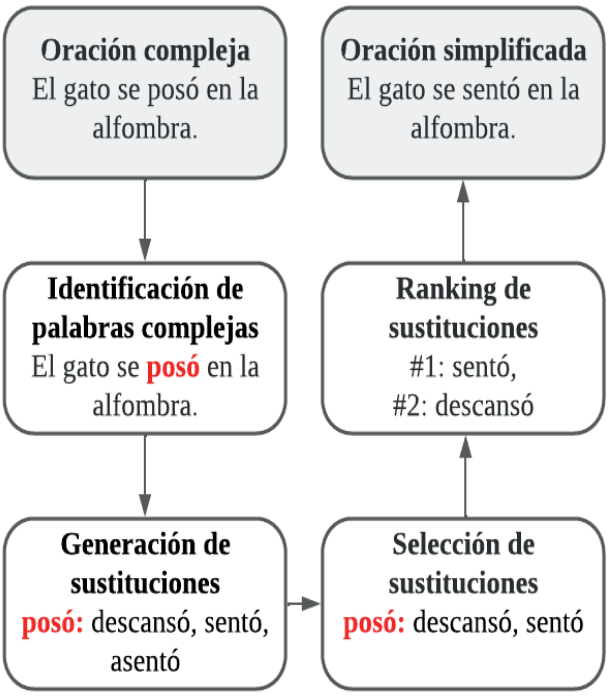
La complejidad relativa del lenguaje se define de acuerdo a sus usuarios, es decir, en función de la dificultad mental para aprender, procesar o utilizar determinados aspectos en el uso de los elementos lingüísticos [26], pero, a la vez, se trata de un concepto subjetivo que depende de un conjunto de factores individuales y de otros más objetivos, como propiedades de percepción (qué tan fácil es detectar una característica lingüística) o la frecuencia en el input (con qué frecuencia se da una característica lingüística o con qué frecuencia aparece en el lenguaje al que se expone la persona) [27].

1.3. Identificación de Palabras Complejas

La Identificación de Palabras Complejas es determinar qué palabras deberían ser simplificadas en un texto, puede formar parte de un proceso más grande que es la simplificación léxica y cuyo objetivo es cambiar palabras y frases complejas por sus equivalentes que sean más simples [28]. La Figura 1 muestra un pipeline del proceso de simplificación léxica donde el subproceso de identificación de palabras complejas es uno de los pasos más relevantes del mismo.

Figura 1

Pipeline de simplificación léxica. Adaptado de [2]



2. Métodos y Materiales

2.1. Tipo de Investigación

Esta investigación es cuantitativa, ya que busca evaluar la efectividad de un modelo de machine learning, así como la evaluación objetiva de resultados mediante métricas numéricas, siguiendo lo que indican [29], quienes sostienen que este enfoque permite una aproximación cuantitativa y objetiva al problema de investigación, haciendo uso del análisis de datos numéricos y la aplicación de técnicas estadísticas.

2.2. Período y Contexto

El período de esta investigación es parte del año 2024. No se centra en un lugar específico, ya que se enfoca en el idioma español y utiliza recursos y conjuntos de datos públicos.

2.3. Población y Muestra

La población de esta investigación se centra en las palabras en español, que de acuerdo con el Diccionario de la Real Academia Española [30], existen 93000 lemas, lo que sugiere que la cantidad de palabras en este idioma es aún mayor y más diversa.

La muestra seleccionada es el conjunto de datos del CWI Shared Task 2018 en español, debido a su anotación por hablantes nativos, enfoque general y reconocimiento internacional, adecuado para este estudio, especialmente considerando la escasez de recursos similares en esta lengua.

2.4. Recolección de Datos

En esta investigación se utilizó el método de recolección de datos secundarios, ya que se reutilizó el conjunto de datos del CWI Shared Task 2018, es decir, se empleó información ya recopilada y procesada por otros investigadores, adaptándola a los objetivos de este estudio [31].

Dicho conjunto de datos es un corpus en inglés, español, alemán y francés, que se creó para una competencia centrada en identificar palabras complejas, sirviendo como base en una tarea compartida que desafió a investigadores y programadores a desarrollar modelos eficaces para la identificación de palabras complejas [32].

En este estudio, se usó el conjunto de datos correspondiente al idioma español. La Tabla 1 presenta el número de instancias para cada idioma en los conjuntos de entrenamiento, desarrollo y prueba del *CWI Shared Task 2018*.

Tabla 1

Número de instancias del conjunto de datos del CWI Shared Task 2018. Adaptada de [32]

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

Tabla 2

Campos del conjunto de datos CWI Shared Task 2018. Adaptada de [32]

Columna	Descripción
HIT ID	Oraciones con el mismo *ID* pertenecen al mismo *HIT*
Oración	Oración original con palabra o frase objetivo
Inicio de Palabra	Índice de inicio de la palabra objetivo
Fin de Palabra	Índice final de la palabra objetivo
Palabra Objetivo	Palabra o frase 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)

La Tabla 2 presenta los campos y sus respectivas descripciones del conjunto de datos del CWI Shared Task 2018.

2.5. Definición de Variables

2.5.1. Machine Learning

De acuerdo con [33], machine learning es el campo de estudio que otorga a las computadoras la habilidad de aprender sin ser explícitamente programadas. En [34], se menciona que un modelo de machine learning es una función entrenada con datos para hacer predicciones o decisiones sobre datos nuevos y se diseñan para abordar diferentes tipos de problemas: regresión (predicción de valores continuos), clasificación (categorización en clases discretas) o ambos.

2.5.2. Embeddings de Nueva Generación

En [35], se indica que se ha lanzado una nueva generación de embeddings, que son representaciones numéricas de conceptos en textos. Los nuevos modelos, text-embedding-3-small y text-embedding-3-large, superan a sus predecesores en rendimiento. Ambos modelos cuentan con una técnica que permite ajustar el equilibrio entre el rendimiento y el costo, acortando los embeddings sin afectar su capacidad de representación.

2.5.3. Corpus del Español del Siglo XXI

Según menciona la Real Academia Española en [20], el Corpus del Español del Siglo XXI (CORPES XXI), desarrollado por esta, es una extensa base de datos lingüística que contiene textos escritos y orales del español actual de varios países, sirviendo como herramienta fundamental para investigaciones lingüísticas. En su versión 1.1 de abril de 2024, el corpus alcanzó más de 410 millones de formas, un aumento notable desde los 395 millones de la versión 1.0 de mayo 2023. La novedad más destacada es la incorporación de un diccionario de frecuencias léxicas que cuenta con 116707 lemas y 387429 formas [36].

2.5.4. Corpus de Referencia del Español Actual

De acuerdo con [37], el Corpus de Referencia del Español Actual (CREA) es un conjunto de textos digitales que representa el uso contemporáneo del español. Cuenta con varios listados de frecuencias, el más amplio de estos actualmente contiene 737779 instancias con su frecuencia absoluta y normalizada.

2.5.5. Wordfreq

En [38], se desarrolló Wordfreq, una biblioteca de Python que permite consultar la frecuencia de uso de palabras en más de 40 idiomas, utilizando diversas fuentes de datos y una escala logarítmica Zipf.

2.5.6. Spanish Billion Word Corpus and Embeddings

El Spanish Billion Word Corpus and Embeddings (SBW) es un recurso lingüístico que consiste en un corpus no anotado de aproximadamente 1.5 mil millones de palabras en español, creado a partir de diversas fuentes en la web e incluye vectores de palabras generados con el algoritmo word2vec utilizando el modelo skip-gram, evaluados mediante pruebas de relaciones de palabras [39].

2.6. Métodos de Procesamiento y Análisis

Para el desarrollo se usó Python 3. Además, se utilizó la semilla aleatoria de valor 42 en todo el proceso para mejorar su reproducibilidad. Dicho proceso incluyó la ingeniería de características en la cual, al inicio, se propuso, se procesó y se obtuvieron los valores numéricos correspondientes a las características mencionadas a continuación.

Para la obtención de los embeddings de palabras objetivos y de oraciones que las contienen, se usó el modelo text-embedding-3-large [35] especificando múltiples dimensiones. Para este estudio, dichas consultas fueron realizadas en septiembre del 2024.

Para la obtención de la frecuencia en CORPES XXI versión 1.1, se utilizó su diccionario de frecuencias [40]. Se realizó la identificación de la etiqueta de parte del discurso (POS) de la palabra en la oración que la contiene, la cual se utiliza para obtener una categoría gramatical a partir de un mapa que asocia cada POS con una categoría dentro de dicho diccionario. Con la palabra y la categoría obtenida, se busca su frecuencia en el diccionario; si no se encuentra con la categoría específica, se intenta con una genérica. En el caso de frases, se toma la frecuencia más baja de las palabras que las componen.

Para la obtención de la frecuencia en CREA, se utilizó su lista de frecuencias [41]. Se manejaron palabras individuales y frases; para ello, se aplicaron técnicas de normalización, como la eliminación de tildes y la lematización, para mejorar la coincidencia de palabras.

En el caso de frases, se utilizó la frecuencia más baja de sus palabras.

Para la obtención de la frecuencia Wordfreq, se usó dicha biblioteca en su versión 3.1.1 [42]. En el caso de frase objetivo, se manejó con la frecuencia más baja de sus palabras.

Para procesar la frecuencia del Spanish Billion Word Corpus and Embeddings, se usó el modelo preentrenado SBW-vectors-300-min5.bin [43] en español. Se utilizó el atributo o el método 'count' para acceder a las frecuencias del vocabulario de acuerdo con las posibles variaciones en la estructura de dicho modelo, y se manejaron casos en los que las palabras o frases no están en su vocabulario.

En las propiedades léxicas de palabras, se calculó la longitud mediante el número de caracteres que componen la palabra o frase objetivo. Para el número de palabras, se contabiliza la cantidad de estas en la frase objetivo, o se considera 1 si es una palabra individual. La presencia de mayúsculas se representa de forma binaria, donde 1 indica que la palabra contiene al menos una letra mayúscula, y 0, que está completamente en minúsculas.

En la selección de la combinación óptima de características para el modelo se utilizó XGBClassifier, ya que, de acuerdo con [44], XGBoost supera a modelos tradicionales, ofreciendo mejor interpretación de interacciones complejas y reducción de redundancia.

Siguiendo parte del método de evaluación realizado por [45], la Tabla 3 presenta los resultados de algunas combinaciones de características evaluadas mediante validación cruzada de 5 pliegues con el algoritmo XGBClassifier en el conjunto de entrenamiento. Los resultados muestran que los embeddings de 500 dimensiones de la palabra objetivo logran el mayor Macro F1 entre todas las evaluadas. Sin embargo, la inclusión de embeddings de la oración que la contiene disminuye el rendimiento del modelo, por lo que fue descartada. Se decidió agregar las características basadas en propiedades léxicas antes mencionadas con la finalidad de incrementar el rendimiento del modelo (longitud, número de palabras y un indicador binario si contiene mayúsculas la palabra o frase objetivo). La combinación que ofrece el mejor rendimiento en la selección de características, y por tanto la seleccionada, consiste en: embeddings de 500 dimensiones, frecuencias basadas en recursos lingüísticos (CORPES XXI, CREA, Wordfreq y Spanish Billion Word Corpus and Embeddings), y propiedades léxicas.

Para la selección de un posible mejor algoritmo para la tarea y basado en los experimentos de [46], 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

Tabla 3

Resultados de validación cruzada 5 pliegues de combinaciones de características en el conjunto de entrenamiento

Categoría	Características	Accuracy	Precision	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**

Tabla 4

Resultados de validación cruzada 5 pliegues de algoritmos y características seleccionadas en conjunto de entrenamiento

Algoritmo	Accuracy	Precision	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**

LGBMClassifier, utilizando sus hiperparámetros por defecto y las características seleccionadas anteriormente. Tras esta evaluación, el LGBMClassifier obtuvo el mejor Macro F1. En la Tabla 4 se muestran dichos resultados.

3. Resultados

Se desarrolló un modelo de machine learning que integra embeddings de nueva generación, frecuencias de palabras en recursos lingüísticos del español actual y propiedades léxicas para la identificación de palabras complejas en este idioma.

En una etapa final, el algoritmo que se seleccionó fue el LGBMClassifier, y fue entrenado en el conjunto de entrenamiento con la combinación óptima de características anteriormente encontrada, y luego fue evaluado en el conjunto de prueba utilizando varias métricas de rendimiento.

Como puede apreciarse en la Tabla 5, el modelo dio lugar a un Accuracy del 0.8115, lo que explica un buen rendimiento global sobre la tarea tratada. A su vez, también se obtuvieron un Precision de 0.8119 y un Recall del 0.7932, los cuales sugieren un buen equilibrio entre la capacidad del modelo para identificar correctamente las palabras complejas y su capacidad para no dejar de lado a estas. La métrica de Macro F1, que busca un cierto equilibrio entre el Precision y el Recall, dado que este considera todas las clases como igualmente importantes, fue de 0.7993, indicando un rendimiento bastante constante del modelo en términos generales.

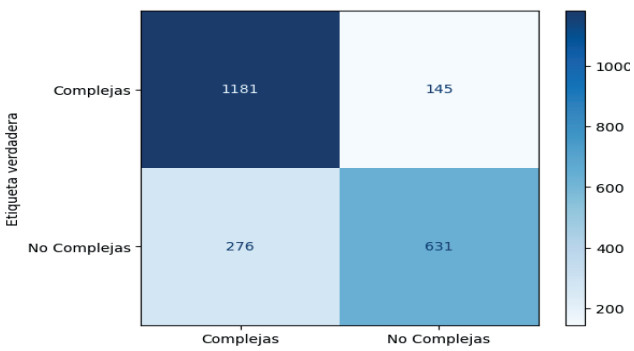
Tabla 5

Resultados de la evaluación en el conjunto de prueba

Métrica	Resultado
Accuracy	0.8115
Precision	0.8119
Recall	0.7932
Macro F1	**0.7993**

Figura 2

Matriz de confusión del modelo desarrollado



Para ilustrar mejor el rendimiento del modelo desarrollado en este estudio, en la Figura 2 se muestra la matriz de confusión de la evaluación del modelo.

La matriz de confusión muestra el desempeño del modelo, indicando que ha clasificado correctamente 1181 palabras complejas (verdaderos positivos) y 631 no complejas (verdaderos negativos). Por el contrario, el modelo ha clasificado erróneamente un número de 276 palabras no complejas como complejas (falsos positivos) y no ha reconocido una cifra de 145 palabras complejas (falsos negativos). Esto indica un buen rendimiento general en la identificación de palabras complejas.

Para evaluar la importancia de las diferentes características en el modelo de identificación de palabras complejas en español, se realizó un análisis de importancia de características. Este análisis nos permite comprender qué características tienen mayor influencia en las predicciones del modelo.

En la Figura 3 se presenta un gráfico de barras que ilustra la importancia de cada característica en el modelo propuesto (para los embeddings se han agregado las importancias de cada una de las 500 dimensiones en un solo valor). Estos resultados resaltan la importancia de los embeddings text-embedding-3-large en este modelo de identificación de palabras complejas en español. Su dominancia en el gráfico sugiere que esta característica captura información semántica muy importante para la tarea de identificación de palabras complejas. Las demás características seleccionadas, como las frecuencias basadas en recursos lingüísticos y las propiedades léxicas también son relevantes, pero en un grado menor en comparación con dichos embeddings.

4. Discusión

El modelo que se desarrolló en este estudio, basado en el conjunto de datos en español de la tarea compartida de Identificación de Palabras Complejas (CWI 2018) ha demostrado un buen rendimiento en esta tarea, superando al rendimiento obtenido por varios modelos recientes. En la Tabla 6 se muestran los Macro F1 de modelos recientes, en la que este estudio alcanza el mejor desempeño.

Esta evaluación confirma la importancia de incorporar embeddings de nueva generación (como text-embedding-3-large) en modelos de procesamiento de lenguaje natural, especialmente para tareas desafiantes como la identificación de palabras complejas, ya que capturan información importante relacionada con la complejidad de estas, que se puede complementar con sus frecuencias basadas en recursos lingüísticos

Figura 3

Importancia de las características en el modelo desarrollado

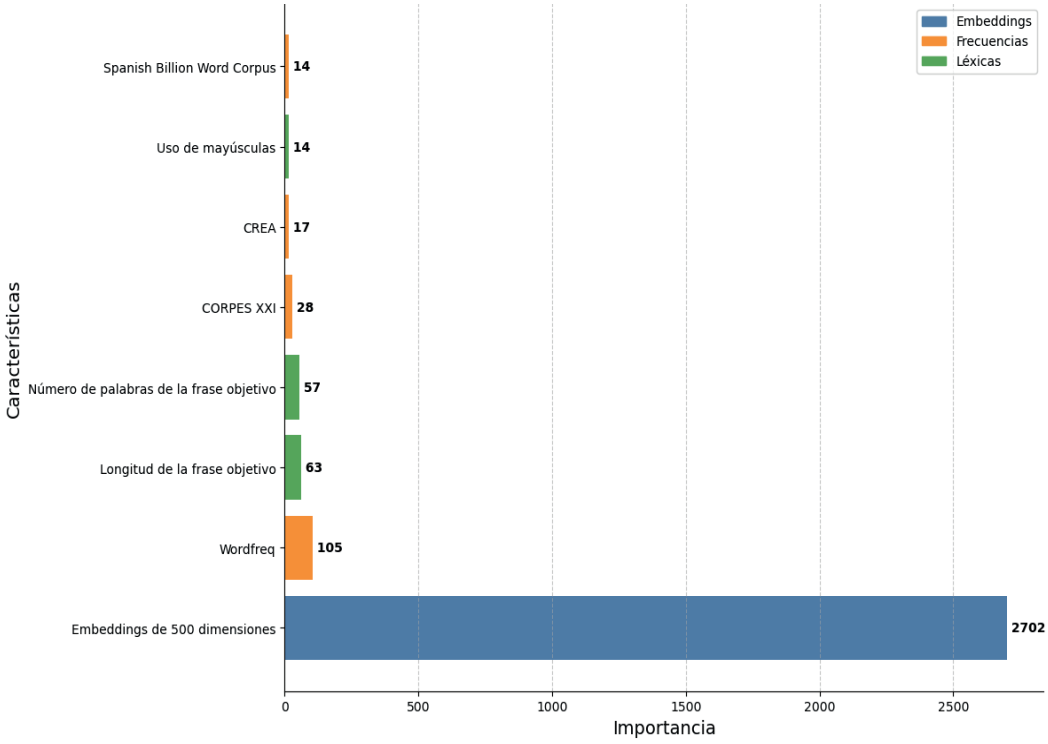


Tabla 6

Macro F1 de modelos para la tarea binaria de identificación de palabras complejas entrenados y evaluados recientemente con el conjunto de datos CWI 2018 en español

Estudio	Modelo/Enfoque	Macro F1
Este estudio	*LGBMClassifier* con *embeddings* de nueva generación, frecuencias en recursos lingüísticos del español actual y propiedades lingüísticas	0.7993
[47]	Redes neuronales convolucionales con características lingüísticas y morfológicas	0.7970
[45]	*SVM* con *kernel* lineal y características como la longitud de palabras, *booleano* para indicar si contiene mayúsculas, *embeddings Word2vec*, *embeddings* de *BERT*, y un *booleano* para la existencia de la palabra en el diccionario *E2*R	0.7920
[12]	Evaluación que empleó dos familias de modelos de lenguaje grandes (*LLM*): *Llama-2* y *ChatGPT-3.5*, en la cual *ChatGPT-3.5-turbo-ft* obtuvo el mejor puntaje	0.7810
[48]	Regresión logística con 25 características	0.7760
[49]	Regresión logística con *embeddings* de *BERT* y otras características	0.7700
[50]	*Random Forest* con frecuencias basadas en corpus de aprendizaje. Logró el mejor puntaje de los equipos que participaron en la tarea compartida de clasificación binaria en español *CWI 2018*.	0.7699
[51]	*SVM* con *kernel* de función de base radial y características de longitud, frecuencia, formato, inclusión en *E2R* y *embeddings Word2Vec* y *FastText*	0.7497

del español actual (CORPES XXI, CREA, Wordfreq, Spanish Billion Word Corpus and Embeddings) y características basadas en propiedades léxicas (como la longitud, el número de palabras o indicadores de mayúsculas) para mejorar la precisión en esta tarea, proporcionando una base sólida para el rendimiento del modelo con un número relativamente pequeño de características.

El estudio muestra un potencial significativo para mejorar herramientas de simplificación de textos y sistemas de apoyo a la lectura, y demuestra que es posible seguir mejorando la precisión de esos modelos mediante la capacidad representativa de nuevos embeddings y de los recursos lingüísticos que se van mejorando y actualizando constantemente.

5. Conclusiones

Este estudio contribuye al campo de la identificación de palabras complejas en español, en cuanto a investigaciones en este idioma. Demuestra cómo el uso de embeddings de nueva generación, frecuencias basadas en recursos léxicos del español actual y algunas propiedades léxicas permiten construir modelos de *CWI* en español más precisos.

El modelo desarrollado en este estudio, basado en el algoritmo LGBMClassifier y las características mencionadas anteriormente, obtuvo un Macro F1 de 0.7993, lo cual superó el rendimiento de algunos modelos anteriores que emplearon diversas arquitecturas. Su buen desempeño, en comparación con algunos basados en redes neuronales y modelos de lenguaje grandes, citados anteriormente, indica que las características y el diseño de los modelos pueden competir e incluso superar, en esta tarea, a las arquitecturas más complejas.

El uso de estos recursos actuales sugiere que hay oportunidades para futuros estudios que mejoren la precisión en esta tarea, la exploración en otros idiomas, dialectos, dominios y su potencial integración en sistemas de simplificación léxica y herramientas de accesibilidad lingüística cada vez más precisas.

6. Referencias

[1] D. Khurana, A. Koli, K. Khatter, y S. Singh, "Natural language processing: state of the art, current trends and challenges", Multimed Tools Appl, vol. 82, núm. 3, pp. 3713–3744, ene. 2023, doi: 10.1007/s11042-022-13428-4.

[2] G. H. Paetzold y L. Specia, "A Survey on Lexical Simplification", Journal of Artificial Intelligence Research, vol. 60, pp. 549–593, 2017.

[3] National Geographic España, "Cuál es el futuro del español, la segunda lengua más hablada del mundo". Consultado: el 21 de junio de 2024. [En línea]. Disponible en: https://www.nationalgeographic.com.es/mundo-ng/cual-es-futuro-espanol-segunda-lengua-mas-hablada-mundo_22113

[4] E. Rennes, M. Santini, y A. Jonsson, "The Swedish Simplification Toolkit: Designed with Target Audiences in Mind", en Proceedings of the 2nd READI Workshop @ LREC2022, European Language Resources Association (ELRA), 2022, pp. 31–38. [En línea]. Disponible en: https://aclanthology.org/2022.readi-1.5.pdf

[5] S. Aluísio y C. Gasperin, "Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplification of Portuguese Texts", en Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, T. Solorio y T. Pedersen, Eds., Los Angeles, California: Association for Computational Linguistics, jun. 2010, pp. 46–53. [En línea]. Disponible en: https://aclanthology.org/W10-1607

[6] S. Gooding y E. Kochmar, "CAMB at CWI Shared Task 2018: Complex Word Identification with Ensemble-Based Voting", en Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, J. Tetreault, J. Burstein, E. Kochmar, C. Leacock, y H. Yannakoudakis, Eds., New Orleans, Louisiana: Association for Computational Linguistics, jun. 2018, pp. 184–194. doi: 10.18653/v1/W18-0520.

[7] E. Loginova y D. Benoit, "Structural information in mathematical formulas for exercise difficulty prediction: a comparison of NLP representations", en Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, y T. Zesch, Eds., Seattle, Washington: Association for Computational Linguistics, jul. 2022, pp. 101–106. doi: 10.18653/v1/2022.bea-1.14.

[8] D. Reinsel, J. Gantz, y J. Rydning, "The Digitization of the World: From Edge to Core", nov. 2018.

[9] G. H. McLaughlin, "SMOG grading: A new readability formula.", Journal of Reading, vol. 12, núm. 8, pp. 639–646, 1969.

[10] E. Dale y J. S. Chall, "A Formula for Predicting Readability", Educational Research Bulletin, vol. 27, núm. 1, pp. 11–28, 1948, [En línea]. Disponible en: http://www.jstor.org/stable/1473169

[11] J. A. Ortiz-Zambrano, C. Espin-Riofrio, y A. Montejo-Ráez, "Transformers for Lexical Complexity Prediction in Spanish Language[Transformers para la Predicción de la Complejidad Léxica en Lengua Española]", Procesamiento del Lenguaje Natural, vol. 69, pp. 177–188, sep. 2022, doi: 10.26342/2022-69-15.

[12] Anonymous, "Investigating Large Language Models for Complex Word Identification in Multilingual and Multidomain Setups", en Submitted to ACL Rolling Review - April 2024, 2024. [En línea]. Disponible en: https://openreview.net/forum?id=GnGXbekH7M

[13] M. Shardlow, K. North, y M. Zampieri, "Multilingual Resources for Lexical Complexity Prediction: A Review", en Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, Torino, Italia: ELRA Language Resource Association, may 2024, pp. 51–59. [En línea]. Disponible en: https://aclanthology.org/2024.determit-1.5

[14] R. Alarcón, L. Moreno, y P. Martínez, "Exploration of Spanish Word Embeddings for Lexical Simplification", en Proceedings of the First Workshop on Current Trends in Text Simplification (CTTS 2021), Online: SEPLN, sep. 2021.

[15] S. Bott, H. Saggion, N. P. Rojas, M. S. Salazar, y S. C. Ramirez, "MultiLS-SP/CA: Lexical Complexity Prediction and Lexical Simplification Resources for Catalan and Spanish", abr. 2024, Consultado: el 21 de septiembre de 2024. [En línea]. Disponible en: https://arxiv.org/abs/2404.07814v1

[16] J. Degraeuwe y P. Goethals, "LexComSpaL2: A Lexical Complexity Corpus for Spanish as a Foreign Language", en Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, y N. Xue, Eds., Torino, Italia: ELRA and ICCL, may 2024, pp. 10432–10447. [En línea]. Disponible en: https://aclanthology.org/2024.lrec-main.912

[17] J. A. Ortiz-Zambrano, C. Espin-Riofrio, y A. Montejo-Ráez, "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, núm. 71, pp. 247 – 259, 2023, doi: 10.26342/2023-71-19.

[18] R. Almeida, H. Tissot, y M. D. Del Fabro, "C3SL at SemEval-2021 Task 1: Predicting Lexical Complexity of Words in Specific Contexts with Sentence Embeddings", en Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, y X. Zhu, Eds., Online:

Association for Computational Linguistics, ago. 2021, pp. 683–687. doi: 10.18653/v1/2021.semeval-1.88.

[19] R. Stodden y G. Venugopal, “RS_GV at SemEval-2021 Task 1: Sense Relative Lexical Complexity Prediction”, en Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, y X. Zhu, Eds., Online: Association for Computational Linguistics, ago. 2021, pp. 640–649. doi: 10.18653/v1/2021.semeval-1.82.

[20] RAE, “CORPES XXI | Real Academia Española”. Consultado: el 7 de julio de 2024. [En línea]. Disponible en: https://www.rae.es/banco-de-datos/corpes-xxi

[21] D. Alfter, “Complexity and Indecision: A Proof-of-Concept Exploration of Lexical Complexity and Lexical Semantic Change”, en Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change, Online: Association for Computational Linguistics, ago. 2024, pp. 137–143. doi: 10.18653/v1/2024.lchange-1.14.

[22] E. Rozi et al., “Stanford MLab at SemEval-2021 Task 1: Tree-Based Modelling of Lexical Complexity using Word Embeddings”, en Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, y X. Zhu, Eds., Online: Association for Computational Linguistics, ago. 2021, pp. 688–693. doi: 10.18653/v1/2021.semeval-1.89.

[23] M. D. Jiménez-López y A. Torrens Urrutia, “Sobre el concepto de complejidad en Lingüística- The concept of complexity in Linguistics”, sep. 2018.

[24] M. Miestamo, “Grammatical complexity in cross-linguistic perspective”, Language Complexity: Typology, Contact, Change, pp. 23–42, sep. 2008, doi: 10.1075/slcs.94.04mie.

[25] G. Pallotti, “A simple view of linguistic complexity”, Second Lang Res, vol. 31, núm. 1, pp. 117–134, 2015, doi: 10.1177/0267658314536435.

[26] K. North, M. Zampieri, y M. Shardlow, “Lexical Complexity Prediction: An Overview”, ACM Comput Surv, vol. 55, núm. 9, p. 40, mar. 2023, doi: 10.1145/3557885.

[27] J. Goldschneider y R. DeKeyser, “Explaining the ‘Natural Order of L2 Morpheme Acquisition’ in English: A Meta-analysis of Multiple Determinants”, Lang Learn, vol. 51, pp. 1–50, sep. 2001, doi: 10.1111/1467-9922.00147.

[28] G. Paetzold y L. Specia, “SemEval 2016 Task 11: Complex Word Identification”, en Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, y T. Zesch, Eds., San Diego, California: Association for Computational Linguistics, jun. 2016, pp. 560–569. doi: 10.18653/v1/S16-1085.

[29] M. N. K. Saunders, A. Thornhill, y P. Lewis, Research Methods for Business Students, 8th ed. Pearson, 2019.

[30] RAE, “Presentación del «Diccionario de la lengua española» y sus ediciones | Real Academia Española”. Consultado: el 1 de agosto de 2024. [En línea]. Disponible en: https://www.rae.es/obras-academicas/diccionarios/presentacion-del-diccionario-de-la-lengua-espanola-y-sus-ediciones

[31] R. Hernández Sampieri, C. Fernández Collado, y P. Baptista Lucio, Metodología de la investigación, 6a ed. México D.F.: McGraw-Hill, 2014.

[32] “CWI Shared Task 2018 - Datasets”. Consultado: el 1 de agosto de 2024. [En línea]. Disponible en: https://sites.google.com/view/cwisharedtask2018/datasets

[33] A. L. Samuel, “Some Studies in Machine Learning Using the Game of Checkers”, IBM J Res Dev, vol. 3, núm. 3, pp. 210–229, 1959, doi: 10.1147/rd.33.0210.

[34] M. J. Islam, Machine Learning Model Serving Patterns and Best Practices: A Definitive Guide to Deploying, Monitoring, and Providing Accessibility to ML Models in Production. Packt Publishing, 2022.

[35] OPENAI, “New embedding models and API updates | OpenAI”. Consultado: el 8 de agosto de 2024. [En línea]. Disponible en: https://openai.com/index/new-embedding-models-and-api-updates/

[36] RAE, “CORPES”. Consultado: el 2 de septiembre de 2024. [En línea]. Disponible en: https://www.rae.es/corpes/

[37] RAE, “CREA | Real Academia Española”. Consultado: el 8 de septiembre de 2024. [En línea]. Disponible en: https://www.rae.es/banco-de-datos/crea

[38] R. Speer, “rspeer/wordfreq: v3.0”, septiembre de 2022, Zenodo. doi: 10.5281/zenodo.7199437.

[39] C. Cardellino, “Spanish Billion Words Corpus and Embeddings”, marzo de 2016.

[40] Real Academia Española, “Diccionario de frecuencias CORPES XXI”, 2024. [En línea]. Disponible en: https://www.rae.es/corpes/assets/rae/files/corpes/diccionario_frecuencias_corpes_alfa.tsv

[41] Real Academia Española, “Lista total de frecuencias”, 2024. [En línea]. Disponible en: https://corpus.rae.es/frec/CREA_total.zip

[42] R. Speer, "wordfreq: Word frequencies for many languages", 2023, Python Package Index. [En línea]. Disponible en: https://files.pythonhosted.org/packages/4c/cd/9581ff0ea2c581012d0caae4bba024f3ff6b46e030a55dde-c1ce545e2caf/wordfreq-3.1.1.tar.gz

[43] C. Cardellino, “Spanish Billion Words Corpus and Embeddings (SBWCE)”, 2016. [En línea]. Disponible en: https://cs.famaf.unc.edu.ar/~ccardellino/SBWCE/SBW-vectors-300-min5.bin.gz

[44] A. Alsahaf, N. Petkov, V. Shenoy, y G. Azzopardi, “A framework for feature selection through boosting”, Expert Syst Appl, vol. 187, p. 115895, 2022, doi: https://doi.org/10.1016/j.eswa.2021.115895.

[45] R. Alarcon, L. Moreno, y P. Martínez, “Lexical Simplification System to Improve Web Accessibility”, IEEE Access, vol. 9, pp. 58755–58767, 2021, doi: 10.1109/ACCESS.2021.3072697.

[46] S. Gooding y E. Kochmar, “CAMB at CWI Shared Task 2018: Complex Word Identification with Ensemble-Based Voting”, en Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, J. Tetreault, J. Burstein, E. Kochmar, C. Leacock, y H. Yannakoudakis,

Eds., New Orleans, Louisiana: Association for Computational Linguistics, jun. 2018, pp. 184–194. doi: 10.18653/v1/W18-0520.

[47] K. C. Sheang, “Multilingual Complex Word Identification: Convolutional Neural Networks with Morphological and Linguistic Features”, en Proceedings of the Student Research Workshop Associated with RANLP 2019, V. Kovatchev, I. Temnikova, B. Šandrih, y I. Nikolova, Eds., Varna, Bulgaria: INCOMA Ltd., sep. 2019, pp. 83–89. doi: 10.26615/issn.2603-2821.2019_013.

[48] P. Finnimore et al., “Strong Baselines for Complex Word Identification across Multiple Languages”, en 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), J. Burstein, C. Doran, y T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, jun. 2019, pp. 970–977. doi: 10.18653/v1/N19-1102.

[49] J. Pimienta Castillo, “Multilingual lexical simplification”, 2021, Consultado: el 8 de septiembre de 2024. [En línea]. Disponible en: http://repositori.upf.edu/handle/10230/49224

[50] T. Kajiwara y M. Komachi, “Complex Word Identification Based on Frequency in a Learner Corpus”, en *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, J. Tetreault, J. Burstein, E. Kochmar, C. Leacock, y H. Yannakoudakis, Eds., New Orleans, Louisiana: Association for Computational Linguistics, jun. 2018, pp. 195–199. doi: 10.18653/v1/W18-0521./9581ff0ea2c581012d0caae4bba024f3ff6b46e030a55ddec1ce545e2caf/wordfreq-3.1.1.tar.gz

[51] Alarcon, L. Moreno, I. Segura-Bedmar, y P. Martínez, “Lexical simplification approach using easy-to-read resources”, Procesamiento del Lenguaje Natural, vol. 63, pp. 95–102, sep. 2019, doi: 10.26342/2019-63