UNIVERSIDAD DE GUAYAQUIL

FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

EVALUACIÓN DE DESEMPEÑO DE LOS MODELOS TRANSFORMADORES PARA LA PREDICCION DE LA COMPLEJIDAD LÉXICA PARA EL IDIOMA ESPAÑOL

PROYECTO DE TITULACIÓN

Previa a la obtención del Título de:

INGENIERO EN SISTEMAS COMPUTACIONALES

AUTORES: Diana Geovanna Aroca Pincay

Diego Gabriel Bernal Yucailla

TUTORA: M.Sc. Jenny Alexandra Ortiz Zambrano

GUAYAQUIL – ECUADOR

2022

REPOSITORIO NACIONAL EN CIENCIAS Y TECNOLOGÍAS

FICHA DE REGISTRO DE TRABAJO DE TITULACIÓN

TÍTULO: *"Evaluación de Desempeño de los modelos transformadores para la predicción de la complejidad léxica para el idioma español"*

AUTOR(ES): Diana Geovanna Aroca Pincay Diego Gabriel Bernal Yucailla	**REVISOR(A):** Ing. Alcides Reyes Guerra
INSTITUCIÓN: Universidad de Guayaquil	**FACULTAD: Ciencias Matemáticas y Físicas**
CARRERA: Ingeniería en Sistemas Computacionales	
FECHA DE PUBLICACIÓN:	**N° DE PAGS:** 133
AREA TEMÁTICA: Procesamiento del lenguaje natural	
PALABRAS CLAVES: Transformers, Datasets, Fine-tuning, Machine Learning, Algoritmos, Prediccion de la complejidad Léxica	

RESUMEN: La presente investigacion plantea la exploracion y evaluacion de los diferentes modelos de *Transformers* aplicados para el idioma español las cuales son BERT, XML-RoBERTa y RoBERTa-Large-BNE con el objetivo de evaluar el nivel de prediccion de las palabras complejas de los textos en español. Estos modelos pre-entrenados serán ejecutados con un corpus previamente creado de los textos Universitarios en español- ClexIS^2 aplicando el algoritmo pre-entrenado generico del codigo abierto de cada *Transformers.* Esto permitiendo la generacion de embedding(incrustaciones-encodings) para la creacion de los Datasets que previamente seran enntrenados por los diferentes algoritmos de Machine Learning, obteniendo la prediccion de la complejidad Léxica(LCP).Ya realizado la exploracion de los Transformers se procedera a la realizacion del ajuste fino a cada modelo mediante la ejecucion de la tecnica de Fine-Tuning sobre los modelos pre-entrenados para la generacion de los Datasets basados en las nuevas representaciones numericas, que a su vez tambien procederan ser entrenados y evaluados por los diferentes algoritmos para obtener la prediccion de la complejidad Léxica.Ya obtenido los resultados de las diferentes combinaciones de los modelos pre-entrenados y ajustados de los modelos *Transformers* se evaluara su desempeño para determinar los mejores resultados de la prediccion de la complejidad Lexica del idioma español

Palabras clave: Transformers, Datasets, Fine-tuning, Machine Learning, Algoritmos, Prediccion de la complejidad Léxica

N° DE REGISTRO:	**N° DE CLASIFICACIÓN:**
DIRECCIÓN URL: (PROYECTO DE TITULACION EN LA WEB)	

ADJUNTO PDF	**SI** [x]	**NO** [ ]
CONTACTO CON AUTORES: Diana Geovanna Aroca Pincay Diego Gabriel Bernal Yucailla	**Teléfono:** 0980538332 0994918537	**Email:** **Diana.arocap@ug.edu.ec,** **DiianaAroca96@hotmail.com ;** **Diego.bernaly@ug.edu.ec**
CONTACTO DE LA INSTITUCIÓN	**Nombre:** Ab. Juan Chávez Atocha	
	Teléfono: 2307729	
	Email: juan.chaveza@ug.edu.ec	

APROBACIÓN DEL TUTOR

En mi calidad de Tutora del Trabajo de Titulación, "Evaluación de Desempeño de los modelos transformadores para la predicción léxica para el idioma español" elaborado por los Sres. **Aroca Pincay Diana Geovanna** y **Bernal Yucailla Diego Gabriel** , **estudiantes no titulados** de la Carrera de Ingeniería en Sistemas Computacionales, Facultad de Ciencias Matemáticas y Físicas de la Universidad de Guayaquil, previo a la obtención del Título de Ingeniero(a) en Sistemas Computacionales, me permito declarar que luego de haber orientado, estudiado y revisado, la **apruebo** en todas sus partes.

Atentamente,

M.Sc. Jenny Alexandra Ortiz Zambrano

TUTORA

DEDICATORIA

Dedico este trabajo de titulación, a Dios, a mis padres: Rosario Pincay - Geovanny Aroca, a mi tía Matilde Pincay, a mi familia, a mi enamorado, que siempre estuvieron apoyándome en cada proceso durante toda mi carrera.

Diana Geovanna Aroca Pincay

Dedico este trabajo de titulación a Dios, a mi madre América Yucailla, a mi familia, a mis amigos por el apoyo, confianza y motivación brindad en el transcurso de mi carrera académica.

Diego Gabriel Bernal Yucailla

AGRADECIMIENTO

Agradezco a Dios por darme las fuerzas que tanto necesitaba para seguir adelante, a mis padres por sus apoyos incondicional, ellos mas que todo se merecen todo esto por estar ahí en cada momento y transcurso de toda mi carrera académica, a mi familia, a mis hermanas porque por ellas no me he dado por vencida para que sigan mi ejemplo y no den marcha atrás, a mi enamorado que me ha ayudado y apoyado en todo en el transcurso de la carrera.

También agradezco a mi compañero de tesis, Diego Bernal por la experiencia y conocimientos compartidos durante este proceso, a M.Sc Jenny Ortiz Zambrano que con su experiencia, conocimiento y dedicación guío el desarrollo de la tesis.

Diana Geovanna Aroca Pincay

AGRADECIMIENTO

Agradezco a Dios, por la fuerza brindada en el transcurso de mi carrera, Agradezco a mi Madre America Yucailla, a mi padre Victor Bernal, a mi Tía Martha Yumiseba y a toda mi familia que me han aportado en mi crecimiento profesional con fuerzas y alientos en este largo proceso. También agradezco profundamente a mis amigos encontrados en este largo camino que aportaron en mi crecimiento como persona y como profesional.

También agradezco a mi compañera de tesis, Diana Aroca por el esfuerzo y dedicación durante este proceso, a M.Sc Jenny Ortiz Zambrano que con su experiencia, conocimiento y dedicación guío el desarrollo de la tesis.

Diego Gabriel Bernal Yucailla

TRIBUNAL PROYECTO DE TITULACIÓN

Ing. Douglas Iturburu Salvador, M.Sc.
DECANO DE LA FACULTAD
CIENCIAS MATEMÁTICAS Y FÍSICAS

Ing. Lorenzo Cevallos Torres, Mgs.
DIRECTOR DE LA CARRERA DE
INGENIERÍA EN SISTEMAS
COMPUTACIONALES

M.Sc. Jenny Alexandra Ortiz Zambrano
PROFESORA TUTORA DEL PROYECTO
DE TITULACIÓN

Ing. Alcides Reyes Guerra
PROFESOR REVISOR DEL PROYECTO
DE TITULACIÓN

Ab. Juan Chávez Atocha, Esp.
SECRETARIO

DECLARACIÓN EXPRESA

"La responsabilidad del contenido de este Proyecto de Titulación, me corresponden exclusivamente; y el patrimonio intelectual de la misma a la UNIVERSIDAD DE GUAYAQUIL".

Diana Geovanna Aroca Pincay

Diego Gabriel Bernal Yucailla

CESIÓN DE DERECHOS DE AUTOR

Ingeniero

Douglas Iturburu Salvador, M.Sc.

DECANO DE LA FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS

Presente.

A través de este medio indico a usted que procedo a realizar la entrega de la cesión de derechos de autor en forma libre y voluntaria del trabajo de titulación ***"Evaluación de Desempeño de los modelos transformadores para la predicción de la complejidad léxica para el idioma español",*** realizado como requisito previo para la obtención del Título de Ingeniero(a) en Sistemas Computacionales de la Universidad de Guayaquil.

Guayaquil, 18 de Marzo del 2022.

Diana Geovanna Aroca Pincay
C.I. N° 0954536041

Diego Gabriel Bernal Yucailla
C.I. N° 0952491025

UNIVERSIDAD DE GUAYAQUIL

FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

EVALUACIÓN DE DESEMPEÑO DE LOS MODELOS TRANSFORMADORES PARA LA PREDICCION DE LA COMPLEJIDAD LÉXICA PARA EL IDIOMA ESPAÑOL

Proyecto de Titulación que se presenta como requisito para optar por el título de

INGENIERO(A) EN SISTEMAS COMPUTACIONALES

Autores: Diana Geovanna Aroca Pincay

C.I. N° 0954536041

Diego Gabriel Bernal Yucailla

C.I. N° 0952491025

Tutora: M.Sc. Jenny Alexandra Ortiz Zambrano

Guayaquil, 18 de marzo del 2022

CERTIFICADO DE ACEPTACIÓN DEL TUTOR

En mi calidad de Tutora del Proyecto de Titulación, nombrado por el Consejo Directivo de la Facultad de Ciencias Matemáticas y Físicas de la Universidad de Guayaquil.

CERTIFICO:

Que he analizado el Proyecto de Titulación presentado por los estudiantes **Diana Geovanna Aroca Pincay, Diego Gabriel Bernal Yucailla**, como requisito previo para optar por el Título de Ingenieros en Sistemas Computacionales cuyo proyecto es:

Evaluación de Desempeño de los modelos transformadores para la predicción de la complejidad léxica para el idioma español

Considero aprobado el trabajo en su totalidad.

Presentado por:

Aroca Pincay Diana Geovanna	0954536041 Cédula de identidad N°
Bernal Yucailla Diego Gabriel	0952491025 Cédula de identidad N°

Tutora: M.Sc. Jenny Alexandra Ortiz Zambrano

Firma

Guayaquil, 18 de Marzo del 2022

UNIVERSIDAD DE GUAYAQUIL
FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS
CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

AUTORIZACIÓN PARA PUBLICACIÓN DE PROYECTO DE TITULACIÓN EN FORMATO DIGITAL

1. Identificación del Proyecto de Titulación

Nombre del Estudiante: Diana Geovanna Aroca Pincay	
Dirección: Urb. Málaga 2 Mz V 5	
Teléfono: 0980538332	**Email:Diana.arocap@ug.edu.ec,**
Nombre del Estudiante: Diego Gabriel Bernal Yucailla	
Dirección: Victor Hugo Briones ·418 y Colón	
Teléfono:0994018537	**Email: Diego.bernaly@ug.edu.ec**
Facultad: Ciencias, Matemáticas y Física	
Carrera: Ingeniería en Sistemas Computacionales	
Proyecto de Titulación al que opta: Ingeniero en Sistemas Computacionales	
Profesor(a) Tutor(a): MSc. Jenny Alexandra Ortiz Zambrano	
Título del Proyecto de Titulación: Evaluación de Desempeño de los modelos Transformadores para la predicción de la complejidad léxica para el idioma español.	
Palabras Claves: Transformers, Datasets, Fine-tuning, Machine Learning, Algoritmos, Prediccion de la complejidad Léxica	

2. Autorización de Publicación de Versión Electrónica del Proyecto de Titulación

A través de este medio autorizo a la Biblioteca de la Universidad de Guayaquil y a la Facultad de Ciencias Matemáticas y Físicas a publicar la versión electrónica de este Proyecto de Titulación.

Publicación Electrónica:

Inmediata		Después de 1 año	

Firma Estudiante:

Aroca Pincay Diana Geovanna	0954536041 Cédula de identidad N°
Bernal Yucailla Diego Gabriel	0952491025 Cédula de identidad N°

3. Forma de envío:

El texto del Proyecto de Titulación debe ser enviado en formato Word, como archivo .docx, .RTF o .Puf para PC. Las imágenes que la acompañen pueden ser: .gif, .jpg o .TIFF.

DVDROM ☐ CDROM ☐

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

ABREVIATURAS

CC.MM.FF	Facultad de Ciencias Matemáticas y Físicas
NLTK	Lenguaje natural de caja de Herramienta
NLP	Procesamiento de Lenguaje natural
SVM	Maquina de Vectores de Soporte
XML	Lenguaje de Marcado Extensible
BNE	Biblioteca Nacional de España
Ing.	Ingeniero
GBR	Regresión de Aumento de Gradiente.
M.Sc.	Máster
ABR	Regresión de AdaBoost
UG	Universidad de Guayaquil
IA	Inteligencia Artificial
MAE	Error Absoluto Medio

SIMBOLOGÍA

s	Desviación estándar
e	Error
E	Espacio muestral
E(Y)	Esperanza matemática de la v.a. y
s	Estimador de la desviación estándar
e	Exponencial

UNIVERSIDAD DE GUAYAQUIL
FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS
CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

EVALUACIÓN DE DESEMPEÑO DE LOS MODELOS TRANSFORMADORES PARA LA PREDICCION DE LA COMPLEJIDAD LÉXICA PARA EL IDIOMA ESPAÑOL

Autores: Diana Geovanna Aroca Pincay
C.I. N° 0954536041
Diego Gabriel Bernal Yucailla
C.I. N° 0952491025
Tutora: M.sC. Jenny Alexandra Ortiz Zambrano

RESUMEN

La presente investigacion plantea la exploracion y evaluacion de los diferentes modelos de *Transformers* aplicados para el idioma español las cuales son BERT, XML-RoBERTa y RoBERTa-Large-BNE con el objetivo de evaluar el nivel de prediccion de las palabras complejas de los textos en español. Estos modelos pre-entrenados serán ejecutados con un corpus previamente creado de los textos Universitarios en español- ClexIS^2 aplicando el algoritmo pre-entrenado generico del codigo abierto de cada *Transformers*. Esto permitiendo la generacion de embedding(incrustaciones-encodings) para la creacion de los Datasets que previamente seran enntrenados por los diferentes algoritmos de Machine Learning, obteniendo la prediccion de la complejidad Léxica(LCP).Ya realizado la exploracion de los Transformers se procedera a la realizacion del ajuste fino a cada modelo mediante la ejecucion de la tecnica de Fine-Tuning sobre los modelos pre-entrenados para la generacion de los Datasets basados en las nuevas representaciones numericas, que a su vez tambien procederan ser entrenados y evaluados por los diferentes algoritmos para obtener la prediccion de la complejidad Léxica.Ya obtenido los resultados de las diferentes combinaciones de los modelos pre-entrenados y ajustados de los modelos *Transformers* se evaluara su desempeño para determinar los mejores resultados de la prediccion de la complejidad Lexica del idioma español

Palabras clave: Transformers, Datasets, Fine-tuning, Machine Learning, Algoritmos, Prediccion de la complejidad Léxica

UNIVERSIDAD DE GUAYAQUIL
FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS
CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

EVALUACIÓN DE DESEMPEÑO DE LOS MODELOS TRANSFORMADORES PARA LA PREDICCION DE LA COMPLEJIDAD LÉXICA PARA EL IDIOMA ESPAÑOL

Author(s): Diana Geovanna Aroca Pincay
C.I. N° 0954536041
Diego Gabriel Bernal Yucailla
C.I. N° 0952491025
Tutora: Ing. Jenny Alexandra Ortiz Zambrano

ABSTRACT

This research proposes the exploration and evaluation of the different Transformers models applied to the Spanish language, which are BERT, XML-RoBERTa and RoBERTa-Large-BNE with the aim of evaluating the level of prediction of complex words in texts in Spanish. Spanish. These pretrained models will be executed with a previously created corpus of University texts in Spanish - ClexIS2 applying the open source generic pretrained algorithm of each Transformers. This allowed the generation of embedding (embeddings-encodings) for the creation of the Datasets that were previously trained by the different Machine Learning algorithms, obtaining the Lexical Complexity Prediction (LCP). Once the exploration of the Transformers has been carried out, we will proceed to the realization of the fine adjustment to each model by means of the execution of the Fine-Tuning technique on the pre-trained models for the generation of the Datasets based on the new numerical representations, which in turn will also proceed to be alterations and evaluated by the different algorithms to obtain the prediction of the Lexical complexity. You have already obtained the results of the different combinations of the pre-trained and adjusted models of the Transformers models, their performance will be evaluated to determine the best results of the prediction of the Lexical complexity of the language Spanish

Key words: Transformers, Datasets, Fine-Tuning, Machine Learning, Algorithms, Lexical Complexity Prediction

INTRODUCCIÓN

La lectura como fuente principal de la compresión de los textos no solo abarca palabras o secciones que contiene un texto sino todo su contenido, el mismo que muchas veces se torna de difícil comprensión para el lector. Por tanto, se requiere una interpretación adecuada del contenido de los textos para poder crear representaciones mentales coherentes y de esta manera lograr que cada lector encuentre significados que ayuden a construir su conocimiento para poder captar su contenido (Van den Broek, 2010).

(Campos Saavedra et al., 2014) manifiesta que la facilidad o dificultad en la que un texto logra ser leído y comprendido se lo entiende como lecturabilidad. Entender y conocer cómo de adecuado es un texto para una persona es una problemática que todavía no está resuelta. Se ha investigado sobre la legibilidad de un texto para cada individuo, pero no es una tarea sencilla, puesto que cada lector presenta destrezas diferentes (Lopez-Anguita et al., 2018). Es importante recalcar que las palabras inusuales pueden influir en la creación de barreras de accesibilidad en las personas con discapacidad intelectual provocado por la alta dificultad presentada en la lectura y en la compresión de texto.(Alarcon, 2020).

Para enfrentar esta problemática se ha recurrido a la necesidad del uso de algoritmos de Machine Learning y técnicas de Deep Learning como las redes neuronales que permitan predecir la complejidad de las palabras a partir de un conjunto de datos (Shardlow, Evans, & paetzoold, 2021). El objetivo de este trabajo es predecir el nivel de complejidad más cercano a la realidad de las palabras mediante la aplicación de un *Transformer* lingüísticos y la ejecución de la técnica de Fine-Tuning para contribuir en investigaciones de la Predicción de

las palabras complejas en el idioma español y contribuir a la reducción de las barreras de la compresión lectora en los textos académicos universitarios de la Carrera de Ingeniería en Sistemas Computacionales y Software de la Universidad de Guayaquil.

Capítulo I, presenta la definición del problema en un contexto específico, situación conflicto nudos críticos, delimitación del problema, evaluación del problema, las causada y consecuencias, formulación del problema, los objetivos planteados tanto en general y específicos, el alcance del proyecto, justificación e importancia y las limitaciones del estudio.

Capítulo II, expone los antecedentes del estudio, fundamentaciones teóricas, revisiones sistemáticas, meta-análisis y preguntas a contestarse, se definirá las variables de la investigación y las decisiones conceptuales.

Capítulo III, contempla la modalidad de la investigación, tipo de la investigación, diseño metodológico de la investigación, la metodología, población, técnicas de recolección de datos para poder determinar la factibilidad de este proyecto, técnicas estadísticas para el procesamiento de la información, beneficiarios directos e indirectos del proyecto, entregables del proyecto, propuesta, criterios de validación de la propuesta y los resultados de esta investigación.

CAPÍTULO I

PLANTEAMIENTO DEL PROBLEMA

Descripción de la situación problemática

Ubicación del problema en un contexto

La Comprensión léxica es un complemento clave de la lengua, que nos otorga el significado y forma la base de nuestra comunicación. A partir de un conocimiento del vocabulario limitado siempre restringe la habilidad de comunicar de manera efectiva, por lo cual un desarrollo del vocabulario mejora enormemente la competencia comunicativa. (Aristu Ollero & Torres Ríos, 2021). Este es un factor importante en todos los niveles de educación porque delinea o estructura su nivel profesional, de modo que los estudiantes puedan comprender el contenido de lo que está escrito en los textos académicos.

Estas barreras e la comprensión lectora pueden obedecer a diferentes factores, tales como: el empleo de un vocabulario especializado, uso de palabras poco frecuentes, oraciones muy largas, entre otros, lo cual se torna en una limitación para el aprendizaje para los estudiantes universitarios, que son individuos que cuentan con un nivel alto en educación superior, y conocimientos especializados en diferentes áreas de estudio, pero aun así, estos están formando parte de los grupos de personas con discapacidad lectora. (Alarcon & Moreno, 2020)

Las palabras que se encuentran en los textos universitarios son escritas en su mayoría con un léxico especializado, ya que el contenido de los textos va acorde con la asignatura. De esta forma, las palabras técnicas y especializadas constituyen una barrera inclusive para

los estudiantes que tienen una formación académica superior. (Ortiz Zambrano, MontejoRáez, Lino Castillo, Gonzalez Mendoza, & Cañizales Perdomo, 2019)

Situación conflicto nudos críticos

Diversas fuentes se mencionan que la sociedad actual se caracteriza por interpretar o producir textos y contenidos. El desenvolvimiento dentro del orden social, política, cultural y demás ámbitos se encuentra mediado por la capacidad para utilizar códigos y competencias para manejar el carácter especializado de los conocimientos, lo que resulta fundamental que el entendimiento de un texto es de gran relevancia para el desarrollo individual y colectivo de una sociedad. (Garcia Garcia, Arevalo Duarte, & Hernandez Suarez, 2018)**.**

Por otra parte, según las investigaciones de (Alban, Arguello, & & Molina, 2020)**.** se refleja la preocupación de la docencia por la visible dificultad que enfrentan las estudiantes provocadas por el bajo nivel de compresión lectora. Mostrando las premisas iniciales que la comprensión lectora es parte integral de un concepto mayor de la lectura, escritura, la cual, a su vez, produce un sistema económico social de desigualdad social y cultural de clase. (Ortiz Zambrano & Montejo Ráez, 2017)

El problema surge cuando el estudiante se enfrenta al desconocimiento de las palabras que le resultan de difícil comprensión, lo que conlleva a la docencia ir más allá de sus obligaciones académicas para que el estudiante llegue a comprender lo leído y producir el conocimiento sobre el objeto de estudio de la disciplina enseñada. (Ortiz-Zambrano & Montejo-Ráez, 2021).

Delimitación del problema

La presente investigación se enmarca en el campo del Procesamiento del Lenguaje Natural, específicamente en el área de la Simplificación Léxica orientada al idioma español. El proyecto pretende investigar la predicción de la complejidad de las palabras que se encuentran en el contenido de los textos académicos de las carreras de Ingeniería en Sistemas Computacionales y la carrera de Software, ambas de la Facultad de Ciencias Matemáticas de la Universidad de Guayaquil.

Plantea la ejecución de un modelo de lenguaje enmascarado basado en *Transformers* para el idioma español denominado RoBERTa-large-BNE, y la aplicación de una técnica de afinamiento de los datos *Fine-Tuning* para usarla como inicialización para un nuevo modelo entrenados con el mismo conjunto de datos y lograr mejores resultados en la predicción de las palabras difíciles.

A continuación, se detalla las fases que componen esta investigación:

1. Varios Transformers como BERT, XLM-RoBERTa y RoBERTa-Large-BNE son aplicados en un conjunto de datos en español provenientes de los textos académicos (Zambrano, 2021)
2. Se ejecuta un conjunto de algoritmos de Machine Learning para determinar el nivel de Predicción de la complejidad de las palabras sobre el conjunto de datos proporcionado de la aplicación de las redes neuronales.
3. Se aplica la técnica de Fine-Tuning, con el objetivo de obtener un mejor resultado en el nivel de predicción de la complejidad de las palabras.

4. Finalmente, el modelo ajustado es ejecutado sobre el conjunto de datos. Luego, mediante la ejecución de diferentes algoritmos de Machine Learning poder obtener la predicción de la complejidad de las palabras.
5. Evaluar los resultados provenientes de los diferentes modelos de lenguajes de Transformers (BERT, XML-RoBERTa y RoBERTa-large BNE) versus los oresultados btenidos del nuevo modelo ajustado.

Tabla 1: Delimitación del problema

Delimitador	Descripción
Campo	Procesamiento del Lenguaje Natural.
Área	Simplificación Léxica.
Aspecto	Predicción de la complejidad de las palabras en los textos académicos universitarios en el idioma español.
Tema	Evaluación de desempeño de los modelos de *Transformers* para la predicción de la complejidad léxica para el idioma español.

Nota: En esta tabla se plantean los términos de análisis aplicados para la delimitación del problema conforme al contexto en donde se desarrolla la problemática.

Evaluación del Problema

Los aspectos generales de evaluación son:

- **Delimitado:** Actualmente, es evidente que la comprensión de un texto en estudios de computación en idioma español les resulta difícil a muchos de los estudiantes de la Universidad de Guayaquil específicamente de la Carrera de Ingeniería en Sistemas Computacionales y la carrera de Software; esto es debido a la complejidad de las palabras con las que son escrito los documentos.
- **Original:** El tema propuesto es auténtico, ya que las investigaciones demuestran que esta área de la Predicción de las Palabras complejas y el uso de *Transformers* en los

textos académicos universitarios no ha sido aplicada, particularmente en textos de estudios de computación.

- **Concreto:** Existe una problemática que es real y que está presente en los estudiantes de las carreras en estudios computacionales. Concretamente la problemática está en la presencia de palabras complejas en el contenido de los textos que utilizan los estudiantes para su formación (Mite-Baidal, y otros, 2018).En el contenido se los textos se ha demostrado que está formado por palabras técnicas, sofisticadas, entre otros, las mismas que se vuelven complejas formando barreras que complican su comprensión lectora. (Castillo, K. N. L., Mendoza, O. R. G., & Perdomo, B. C. C., 2019).
- **Relevante:** Su relevancia está en contribuir con las investigaciones en la predicción de la complejidad de las palabras dirigida a los textos de computación de los estudiantes universitarios, lo que permitirá presentar soluciones que aporten al conocimiento de las palabras difíciles mediante propuestas de Simplificación Léxica.
- **Contextual:** el problema se presenta en el contexto de la educación superior, específicamente en los estudiantes de las carreras de Ingeniería en Sistemas Computacionales, y la carrera de Software. Los estudios se realizaron en los primeros 4 semestres de la carrera de Software, y del quinto al octavo semestre de la carrera de Sistemas (Zambrano, J. A. O., & Montejo-Ráez, A, 2021).
- **Factible:** Este proyecto es realizable, debido a que la propuesta de desarrollo se lo realizará mediante el empleo de software libre y lenguaje de programación que permite la aplicación de diferentes modelos de *Transformers* que aportan con sus características al estudio de las palabras complejas y su aplicación mediante el uso de los diferentes algoritmos de Machine Learning.

- **Variables:** Se pueden identificar las variables dependientes e independientes
    - **Variable Dependiente:** Evaluación de desempeño aplicando modelos de *Transformers*.
    - **Variable Independiente:** Predicción de la Complejidad Léxica de las palabras en idioma español.

Causas y consecuencias del problema

El problema principal predomina en la falta de entendimiento de los textos académicos del idioma español por su nivel de complejidad léxica. Esto es debido a diferentes factores, a continuación, se muestra en la tabla de causas y consecuencias.

Tabla 2: Matriz de causas y consecuencias del problema

Causas	Consecuencias
C1. Dificultad de comprender textos académicos en español.	E1. Nula o errónea compresión del texto académico.
C2. Desconocimiento de términos o palabras científicas.	E2. Incorrecta interpretación léxica
C3. Desmotivación para aprender de textos académicos en español.	E3. Bajo nivel académico y desconocimiento.
C4. Carencia de estudios de los textos académicos por su nivel de complejidad lingüística.	E4. Dificultad de elección de textos académicos.
C5. Desmotivación para aprender nuevas palabras del español.	E5. Crecimiento de la desigualdad estudiantil y social.

Formulación del problema

A las personas les resulta difícil comprender cuando se encuentran con términos nuevos al momento de realizar alguna lectura ya que se encuentran con palabras homófonas o técnicas (Arroba & Pozo, 2021).

El aprendizaje de un tema puede resultar monótono cuando no se lo interpreta a la perfección ya que descubren palabras nuevas con términos de alto grado de complejidad léxica, por ese motivo se provoca un déficit de gran relevancia que afecta al lector, por ese motivo formulamos la siguiente pregunta:

¿Cómo una Evaluación de desempeño de los modelos de transformadores para la predicción de la complejidad léxica para el idioma español puede contribuir con la comprensión lectora?

Objetivos del proyecto

Objetivo general

Implementar los modelos pre-entrenados basados en *Transformers* BERT, XLM-RoBERTa y RoBERTa-Large-BNE y aplicando la técnica de Fine-Tuning evaluar cómo la ejecución del modelo ajustado sobre el modelo pre-entrenado mejora su desempeño en la predicción de la complejidad léxica para el idioma español.

Objetivos específicos

1. Explorar los modelos basados en *Transformers* BERT, XLM-RoBERTa, y RoBERTa-Large-BNE para la construcción de los embeddings (incrustaciones - encodings) aplicados a un conjunto de datos en Español.
2. Generar las incrustaciones numéricas a partir de la ejecución de los modelos pre-entrenados basados en *Transformers* para la construcción de los datasets.
3. Aplicar el ajuste fino mediante la ejecución de la técnica de *Fine-Tuning* para su ejecución sobre los modelos pre-entrenados y la generación de los datasets basados en las nuevas representaciones numéricas.
4. Entrenar y evaluar los conjuntos de datos provenientes de los modelos pre-entrenados como de los modelos ajustados aplicando los diferentes algoritmos de *Machine Learning* (Aprendizaje Supervisado) para obtener la Predicción de la Complejidad Léxica (LCP).
5. Evaluar el desempeño de los modelos basados en Transformers mediante los resultados obtenidos de la LCP para el idioma español.

Alcance del proyecto

Se implementará el transformador neuronal RoBERTa-Large-BNE en los corpus de datos comprendidos de los textos universitarios obtenido previamente del proceso de titulación del periodo 2021 CI, como complemento al conjunto de datos para el pre-entrenamiento de aprendizaje supervisado.

Requiere de los siguientes puntos:

- Implementar el sistema de predicción con el lenguaje de programación Python usando Framework Electron
- Evaluar el sistema de predicción RoBERTa-Large-BNE en comparación al sistema XLM-RoBERTa y BERT
- El sistema será puesto a un pre-entrenamiento con la información brindada por los corpus de datos de los textos universitarios en idioma Español.
- Se medirán los resultados en base a métricas de error en tareas de regresión.

Justificación e importancia

La compresión de textos con complejidad léxica es fundamental para el entendimiento, análisis y una precisión lectora adecuada, fomentar el entendimiento de la complejidad léxica a través de transformadores simplificara y clasificara el nivel de complejidad lo que favorecerá a mejorar el entendimiento léxico del español para el lector con un bajo nivel de compresión cognitivo recalcando un texto de palabras que requiera un nivel de compresión bajo (Arroba & Pozo, 2021).

Una mejor exactitud de compresión de palabras ayudara a la sociedad para continuar avanzando en el área de la educación, la simplificación léxica de textos proporcionada un mejor entendimiento del idioma español, ayudar al entendimiento de palabras proporcionara herramientas para disminuir la barrera de la compresión lectora (Kevin Gaspar, 2021).

Se aplica la identificación de palabras por la necesidad de clasificar por nivel de complejidad contextual en libros académicos de software. Así fomentando la regulación y clasificación de documentos lo que favorecerá la proporción y selección de documentos para el aprendizaje por nivel académico.

Esta clasificación de documentos proporcionara al usuario un mejor entendimiento y visión general de la complejidad del documento de su selección lo que a su vez ayudara al fortalecimiento intelectual, de aprendizaje y compresión lectora.

La complejidad de documentos es un factor importante en el estudio académico en el cual tiene gran relevancia en el proceso de aprendizaje que afecta a este si el documento presenta un nivel avanzado de complejidad que no se encuentra a la par del usuario lo que provocara un déficit de entendimiento y dificultad lectora. (Arroba & Pozo, 2021)

Esto conlleva al planteamiento de la utilización de las métricas de complejidad léxica por lo cual permitiría evaluar la complejidad de los corpus de los textos académicos de un gran número de asignaturas, permitiendo comprender el léxico que frecuentemente utilizan los docentes para impartir sus clases que normalmente resulta sofisticado y técnico que dificultad su compresión. (Zambrano & Montejo Raez, 2021)

El procesamiento del lenguaje natural o NLP (Natural Language Processing) es importante en el campo de la inteligencia artificial debido a que a través de este un sistema puede entender el lenguaje humano con sus diferentes interpretaciones.

El vocabulario de los textos de la carrera de ingeniería en sistemas puede llegar a tener un nivel de complejidad considerable, y es por esto que es importante poder determinar que palabras pueden tener ese nivel de complejidad; y qué mejor que hacerlo a través del uso de algoritmos de redes neuronales (Arroba & Pozo, 2021).

Actualmente es un proceso en el que se puede sumergir y con el que puede experimentarse de muchas maneras, pues tiene pocos años que se ha venido implementando y en nuestra región poco se conoce acerca de los transformadores para el procesamiento del lenguaje natural.

Esto también permite la investigación de aplicación de software libres que faciliten el uso de bibliotecas en el campo de procesamiento del lenguaje Natural lo que conllevara una gran contribución para el análisis de la complejidad de la compresión de textos. (Ortiz Zambrano J. &., 2018)

Lo que se busca a través de este proyecto de investigación es realizar un afinamiento/optimización de los transformadores BERT, XLM-RoBERTa y RoBERTa-Large-BNE en cuanto a características lingüísticas que ayuden en la determinación del grado de complejidad de una palabra para que así los resultados sean más concisos.

Limitaciones del estudio

- El sistema será puesto a un preentrenamiento con la información brindada por el corpus de datos de los textos universitarios en el idioma español.
- Se medirán los resultados en base a métricas de error en tareas de regresión.
- La evaluación solo funciona para corpus en español.
- No requiere del uso de internet.
- La agrupación de datos está fundamentada en variables lingüísticas como la longitud de términos complejos, el token (cantidad de sílabas), la complejidad léxica de la palabra, etc.

CAPÍTULO II

MARCO TEÓRICO

Antecedentes del estudio

El problema de categorización de textos se ha estudiado a fondo en inconvenientes de recuperación de información y labores de minería de datos. Una innovación presente tanto en la sustracción de datos como en el procesamiento del lenguaje natural llamó la atención de estudiosos de todo el planeta para desarrollar sistemas automatizados para la categorización de textos(Qasim et al., 2022).

Para esto en la actualidad el uso de Machine Learning y Redes Neuronales son implementados en trabajos que buscan una solución mediante modelos predictivos que aporten a esta nueva tendencia de interpretación de textos. (Arroba & Pozo, 2021)

Las redes neuronales son sistemas complejos de procesamiento de la información la cual su estructura y funcionamiento se encuentra inspirada en las redes neuronales biológicas. Esta consiste en un conjunto de elementos simples de procesamiento llamados neuronas conectadas entre sí por conexiones que tienen un valor numérico modificable (Asanza, Olivo, & Peñafiel, 2018). En la cual este modelo natural es la base para la arquitectura de los *Transformers.*

Los transformadores basados en redes neuronales como BERT, XLM-RoBERTa y RoBERTa-Large-BNE fueron las herramientas para el preentrenamiento del conjunto de datos para alcanzar una mayor predicción léxica del corpus de los textos académicos.

El advenimiento de los modelos de lenguaje basados en transformadores (es mencionar, basados en arquitecturas de autoatención) ha revolucionado todo el campo del procesamiento del lenguaje natural (PLN). Ya entrenados previamente a través de grandes corpus sin etiquetar y filtrar, podemos aplicar el aprendizaje de transferencia a prácticamente todas las tareas posteriores.(Armengol Estapé Advisor et al., 2021)

La arquitectura *Transformers* recientemente implementada ha revolucionado la escena de la (PLN). Primordialmente orientado a la traducción automática, los investigadores han presentado y demostrado que también sirve como un poderoso Backend de aprendizaje profundo para el entrenamiento previo de modelos de lenguaje grandes en los trabajos fundamentales orientado a la interpretación de textos.(Armengol Estapé Advisor et al., 2021)

(Segura-Bedmar et al., 2021) comenta también el impulso de esta tecnología en muchas de las taras de Procesamiento de lenguaje Natural (PLN), en aplicación de tareas de generación de resúmenes de textos académicos como esto también se ha beneficiado de estas técnicas que a su vez a dado la posibilidad de la implementación de distintos modelos logrando superar los resultados de los textos procesados aportando al entendimiento léxico de diversos textos.

(Valdés-León, 2021) Recalca que la información recogida gracias a la generación de analíticas sobre el desempeño léxico y escritura en textos de educación superior, permite tomar decisiones pedagógicas en corto plazo, si no también planificar investigaciones que comprueben la relación que existe entre ambas competencias. Para ellos, se ejecutó un estudio del caso, de tipo experimental utilizando software libre, presentando como resultado la existencia de una estrecha relación entre léxico académico y calidad del contenido.

(Alarcon, 2020) También se menciona que los textos con contenido técnicos y con palabras inusuales provocan resultado no deseados como barreras de accesibilidad y dificultad en la acción de lectura o compresión del texto. Por ende, se ha realizado un enfoque para la simplificación léxica del texto a través de herramientas de accesibilidad como un soporte sistemático al cumplimiento de estándares de accesibilidad. Como ejemplo se aplicó diferentes técnicas de aprendizaje automático (BERT) utilizando recursos de lectura fácil y lenguaje simple, dando como resultado una contribución orientada a las personas con discapacidad cognitiva.

La utilización de transformadores léxico tienen una amplia gama de oportunidades desde lo académico como lo comercial demostrando la importancia de los textos y su entendimiento para sacar el mejor provecho de estos. (José et al., 2021) demuestra que mediante la utilización del transformador Bert, modelo pre-estrenado para los comentarios en español en la tienda de aplicaciones de Google Play Store, obtiene resultados experimentales que demuestran que después del pre-procesamiento adecuado puede mostrar resultados prometedores en el conjunto de datos de español demostrando una tendencia y alcanzando una precisión de 0.81 de precisión en promedio, incluso con las limitaciones de datos.

Viendo como los cambios tecnológicos avanza a grandes pasos y como el Lenguaje Natural (LN) del ser humano lo que hablan cotidianamente se puede incorporar en las computadoras permitiendo la realización de diversas tareas que ayudan al ser humano a como una compresión de un texto extenso, traducción, redacción y otras actividades.

El estudio que realiza (Beltran & Mojica, 2021) presenta el potencial que se puede aplicar en Ingeniería de Software de la utilización de modelos de GPT-3, BERT con su arquitectura *Transformers* y otros, esto permitiendo la derivación de diversas aplicaciones como la creación de

chatbots y su relación con procesos de ciclo de vida del software. La diversidad de generalidades que se pueden aplicar de LN, del LN computarizado, Redes neuronales, Inteligencia Artificial demuestran el nivel de procesamiento de LN, lo cual permite evaluar las perspectivas de este tipo innovaciones y sus alcances.

(Jara, 2021)Visualiza que el progresivo desarrollo tecnológico está permitiendo generar una gran cantidad de datos de información en formatos digitales y menciona que el para el año 2025 habrá aproximadamente 175 zettabytes de información digital, la cual el mayor porcentaje de información estará en forma no estructurada o texto libre. Por lo cual la necesidad inminente de desarrollar nuevas tecnologías que permitan descubrir automáticamente el conocimiento relevante de dichas fuentes o como apoyo a las tomas de decisiones sobre un texto. Lo cual da apertura a la utilización de técnicas para organizar automáticamente una gran información digital como la clasificación de textos.

La categorización o clasificación de textos nos permite asignar automáticamente etiquetas predefinidas a los textos en base a su contenido, esto a través de algoritmos y redes neuronales en el campo de procesamiento del lenguaje natural (NLP) visto la opción más rentable debido a la gran disponibilidad y a las mejoras capacidad de procesamiento computacional. (Ortiz-Zambranoa & Montejo-Ráezb, 2020)

Fundamentación teórica

Predicción de las palabras

La predicción de palabras es una de las técnicas más comúnmente utilizadas para ayudar, tanto en la comunicación personal, académico y escritura, que favorece a personas naturales o con discapacidades motrices. Las personas con problemas lingüísticos son bastantemente beneficiados ya que reciben una guía con las palabras predichas en su proceso de aprendizaje siendo este un gran apoyo gramatical.(Kevin Gaspar, 2021).

Complejidad Léxica

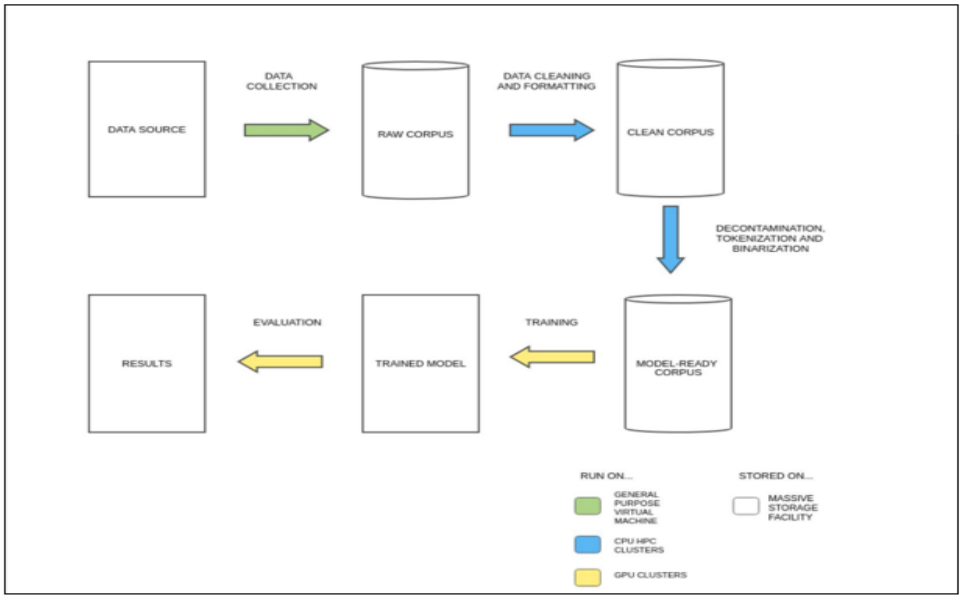
El aprendizaje de una nueva palabra implica nuevas representaciones semánticas y establecer vínculo con estándares conceptuales, algunos preexistentes en la memoria y otros en proceso de elaboración. El crecimiento léxico potencia exponencialmente el aprendizaje de nuevas relaciones semánticas y conceptuales, lo que aporta a un creciente nivel de entendimiento. Por lo cual se afirma que el léxico define claramente un campo de interacción entre el contexto y la cognición, demostrando a su vez que a una mayor complejidad léxica exige, y a la vez refleja una mayor complejidad cognitiva en distintos procesos y niveles.(Aravena & Quiroga, 2018)

Procesamiento del lenguaje Natural

Procesamiento del Lenguaje Natural o Natural Lenguage Processing (NLP) es una ciencia computarizada que vincula la inteligencia artificial y la lingüística que estudia las interacciones a través de los datos entre las computadoras y el lenguaje humano, a través de análisis sintáctico, semántico, pragmático y morfológico, observando y a la vez escribiendo los patrones estructurales, aplicando un formalismo gramatical concreto.(Manjarrés-Betancur & Echeverri-Torres, 2020)

La función principal del preprocesamiento es limpiar el texto sin procesarlo y formatearlo según sea necesario, manteniendo la coherencia del documento para aprender dependencias de largo alcance. La mayoría de los métodos existentes de recopilación y limpieza de datos para la (NLP) se han centrado en la cantidad en lugar de la calidad. Debido a que el enfoque es ser compatible con idiomas y dominios de bajos recursos, el filtrado debe ser lo más detallado posible. (Armengol Estapé Advisor et al., 2021)

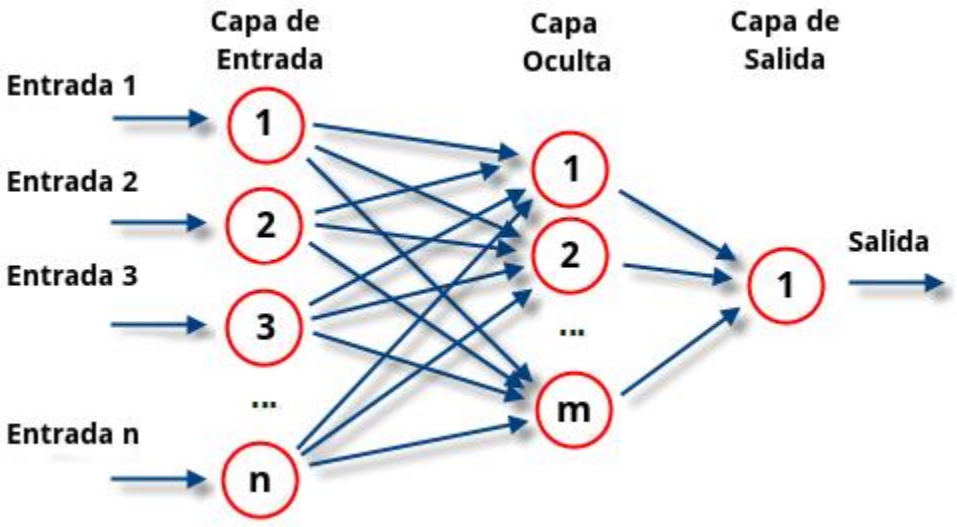
Ilustración 1: Una canalización para el pre-procesamiento de texto sin procesar de gran tamaño y el entrenamiento de modelos de lenguaje a escala



Redes Neuronales

Las Redes Neuronales artificiales tienen una estrecha relación con la Inteligencia Artificial las cuales son redes entrenadas mediante entradas obtenidas a partir de escenarios externos o internos en el sistema y estas entradas se multiplican por pesos asignados aleatoriamente. (Asanza, Olivo, & Peñafiel, 2018)

Ilustración 2: Redes Neuronales



Nota: Imagen recuperada en la página: https://www.atriainnovation.com/que-son-las-redes-neuronales-y-sus-funciones/

Características

Unas de las principales características de las redes neuronales es que se las entiende y visualiza como un esquema computacional distribuido que básicamente se asimila a una estructura de un sistema nervioso de los seres humanos.(Rivas et al., n.d.)

Topologia de red : Posibilita establecer la capacidad representativa de como una proporción de neuronas permanecen distribuidos en capas y distribuidas entre sí.

Regla de aprendizaje: Esta basado en un sistema de aprendizaje, es por ello que tienen la capacidad de aprender a través de un entrenamiento previo.

Tipo de Entrenamiento: Las redes Neuronales poseen dos tipos de entrenamiento, una que durante el inicio del aprendizaje de red se entrena para que los estándares sinápticos se adecuen

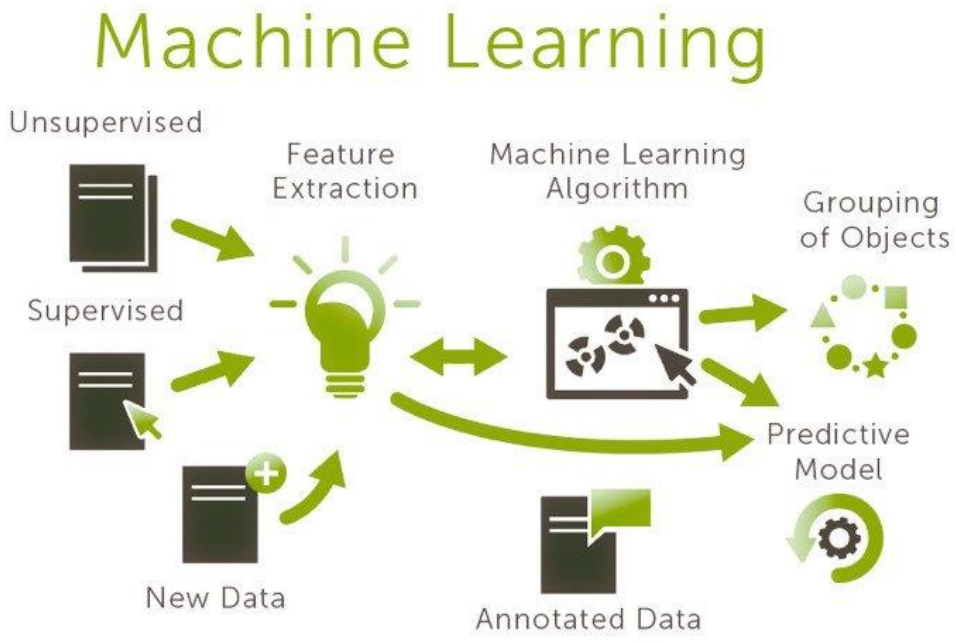
a la red. La segunda estaba basada durante la ejecución donde la red ya pasa ser operativa y a su vez toma un valor de funcionamiento real.

Algoritmos Machine Learning

Este tipo de algoritmos consigue su entendimiento de la comparación de resultados, que es dependiente del ajuste del modelo a usar, donde el mismo está formado de propiedades y etiquetas (también conocidas como costo objetivo), modelo va aprendiendo, y por consiguiente la predicción va haciéndose más positiva. (Reategui & Suarez, 2021)

El machine learning en las organizaciones en un mercado cada vez mas competitivo donde terminos como la automotización de procesos y la eficencia son cada vez mas necesarios y determinantes y donde la inteligencia competitiva avanza a pasos agigantados donde las grandes organizaciones manejan enormes volumenes.

Ilustración 3: Machine Learning



Nota: Imagen recuperada en la página: https://www.smartpanel.com/algoritmos-de-machine-learning/

Sistema de predicción

El sistema de predicción basado en algoritmos Machine Learning tiene diversos algoritmos que permiten edificar modelos capaces de examinar y presagiar datos con la intención de utilizar técnicas para mirar la regresión de datos o factor examinado, en lo que el segundo regresa un costo, con base a las propiedades que se han determinado como etiquetas para su entrenamiento. (Arroba & Pozo, 2021)

TRANSFORMERS

Los *Transformers* son un modelo de aprendizaje profundo introducido en 2017, usado primordialmente en el campo de la (NLP). Los Transformers fueron creados para manejar datos secuenciales como lenguaje natural, los transformadores no necesitan datos secuenciales para ser procesados en orden, por consiguiente, los transformadores no requieren procesar el inicio de una sentencia anterior a procesar el desenlace. (Wolf, y otros, 2019)

La parte más crucial de un Transformador es el mecanismo de atención. Este mecanismo de atención recalca la importancia que otros tokens tienen en una entrada para la codificación de un token dado. El mecanismo posibilita que el Transformador se centre en ciertas palabras tanto en el izquierda y derecha para intentar el término de hoy según la labor de (NLP) que estamos direccionamiento. (Özçift, Akarsu, Yumuk, & Söylemez, 2021)

Otra virtud de la arquitectura Transformers es que el aprendizaje en uno el lenguaje se puede transferir a otros lenguajes por medio del aprendizaje por transferencia. A plena términos, el aprendizaje por transferencia es la iniciativa de tomar el razonamiento adquirido una vez que hacer una labor y aplicarla a una labor distinto. (Ubeda, 2021)

¿Cuál es la función con los transformers?

Existen 2 etapas y son las siguientes:

PRE TRAINING

En la primera etapa, este modelo aplica la configuración del lenguaje de manera habitual, tambien puede adquirir aprendizaje generico del concepto del termino. (Humeau, Shuster, Lachaux, & Weston, 2019)

FINE TUNING

En la segunda etapa, a este modelo se le debe añadir varias partes a la arquitectura y asi ajustar los modelos a los trabajos determinados.

Embeddings

Los Embeddings son una técnica que se basa en representar palabras en vectores de números, lo cual permite mejorar de manera significativa las labores de hallazgo de entendimiento y de recomendación de contenido. (Kalyan & Sangeetha, 2020)

BERT

BERT (Bidirectional Encoder Representations from Transformers) se basa en un modelo para representar datos, de forma que logren alimentar a la red neuronal y se clasificados con una más grande exactitud. sistema que tiene presente el entorno de los vocablos basándose en el entrenamiento bidireccional de un Transformador, un modelo de atención conocido. (Zapata Garcia, 2021)

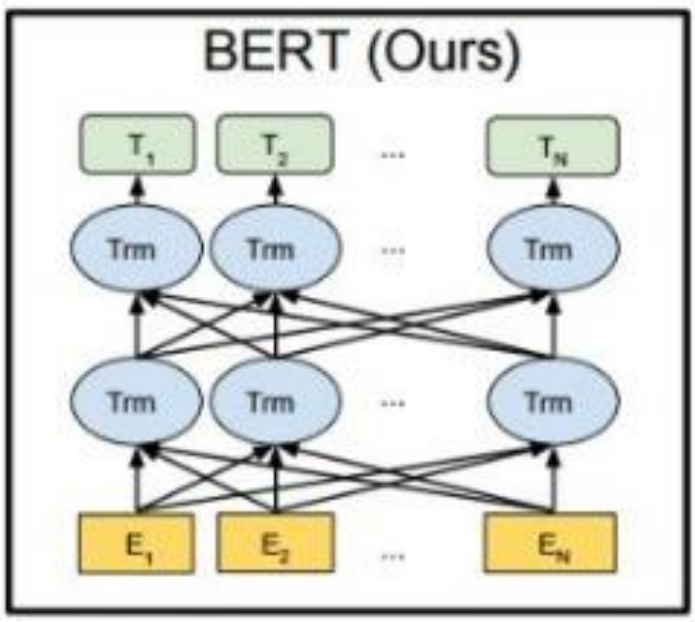
En el trabajo de titulación de (Collarte Gonzale, 2020) se publicó un estudio “BERT: Entrenamiento previo de transformadores bidireccionales profundos para la comprensión del

lenguaje" que abrió un nuevo horizonte de posibilidades en el campo del PLN, al introducir el entrenamiento bidireccional de transformadores.

El propósito primordial de BERT es practicar representaciones bidireccionales desde un grupo de datos sin etiquetar. BERT es sencilla sin embargo poderoso. Un modelo de ajuste fino de BERT solo requiere añadir una capa más para que cada nuevo modelo haga una diversidad de labores.(Qasim et al., 2022)

La base de BERT es en comparación más pequeña en tamaño comparado a otros transformadores, lleva menos tiempo de cálculo y procesamiento, y además es asequible.(Qasim et al., 2022)

Ilustración 4: BERT



Nota: Imagen recuperada en la página:https://programmerclick.com/article/48161635718/

La manifestación de BERT simula unir todas las mejorías de los demás modelos y excluir sus carencias y así poder alcanzar los excelentes rendimientos en varias tareas.

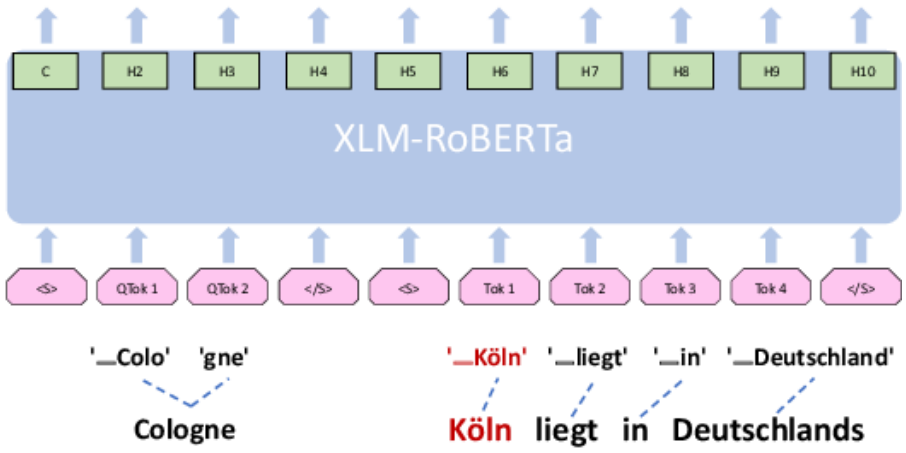
XML-RoBERTa

El modelo XML-RoBERTa busca mejorar el funcionamiento de BERT original al hacer utilizando lotes de ingreso más grandes y usar más datos de ingreso. entrenado en 100 idiomas utilizado un vocabulario de 250 mil palabras (pudiendo ser palabras completas o solo partes de la palabra), a diferencia del vocabulario usado. (Erick Quezada, 2020)

Otras características que menciona (Ubeda, 2021)XML-RoBERTa consigue mejoras fundamentales y estos cambios integran:

- Ejecutar el modelo por más tiempo con lotes mayores y más datos,
- Entrenamiento en secuencias más largas, y
- Cambiando dinámicamente las posiciones enmascaradas a lo largo del pre-entrenamiento.

Ilustración 5: XML-RoBERTa



Nota: Imagen recuperada en la página: https://www.catalyzex.com/paper/arxiv:2101.11112

RoBERTa-Large-BNE

Es un prototipo de lenguaje enmascarado fundamentado en transformadores para el idioma español. Este modelo ha sido pre-entrenado utilizando el corpus español más extenso y conocido hasta la actualidad, con un resultado de 570 GB de texto limpio, compilado a partir de las exploraciones web realizados por la Biblioteca Nacional de España (BNE) (Mendizábal, y otros, 2021)

El corpus de entrenamiento ha sido Tokenizado beneficiado en el modelo RoBERTa original con un tamaño de silabas de 502622 tokens. El entrenamiento precedente de RoBERTa-large- BNE se basa en un entrenamiento de modelo de lenguaje enmascarado que sigue el planteamiento empleado para RoBERTa-large (Mendizábal, y otros, 2021)

Validación Cruzada

Es un método que se utiliza para la estimación del rendimiento estadísticos y que se certifique que son rendimientos independientes como el algoritmo de Random Forest (Messina Valverde, 2018)

Máquina de Vectores de Soporte (SVM)

En el aprendizaje automático, otra técnica en común que puede utilizarse para la regresión, clasificación o demás tareas es una máquina de vectores de soporte, es decir, en un espacio infinito o dimensional, una máquina de vectores de soporte para construir un conjunto de hiperplanos. (Sarker, 2021)

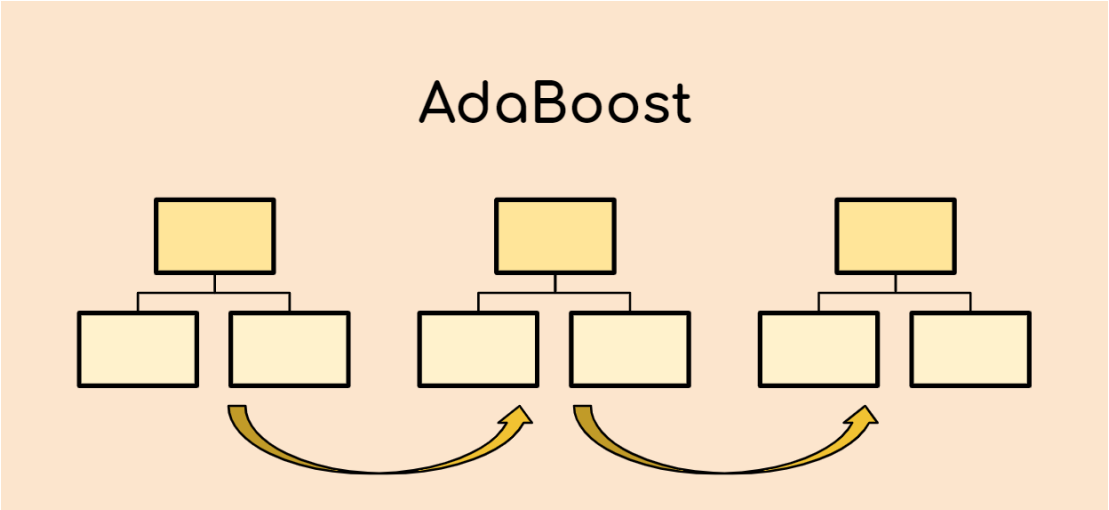
Regresión de Aumento de Gradiente (GBR)

Es una técnica del aprendizaje automático, es utilizado en tareas de clasificación y regresión, esta técnica proporciona un modelo de predicción de manera de conjunto de modelos de predicción débil que suelen ser arboles de decisión, es decir, cuando un árbol de decisión es el aprendiz débil, este algoritmo diferenciado se llama árbol potenciado por gradiente, supera al bosque aleatorio. (Kevin Gaspar, 2021)

Regresor de AdaBoost (ABR)

Es un meta estimador que empieza adaptando un regresor en el grupo de datos originales y luego adapta copias adicionales del regresor en el mismo grupo de datos, es decir, donde los pesos de estas instancias se adaptan de acuerdo con el error de la predicción actual, los regresores posteriores se concentran en los casos complejos. (Patil, 2018)

Ilustración 6: Regresor de AdaBoost (ABR)



Nota: Imagen recuperada en la página: https://towardsdatascience.com/a-mathematical-explanation-of-adaboost-4b0c20ce4382

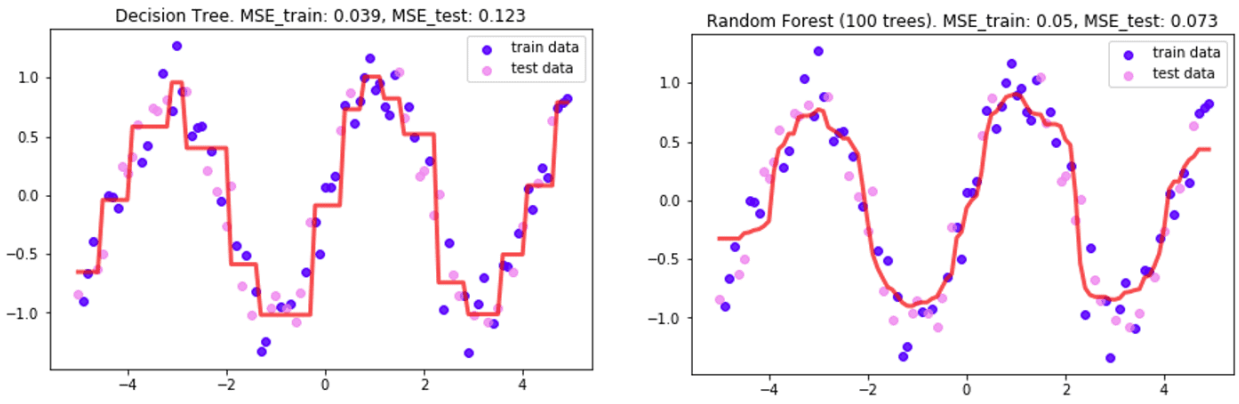
Regresión Vecina (KNR)

Es una técnica no paramétrica, que de forma intuitiva se acerca a la asociación entre las variables independientes y el resultado, el analista debe establecer el tamaño del conjunto o puede elegir mediante validaciones cruzadas. Esta técnica es bastante atractiva, fácilmente se vuelve poco practica cuando la dimensión aumenta. (Kevin Gaspar, 2021)

Regresión de Bosque Aleatorio (RFR)

Es un método de aprendizaje supervisado que usa un método de aprendizaje en grupo para la regresión. Este método de aprendizaje en grupo es una técnica que une predicciones de muchos algoritmos de aprendizaje automático para realizar una predicción más precisa que un solo modelo. (Penadillo Palomino, 2021)

Ilustración 7: Regresión de Bosque Aleatorio (RFR)

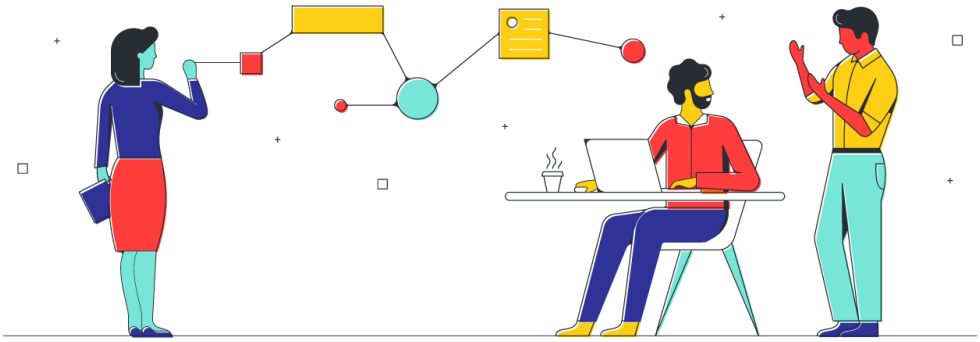


Nota: Imagen recuperada en la página: https://www.iartificial.net/random-forest-bosque-aleatorio/

Árbol de Decisión (DT)

El método de aprendizaje de árbol de decisión (DT) es utilizado para ambas clasificaciones como para tareas de regresión, Un árbol de decisión clasifica la instancia, es decir, verificando el atributo definido por este nodo, empezando desde el nodo raíz del árbol luego descienden las ramas del árbol. Las instancias se clasifican confirmando el atributo determinado por este nodo (Brenes Jiménez, 2020).

Ilustración 8: Árbol de Decisión (DT)



Nota: Imagen recuperada en la página: https://lucidspark.com/es/blog/como-hacer-arboles-de-decisiones

Regresión Pasiva Agresiva (PAR)

Este método pertenece a la categoría de aprendizaje en línea en el aprendizaje automático. El método PAR no es uno de los algoritmos más comunes, pero son utilizados para obtener resultados eficientes que soluciones problemas basados en regresión. (Meza Rodríguez, 2018)

Descenso de Gradiente Estocástico (SGD)

Es un método sencillo pero muy eficiente para adaptar regresores bajo funciones de perdidas prominente como máquinas de vectores de soporte lineales y regresión logística. Este método SGD se ha aplicado a problemas de aprendizaje automático dispersos y a gran escala. (Sánchez, 2021)

Las ventajas del Descenso de Gradiente Estocástico son las siguientes:

- Fácil de implementar.
- Eficiencia

Regresor Lineal Múltiple (RLM)

Este método trata de adaptar los modelos lineales entre una variable dependiente y varias variables independientes, es decir si el número de predictores es mayor, antes de adaptar un modelo de regresión con todos los predictores, se debería usar las técnicas de selección de modelo paso a paso para descartar los predictores que no estén asociados con los resultados. (Florian, 2021)

Frecuencia Absoluta

Se determino frecuencia Absoluta como la cantidad de oportunidades que repite en un grupo de información, las mismas que son fundamentales cuando se estudia un corpus. (Reategui & Suarez, 2021)

Frecuencia Relativa

Medida estadística que es calculada como el cociente de una frecuencia absoluta de un valor de muestra o población (fi), entre el resultado de los valores muestra o población(N) (Arroba & Pozo, 2021)

La fórmula es la siguiente:

$$hi = \frac{fi}{N}$$

- N= Número total de examinación de muestras.
- Hi= Frecuencia relativa de examinación i-ésima.
- Fi= Frecuencia absoluta de examinación i-ésima.

Longitud de Palabra: Es el tamaño de una palabra que hace referencia a la cantidad de bits que tiene una palabra (Arroba & Pozo, 2021).

Herramientas para Desarrollo

Python

Es un lenguaje de programación que está basado principalmente en la interpretación del código desarrollado, posee de forma nativa aspectos prácticos orientados a objetos. Ingresar a este lenguaje de forma eficaz es sencillo ya que permite ser progresivos para así añadir elementos prácticos. (Vidal-Silva, 2021)

Las características del lenguaje son:

1. Es un lenguaje de multifunción, es decir, posee distintos estilos del lenguaje de programación ya sea imperativa, funcional u orienta a objetos.
2. Es muy sencillo, ya que no necesita especificación del tipo de datos, se acoplan las variables en la ejecución del programa.
3. Es un lenguaje que posee objetivos generales, esto quiere decir que no está orientado solo a un fin, sino que puede crear desde scripts, paginas, hasta la innovación de un software.

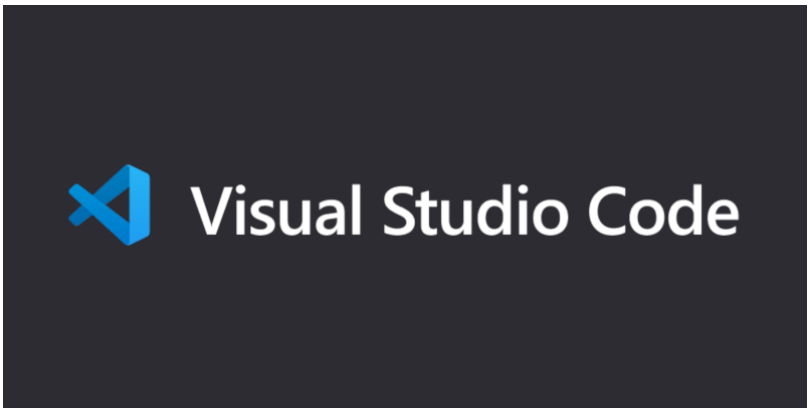
Ilustración 9: PYTHON

Nota: Imagen recuperada en la página: https://www.python.org/

Visual Studio

Es un editor de código libre recomendado por muchos programadores por su potencial y facilidad al dar soporte a muchos lenguajes de programación (Studio, 2019)

Ilustración 10: VISUAL STUDIO



Nota: Imagen recuperada en la página: https://visualstudio.microsoft.com/es/

Librería nltk

NLTK (Natural Languaje ToolKit) es la plataforma jefe para edificar programas en Python con datos del lenguaje humano. Con más de 50 cuerpos y recursos léxicos simples de utilizar, junto con librerías de procesamiento. Esta librería está enfocada en el Procesamiento del Lenguaje Natural para permitir y facilitar la utilización de IA. (Arroba & Pozo, 2021)

Librería Torch

Es una librería enfocada para el cálculo de tensores multidimensionales, se creó por Facebook y utilizado por este, Twitter y Google. Se usa para realizar cálculos numéricos utilizando programación de tensores, además posibilita la ejecución en GPU pudiendo precipitar los cálculos es decir permite crear y entrenar redes neuronales de más eficiente. (Belzunegui Gabilondo, 2020)

Hipótesis / Preguntas científicas a contestarse

¿Cómo una Evaluación de desempeño de los modelos de transformadores para la predicción de la complejidad léxica para el idioma español puede contribuir con la comprensión lectora en la Carrera de Ingeniería en Sistema Computacionales e Ingeniería en Software de la Universidad de Guayaquil?

Variables de la investigación

Las variables son las siguientes:

Variable Independiente: Algoritmo ML

Variable Dependiente: Identificación de complejidad léxica.

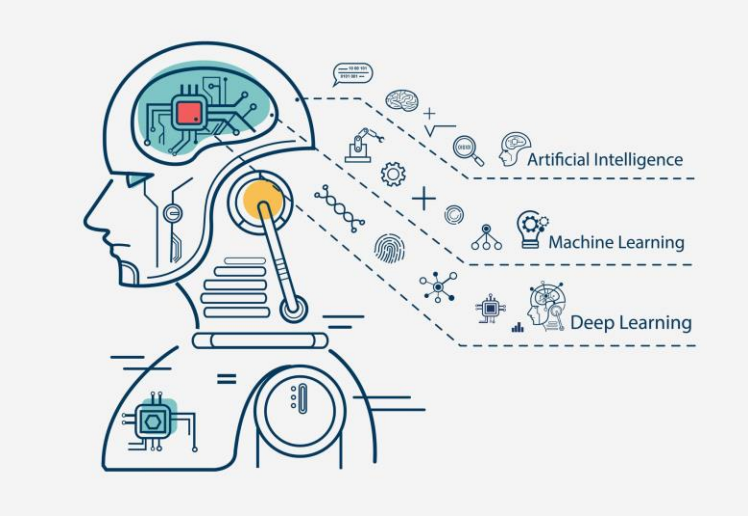
Definiciones conceptuales

Léxico: Es un conjunto de palabras que constituye un lenguaje además funciona en diferentes niveles, lenguaje formal y lenguaje informal. El lenguaje Formal es ampliamente utilizados y aceptados por las instituciones del idioma y el Lenguaje Informal es adaptado a sus necesidades en cada comunidad, generando así un léxico comunitario. (Zagorulko, 2021)

Algoritmo: Un algoritmo es una secuencia de instrucciones elementales que tiene como propósito realizar acciones o programas. (Cornelio, 2019)

Machine Learning: Es conocido como el maestro de reconocimiento de patrones, El aprendizaje automático es una rama de la Inteligencia Artificial que se basa en construir sistemas capaces de entender por sí mismos. (Uddin, Khan, Hossain, & Moni, 2019)

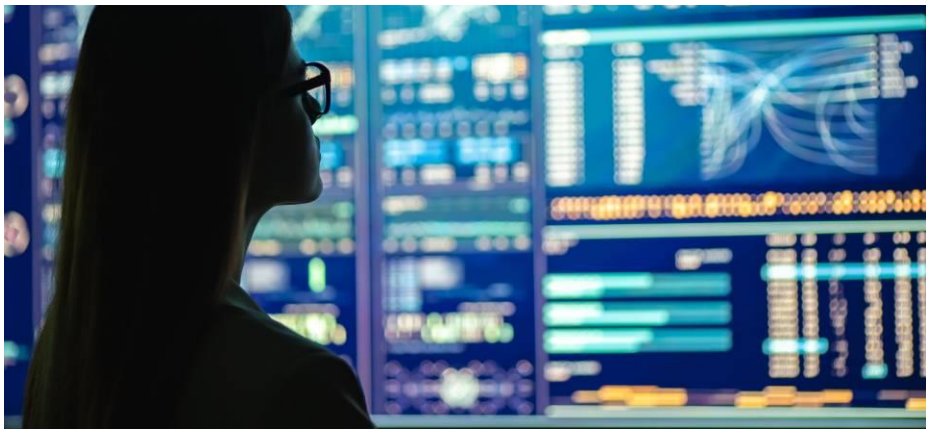
Ilustración 11: Machine Learning



Nota: Imagen recuperada en la página: https://www.atriainnovation.com/el-machine-learning-en-la-industria/

Inteligencia Artificial: Es el aprendizaje de una máquina de presentar el mismo razonamiento, creatividad y capacidad de planear igual que el ser humano. (Ufarte Ruiz & Manfredi Sánchez, 2019)

Ilustración 12: Inteligencia Artifcial



Nota: Imagen Recuperada en la página: https://www.sydle.com/es/blog/inteligencia-artificial-61896222830b254194ec71b1/

Complejidad Léxica: Se lo designa así por la gran diversidad de términos que son usados por editores en la cual se busca la complejidad del desarrollo de la comprensión con los textos descritos en introducciones de diferentes editoriales de investigación. (Borbor Merejildo & Labre Hidalgo, 2021)

Tokenización: Se basa en convertir datos perceptibles en formatos encriptados denominados tokens. La tokenización es una de las mejores estrategias de seguridad de datos que pueden incorporarse en los diferentes sistemas de pagos en Commerce (Corrales Beltrán, 2020)

Comprensión Lectora: La compresión lectora es captar, comprender o asimilar un texto de lectura. Es una actividad compleja y más cuando se está empezando a leer. La dificultad surge por la unión de diferentes procesos cognitivos. (Palacios Villalobos, 2020)

NLP: Procesamiento de lenguaje natural es una rama de la Inteligencia Artificial, ayuda a las computadoras a comprender el lenguaje humano. El procesamiento de lenguaje natural toma componentes prestados de muchas disciplinas añadiendo la ciencia de la computación y la lingüística computacional. (Chulilla Alcalde, 2021)

Métrica: Son valores expresados usando porcentajes o unidades y se adquiere a partir de una herramienta de medición ya establecida que muestra un valor que mide un proceso o actividad. (Gonzalo Fuentes, 2019)

Hand Crafted Features: Las características hechas a mano son tradicionalmente combinadas con un clasificador a través de vectores de soporte que permite que contrasten con las redes neuronales convencionales (Hssayeni, Saxena, Ptucha, & Savakis, 2017)

CAPÍTULO III

METODOLOGÍA DE LA INVESTIGACIÓN

El objetivo de este proyecto de titulación proporcionará para proseguir con los estudios en el entorno de la simplificación léxica en el cual se van a introducir los algoritmos de Machine Learning.

Este proyecto interpreta una investigación de tipo descriptiva la cual se permitió realizar una recolección de información en el entorno de la simplificación léxica para la identificación de palabras complejas.

Modalidad de la investigación

Aproximadamente el 30% del trabajo de investigación se realizó mediante investigación documental y el 70% se realizó mediante revistas bibliográficas electrónicas con el objetivo de extraer material relevante que sustentara el proyecto.

Se estudiaron múltiples fuentes a través de la investigación bibliográfica con el objetivo de tener una comprensión más amplia del tema tratado en este proyecto de tesis y de igual forma, la investigación documental determinó la dificultad del tema de estudio.

La averiguación bibliográfica se hace por medio de la utilización de datos secundarios que sirven como primordial fuente de recolección de información, donde se consigue resolver inconvenientes por medio de 2 aspectos. Interacción de los datos existentes que provienen de distintas fuentes y generando una visión panorámica y sistemática fabricadas en varias fuentes

dispersas que tiene un asunto en común. En resumen, la averiguación bibliográfica radica en la indagación, recopilado, organización, valoración, crítica e información de los datos bibliográficos.

Tipo de investigación

Investigación descriptiva

La investigación descriptiva es un tipo de investigación que se enfoca en explicar la población, situación o evento bajo investigación. También obtiene información sobre el evento o escenario que le interesa a través de enfoques como la observación y la encuesta, entre otros. (Alban, Arguello, & & Molina, 2020)

Diseño metodológico de la investigación

El diseño metodológico usado en esta averiguación esta con base en el Análisis de Casos. Se seleccionaron dos de sus técnicas:

Estudio de Discurso

Esta técnica permitió examinar el contenido de las clases impartidas por los docentes cuya fuente estaba almacenada por medio de grabaciones situadas en la nube de cada docente. Para eso se llevó a cabo del siguiente proceso.

- Estudio del grupo de datos implementando medidas de rendimiento
- Construcción de un sistema que permitió el etiquetado de los vocablos complicados.

Estudio Documental

Esta técnica permitió recopilar información científica perteneciente de: artículos, libros, revistas, entre otros. Se hizo una revisión documental por medio de la aplicación de diferentes motores de averiguación en algunas fuentes científicas de los últimos 5 años sobre diferentes temáticas relacionadas en el presente tema de indagación, ejemplo:

- Métricas de dificultad léxica para el idioma español.
- Matriz de confusión como procedimiento de validación

Metodología de investigación

Análisis de casos es un estudio en dificultad de una situación especial, utilizado para reducir un campo bastante largo de indagación hasta lograr un asunto de manera sencilla investigable. (Concepcion-Toledo, 2019)

Se detallan en 3 fases que detallan al Estudio de Casos:

- Fase Analítica
- Fase Teórica
- Fase trabajo de Campo

Fase Analítica

El análisis de contenido es una técnica de investigación cualitativa para aprender. Se utiliza para investigar datos cualitativos de manera objetiva, sistemática y cuantitativa. Saca conclusiones válidas y fiables sobre el entorno. Se han aplicado y posibilitado técnicas de análisis interno o de contenido, genera características de idioma diferentes a través de esa aplicación que las que ya

están almacenadas, existen nuevas características del lenguaje obtenidas mediante la aplicación de dos métodos basados en redes neuronales: XLM-RoBERTa y BERT. (Reategui & Suarez, 2021)

Fase Teórica:

Este trabajo se consideró las experimentaciones realizadas en la investigación de predicción de palabras compuestas en la aplicación de Aprendizaje Automático y redes neuronales. Los estudios previos muestran varios resultados de los algoritmos que determina la mejor predicción en función de las características lingüística de palabras. (Arroba & Pozo, 2021)

Fase Trabajo de Campo:

La exploración de contenido es una técnica de indagación cualitativa para aprender y analizar datos cualitativos de manera objetiva, sistemática y cuantitativa, utilizada para hacer inferencias validas y confiables con su ámbito. Tras los resultados de la implementación del método Bert y XLM Roberta, el algoritmo de Machine Learning- Random Forest Regressor para obtener niveles de dificultad de las palabras después del aprendizaje. Se realizaron varios experimentos para obtener predicciones mucho más preciso.

Población y muestra

Población.

La población se tomó en consideración a los alumnos que participaron en el proceso de etiquetación de los vocablos complicados de los materiales de las clases grabadas en línea que dieron los profesores de la carrera de Ingeniería en Sistemas Computacionales e Ingeniería en Software.

En las siguientes tablas se muestran los alumnos que forman parte de la población por los diversos semestres de las carreras.

Tabla 3. 1: Carrera de Ingeniería en Sistemas Computacionales.

SEMESTRE	**POBLACIÓN**
Quinto	3
Sexto	3
Séptimo	3
Octavo	3
TOTAL	**12**

Nota: Se muestra el resultado de los alumnos de la carrera de Ingeniería en Sistemas Computacionales.

Tabla 4. 2: Carrera de Ingeniería en Software.

SEMESTRE	**POBLACIÓN**
Primero	3
Segundo	3
Tercero	3
Cuarto	3
TOTAL	**12**

Nota: Se muestra el resultado de los alumnos de la carrera de Ingeniería en Software.

De los cuales se obtuvo 23 características manuales (Hand Crafted Features) que previamente serán utilizados para el pre-entrenamiento de los Transformers y cuáles son las siguientes:

Unidades Manuales

Las unidades manuales son el conjunto de datos que se basaron en las características que arrojaron las palabras cuyo origen eran los textos académicos del corpus en español CLEXIS2 (en proceso de publicación). (Kevin Gaspar, 2021)Tales características que en su totalidad consta de 23 unidades fueron entrenadas de manera manual (Hand Crafted Features) y son fundamentales para el trabajo de investigación ya que tales datos también serán de utilidad para el pre-entrenamiento y afinamiento de los modelos Transformers VER, XML-RoBERTa y RoBERTa-Large-BNE.

Tabla 5: Unidades de Análisis

Unidad de Análisis	*Detalle*
Abs_frecuency	Frecuencia absoluta de la palabra
Rel_frecuency	Frecuencia relativa de la palabra
Length	Longitud de la palabra.
Number_syllables	Número de silabas.
Token_possition	Posición de la palabra.
Number_token_sentences	Número de palabras en la oración.
Number_synonyms	Número de sinónimos.
Number_hyponyms	Número de hipónimos.
Number_hypernyms	Número de hiperónimos
Part_of_speech	Tipo de palabra
Freq_relative_word_before	Frecuencia relativa de la palabra anterior
Freq_relative_word_after	Frecuencia relativa de la palabra posterior.
Len_word_before	Longitud de la palabra anterior.

Len_word_after	Longitud de la palabra posterior
Mtld_diversity	La diversidad léxica de la palabra.
PROPN	Número de pronombres.
AUX	Número de auxiliares
VERB	Número de verbos.
ADP	Número de adverbios
NOUN	Número de sustantivos.
NN	Número de Sustantivos, singular o masivo
SYM	Número de símbolos.
NUM	Cantidad de números

Técnicas de recolección de datos.

Para el alcance del primer objetivo planteado se realizaron las ejecuciones de los transformes en español Bert y XML-RoBERTa generando tablas en las que se puede visualizar los entrenamientos ejecutados por los diferentes modelos predictivos, lo que permite la exploración de dichos Transformes y su comportamiento.

Para la creación de las tablas se utilizó un Corpus ClexIS^2 previamente creados con un total de 18.397 registros que fueron utilizados en su totalidad para el entrenamiento de datos. Procediendo en la creación de los Datasets, los transformes utilizados fueron ajustado al modelo entrenado Fine-Tuning presentado los siguientes resultados con sus diferentes algoritmos de Machine Learning.

Tabla 6: Transformers BERT

M.A.E	M.S.E	R.M.S.E	PEARSON	ALGORITMOS	NOMBRE
0.18101883	0.042933414	0.219943405	0.578349084	GBR	GRADIENT BOOSTING REGRESSOR
0.182894393	0.052394812	0.227439228	0.543482632	SVR	SUPER VECTOR REGRESSOR
0.183294857	0.05384932	0.249238128	0.439482075	KNR	KNEIGHBORS REGRESSOR
0.187349022	0.060434287	0.237346256	0.682301878	RFR	RANDOM FOREST REGRESSOR
0.191234863	0.054120122	0.250342975	0.262000142	RLM	REGRESSOR LINEAR M
0.192896872	0.059430096	0.230239254	0.619450305	ABR	ADABOOST REGRESSOR
0.201348784	0.091484935	0.293472315	0.640349264	DT	DECISION TREE
0.201390127	0.066340282	0.258921329	0.162023937	SGR	STOCHASTIC GRADIENTE REGRESSOR
0.239483939	0.085039286	0.300232375	0.053023911	PAR	PASSIVE AGGESSIVE REGRESSOR

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.6 Se muestra los resultados obtenidos por los diferentes algoritmos utilizados en base al entrenamiento de los 18.397 registros, obteniendo el Error Absoluto Medio (M.A.E) que representa la referencia de la precisión de cada algoritmo en contraste con los datos reales, también el valor cuadrático medio (M.S.E) y el error cuadrático medio (R.M.S.E), también se obtiene el valor de Pearson que representa el coeficiente de correlación y por último se distingue los 9 algoritmos de Machine Learning utilizados. Su característica principal que esta tabla fue creada a partir de 768 características Token de inicio (CLS) que es propio del modelo Base de BERT y posteriormente ajustado con el modelo Fine Tuning

Tabla 7: Bert Token modelo Fine-Tuning

M.A.E	M.S.E	R.M.S.E	PEARSON	ALGORITMOS	NOMBRE
0.162294338	0.042948395	0.201837275	0.918349232	SVR	SUPER VECTOR REGRESSOR
0.163267245	0.047232135	0.042923835	0.70834929	GBR	GRADIENT BOOSTING REGRESSOR
0.163283436	0.050124211	0.228421747	0.996728184	ABR	ADABOOST REGRESSOR
0.192034926	0.054082359	0.237239124	0.629237268	KNR	KNEIGHBORS REGRESSOR
0.169832237	0.042802343	0.20712382	0.999832483	RFR	RANDOM FOREST REGRESSOR
0.203243497	0.092384372	0.301199222	0.99777124	DT	DECISION TREE
0.200334255	0.062172193	0.254098135	0.278182774	PAR	PASSIVE AGGESSIVE REGRESSOR
0.186399453	0.052384796	0.222381274	0.397348095	SGR	STOCHASTIC GRADIENTE REGRESSOR
0.165534353	0.043923831	0.207483231	0.473790219	RLM	REGRESSOR LINEAR M

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.7 Se muestra la continuación de los resultados obtenidos por los diferentes algoritmos utilizados en base al entrenamiento de los 18.397 registros. Su característica principal que esta tabla fue creada a partir de 768 características Token finales que es propio del modelo Base de BERT y posteriormente ajustado con el modelo Fine-Tuning

Tabla 8: Bert CLS + Token modelo Fine-Tuning

M.A.E	M.S.E	R.M.S.E	PEARSON	LGORITMO	NOMBRE
0.167231281	0.048644932	0.20019238	0.88623912	SVR	SUPER VECTOR REGRESSOR
0.168043737	0.040923579	0.198921379	0.790129195	GBR	GRADIENT BOOSTING REGRESSOR
0.162110076	0.048730012	0.230018236	0.99999899	ABR	ADABOOST REGRESSOR
0.187702386	0.051023205	0.226425748	0.566342377	KNR	KNEIGHBORS REGRESSOR
0.1592372	0.037901903	0.198239181	0.988301222	RFR	RANDOM FOREST REGRESSOR
0.210239122	0.09402914	0.303726569	0.99023914	DT	DECISION TREE
0.19723013	0.064029187	-0.1	0.347340296	PAR	PASSIVE AGGESSIVE REGRESSOR
1.27239E+15	4.92919E+15	1.77239E+18	0.043827193	SGR	STOCHASTIC GRADIENTE REGRESSOR
0.168349287	0.047349212	0.269303302	0.52723812	RLM	REGRESSOR LINEAR M

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No. 8 Se muestra la continuación de los resultados obtenidos por los diferentes algoritmos utilizados en base al entrenamiento de los 18.397 registros. Su característica

principal que esta tabla fue creada a partir de 768 características Token(CLS) mas 768 características Token finales que es propio del modelo Base de BERT, y posteriormente ajustado con el modelo Fine Tuning.

Tabla 9: Bert ClS +Token +23(Hand Crafted Features) modelo Fine-Tuning

M.A.E	M.S.E	R.M.S.E	PEARSON	LGORITMO	NOMBRE
0.166823128	0.038812981	0.22202398	0.062912846	SVR	SUPER VECTOR REGRESSOR
0.160009359	0.036702388	0.197901155	0.820239183	GBR	GRADIENT BOOSTING REGRESSOR
0.161093889	0.050619275	0.22428396	0.999871205	ABR	ADABOOST REGRESSOR
0.17230143	0.05400912	0.2382312	0.543829141	KNR	KNEIGHBORS REGRESSOR
0.161017645	0.040182363	0.198892168	0.998767321	RFR	RANDOM FOREST REGRESSOR
0.213928171	0.097023964	0.307120381	0.99778114	DT	DECISION TREE
4.30299E+15	9.30989E+34	1.73824E+17	0.320129348	PAR	PASSIVE AGGESSIVE REGRESSOR
1.511E+17	1.01029E+42	5.99012E+15	0.036239186	SGR	STOCHASTIC GRADIENTE REGRESSOR
9.02919E+16	3.73019E+41	3.53E+18	0.531923835	RLM	REGRESSOR LINEAR M

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.9 Se muestra la continuación de los resultados obtenidos por los diferentes algoritmos utilizados en base al entrenamiento de los 18.397 registros.Su característica principal que esta tabla fue creada a partir de 768 características Token(CLS) mas 768 características Token finales que es propio del modelo Base de BERT, adicionalmente se le incremento 23 características manuales(Hand Crafted Features) que fueron generadas por un conjunto de estudiantes de la Facultad de Ciencias Matemáticas y Física en la cual cada estudiante manualmente agrego según su criterio la complejidad de cada palabra y posteriormente ajustado con el modelo Fine Tuning

XML-RoBERTA

Tabla 10: XML-RoBERTA SEP + 23 (Hand Crafted Features) modelo Fine Tuning

M.A.E	M.S.E	R.M.S.E	PEARSON	ALGORITMOS	NOMBRE
0.168237123	0.04879202	0.220291218	0.272918201	SVR	SUPER VECTOR REGRESSOR
0.163029158	0.042092883	0.20621809	0.47901242	GBR	GRADIENT BOOSTING REGRESSOR
0.162007612	0.051928323	0.228729105	0.997930016	ABR	ADABOOST REGRESSOR
0.182938296	0.053702317	0.233029136	0.517328034	KNR	KNEIGHBORS REGRESSOR
0.170001425	0.043912342	0.20772301	0.983948189	RFR	RANDOM FOREST REGRESSOR
0.208301257	0.092831786	0.298902185	0.999301299	DT	DECISION TREE
0.236034933	0.102651329	0.318912035	0.112039231	PAR	PASSIVE AGGESSIVE REGRESSOR
0.17623914	0.048237197	0.213911283	0.201289387	SGR	STOCHASTIC GRADIENTE REGRESSOR
0.178392187	0.048230093	0.211023229	0.200390099	RLM	REGRESSOR LINEAR M

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.10 Se muestra los resultados obtenidos por los diferentes algoritmos utilizados en base al entrenamiento de los 18.397 registro.Su característica principal que esta tabla fue creada a partir de 768 características de inicio de la oración o del Token (SEP) que se representa porcentualmente en un valor acumulado propio del modelo XML-RoBERTA, adicionalmente se le incremento 23 características manuales (Hand Crafted Features)que fueron generadas por un conjunto de estudiantes de la Facultad de Ciencias Matemáticas y Física en la cual cada estudiante manualmente agrego según su criterio la complejidad de cada palabra y posteriormente ajustado con el modelo Fine Tuning.

Tabla 11: XML-RoBERTA Token + 23 (Hand Crafted Features) modelo Fine Tuning

M.A.E	M.S.E	R.M.S.E	PEARSON	LGORITMO	NOMBRE
0.166034819	0.048237126	0.217218391	0.308392339	SVR	SUPER VECTOR REGRESSOR
0.163829103	0.042970187	0.203481292	0.480031227	GBR	GRADIENT BOOSTING REGRESSOR
0.160132594	0.050137294	0.225192898	0.998713376	ABR	ADABOOST REGRESSOR
0.175123099	0.053014378	0.229830297	0.52912202	KNR	KNEIGHBORS REGRESSOR
0.166983232	0.042718249	0.201329001	0.984910088	RFR	RANDOM FOREST REGRESSOR
0.204029182	0.083827106	0.293932124	0.999000009	DT	DECISION TREE
0.239205426	0.106732065	0.316339291	0.148882925	PAR	PASSIVE AGGESSIVE REGRESSOR
0.172056706	0.04612307	0.213463236	0.228734889	SGR	STOCHASTIC GRADIENTE REGRESSOR
0.170439231	0.045382905	0.212291808	0.22543923	RLM	REGRESSOR LINEAR M

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.11 Se muestra los resultados obtenidos por los diferentes algoritmos utilizados en base al entrenamiento de los 18.397 registros. Su característica principal que esta tabla fue creada a partir de 768 características Token(Token-XLM) fin de la oración que se representa porcentualmente en un valor acumulado propio del modelo XML-RoBERTA, adicionalmente se le incremento 23 características manuales(Hand Crafted Features) que fueron generadas por un conjunto de estudiantes de la Facultad de Ciencias Matemáticas y Física en la cual cada estudiante manualmente agrego según su criterio la complejidad de cada palabra y posteriormente ajustado con el modelo Fine-Tuning.

Tabla 12: XML-RoBERTA SEP + Token + 23 (Hand Crafted Features) modelo Fine-Tuning

M.A.E	M.S.E	R.M.S.E	PEARSON	GORITM	NOMBRE
0.168812372	0.048270153	0.219148792	0.293218797	SVR	SUPER VECTOR REGRESSOR
0.165293843	0.043090213	0.206971429	0.493021931	GBR	GRADIENT BOOSTING REGRESSOR
0.162002916	0.052603928	0.226822321	0.998703998	ABR	ADABOOST REGRESSOR
0.17303493	0.05390382	0.230023916	0.528430646	KNR	KNEIGHBORS REGRESSOR
0.168932711	0.043802109	0.208613201	0.986012933	RFR	RANDOM FOREST REGRESSOR
0.205938236	0.08700381	0.296839286	0.999783209	DT	DECISION TREE
0.229901292	0.094239135	0.304479214	0.026839208	PAR	PASSIVE AGGESSIVE REGRESSOR
0.171445322	0.0450329	0.212837438	0.225904009	SGR	STOCHASTIC GRADIENTE REGRESSOR
0.171892016	0.045239129	0.21277022	0.225740126	RLM	REGRESSOR LINEAR M

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.12 Se muestra los resultados obtenidos por los diferentes algoritmos utilizados en base al entrenamiento de los 18.397 registros. Su característica principal que esta tabla fue creada a partir de 768 características Token (Sep) más otras 768 caracteristicas Token(Token-XLM) fin de la oración, por lo cual se representa porcentualmente en dos valores acumulados de cada uno que son propio del modelo XML-RoBERTA, adicionalmente se le incremento 23 características manuales(Hand Crafted Features) que fueron generadas por un conjunto de estudiantes de la Facultad de Ciencias Matemáticas y Física en la cual cada estudiante manualmente agrego según su criterio la complejidad de cada palabra y posteriormente ajustado con el modelo Fine-Tuning

Tabla 13: Mejores resultados Bert con Fine Tuning

M.A.E	M.S.E	R.M.S.E	PEARSON	ALGORITMOS	NOMBRE
0.159237200	0.037901903	0.198239181	0.98830122	RFR	RANDOM FOREST REGRESSOR
0.160009359	0.036702388	0.197901155	0.82023918	GBR	GRADIENT BOOSTING REGRESSOR
0.161017645	0.040182363	0.198892168	0.99876732	RFR	RANDOM FOREST REGRESSOR
0.161093889	0.050619275	0.22428396	0.9998712	ABR	ADABOOST REGRESSOR
0.162110076	0.048730012	0.230018236	0.99999899	ABR	ADABOOST REGRESSOR

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.13 Se muestra los mejores 5 resultados obtenidos de las diferentes combinaciones de entrenamiento de datos con sus respectivos algoritmos, por lo cual se evidencia que los mejores resultados obtenidos fueron a través de los algoritmos de (RFR) Random Forest Regressor con un valor de 0.159237200 de exactitud, siguiendo el algoritmo (GBR) Gradient Boosting Regressor con un valor de 0.160009359 de exactitud y por ultimo el algoritmo (ABR) Adaboost Regressor con un valor 0.161093889 valores representados en Error Absoluto Medio (M.A.E).

Tabla 14: Mejores resultados XML-RoBERTA modelo Fine Tuning

M.A.E	M.S.E	R.M.S.E	PEARSON	ALGORITMOS	NOMBRE
0.160132594	0.050137294	0.225192898	0.998713	ABR	ADABOOST REGRESSOR
0.162002916	0.052603928	0.226822321	0.998704	ABR	ADABOOST REGRESSOR
0.162007612	0.051928323	0.228729105	0.99793	ABR	ADABOOST REGRESSOR
0.163029158	0.042092883	0.20621809	0.479012	GBR	GRADIENT BOOSTING REGRESSOR
0.163829103	0.042970187	0.203481292	0.480031	GBR	GRADIENT BOOSTING REGRESSOR

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.14 Se muestra los mejores 5 resultados obtenidos de las diferentes combinaciones de entrenamiento de datos con sus respectivos algoritmos, por lo cual se evidencia que los mejores resultados obtenidos fueron a través de los algoritmos de (ABR) Adaboost Regressor con un valor de 0.160132594 de exactitud, siguiendo el algoritmo (ABR) Adaboost Regressor con un valor de 0.162002916 de exactitud y por último el algoritmo (GBR) Gradient Boosting Regressor con un valor 0.163029158 valores representados en Error Absoluto Medio (M.A.E).

Transformador Roberta-Large-BNE

Para el alcance del segundo objetivo planteado se realizaron las ejecuciones del *Transformes* en español Roberta-Large-BNE generando tablas en las que se puede visualizar los entrenamientos ejecutados con los diferentes algoritmos y combinaciones de caracteristicas, lo que permite la exploración de dichos Transformes y su comportamiento.

Para la creación de las tablas se utilizo un Corpus previamente creados con un total de 18.397 registros que fueron utilizados en su totalidad para el entrenamiento de datos.

Procediendo en la creación de los Datasets, el transformes utilizado fue ejecutado con el modelo pre-entrenado genérico de código abierto RoBERTa Large BNE presentado los siguientes resultados con sus diferentes algoritmos de Machine Learning.

Tabla 15: RoBERTA-Large-BNE Token

M.A.E	M.S.E	R.M.S.E	PEARSON	ALGORITMOS	NOMBRE
0.16869955369340825	0.046859969527455504	0.21460741191055308	0.5021343770126271	SVR	SUPER VECTOR REGRESSOR
0.16908201788569577	0.044240896454899985	0.20906663113555038	0.6585583936277026	GBR	GRADIENT BOOSTING REGRESSOR
0.1677199972276148	0.055120968586853725	0.23366248784228572	0.9952903721323743	ABR	ADABOOST REGRESSOR
0.1811959121548163	0.05378212654924983	0.23075446848823752	0.5093815384587101	KNR	KNEIGHBORS REGRESSOR
0.17319874216088707	0.04476875704900387	0.21058139065095827	0.9815281107440883	RFR	RANDOM FOREST REGRESSOR
0.2110784953250707	0.09039682539682541	0.30029903415585707	0.9972086402838096	DT	DECISION TREE
252743.8898342543	24501425068299.965	4925473.329509084	0.08959795614609696	PAR	PASSIVE AGGESSIVE REGRESSOR
8.454010826693362e+16	2.948961200402752e+36	1.5757420740773092e+18	0.010468354184863 45	SGR	STOCHASTIC GRADIENTE REGRESSOR

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.15 Se muestra los resultados obtenidos por los diferentes algoritmos utilizados en base al entrenamiento de los 18.397 registros, obteniendo el Error Absoluto Medio (M.A.E) que representa la referencia de la precisión de cada algoritmo en contraste con los datos reales, también el valor cuadrático medio (M.S.E) y el error cuadrático medio (R.M.S.E), también se obtiene el valor de Pearson que representa el coeficiente de correlación y por último se distingue los 9 algoritmos de Machine Learning utilizados. Su característica principal que esta tabla fue creada a partir de 1024 características Token de fin que es propio del modelo Roberta Large BNE.

Tabla 16: RoBERTA-Large-BNE SEP

M.A.E	M.S.E	R.M.S.E	PEARSON	ALGORITMOS	NOMBRE
0.17084372608661 21	0.05079483171309216	0.22247068719413657	0.05020876749405 0906	SVR	SUPER VECTOR REGRESSOR
0.18436869189995217	0.05224648372933648	0.22760666068673827	0.52091939463696 92	GBR	GRADIENT BOOSTING REGRESSOR
0.20032630095037596	0.06592039534499654	0.2565769000047091	0.57646151887542 52	ABR	ADABOOST REGRESSOR
0.19037616873233312	0.05871032833246146	0.24153908035694735	0.44039888544072349	KNR	KNEIGHBORS REGRESSOR
0.19332453285501 85	0.06127642528189903	0.24705936963212 34	0.64125845437229 31	RFR	RANDOM FOREST REGRESSOR
0.20872841145590 66	0.08161586682970726	0.28550696572080 76	0.65399326384778 92	DT	DECISION TREE
6886.5907017799 04	65072674534 9.2292	539251.5940406561	0.07345233291329939	PAR	PASSIVE AGGESSIVE REGRESSOR
19414348200796 6.38	6.782556070806193e+32	1.52027670716656 06e+16	-0.00579382089063 6809	SGR	STOCHASTIC GRADIENTE REGRESSOR

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.16 Se muestra los resultados obtenidos por los diferentes algoritmos utilizados en base al entrenamiento de los 18.397 registros, obteniendo el Error Absoluto Medio (M.A.E) que representa la referencia de la precisión de cada algoritmo en contraste con los datos reales, también el valor cuadrático medio (M.S.E) y el error cuadrático medio (R.M.S.E), también se obtiene el valor de Pearson que representa el coeficiente de correlación y por último se distingue los 9 algoritmos de Machine Learning utilizados. Su característica principal que esta tabla fue creada a partir de 1024 características Token de inicio (SEP) que es propio del modelo Roberta-Large-BNE.

Tabla 17: RoBERTA-Large-BNE SEP + Token

M.A.E	M.S.E	R.M.S.E	PEARSON	ALGORITMOS	NOMBRE
0.16841852607312177	0.04726694913788235	0.21529056287897833	0.44255209981392946	SVR	SUPER VECTOR REGRESSOR
0.160938040946769	0.0421359879190553	0.204729268072132	0.675490106625494	GBR	GRADIENT BOOSTING REGRESSOR
0.169905946273682	0.0598804632566218	0.239165134941683	0.989119133121844	ABR	ADABOOST REGRESSOR
0.18195042400521855	0.0543418134377038 55	0.2321604962589586	0.5092023134646442	KNR	KNEIGHBORS REGRESSOR
0.170582633709841	0.0448102059236488	0.211825808682768	0.9816954917702 18	RFR	RANDOM FOREST REGRESSOR
0.21037725592520115	0.08941454664057402	0.2985186678730904	0.9972086402838096	DT	DECISION TREE
141615.49213829418	7449870447594.413	2701675.7585053844	0.19947095577729496	PAR	PASSIVE AGGESSIVE REGRESSOR
3.3072201925176436e+	5.7240799383273246e+1	6.991424511370212e+17	-0.0011573236546878 42	SGR	STOCHASTIC GRADIENTE REGRESSOR

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.17 Se muestra los resultados obtenidos por la combinación de características, las cuales son 1024 características Token de inicio (SEP) más 1024 características Token fin de la oración de que es propio del modelo Roberta Large BNE

Tabla 18: RoBERTA-Large-BNE SEP + Token + 23(Hand Crafted Features)

M.A.E	M.S.E	R.M.S.E	PEARSON	ALGORITMOS	NOMBRE
0.16308555308402606	0.043557177553162636	0.20704193155948306	0.4883315006602684	SVR	SUPER VECTOR REGRESSOR
0.16860804094672288	0.04323588771905725	0.206729267072131 05	0.685450100625 4142	GBR	GRADIENT BOOSTING REGRESSOR
0.1689049442736826	0.05588036325652185	0.23516513294148375	0.995115132121 8442	ABR	ADABOOST REGRESSOR
0.17717982170036964	0.0512804522722331	0.2253668315321251	0.594708986548 5775	KNR	KNEIGHBORS REGRESSOR
0.1755826387092477	0.04491020992364850 4	0.21089589828270824	0.980690491730 2182	RFR	RANDOM FOREST REGRESSOR
0.21740595781691674	0.0940226136116547	0.3060424907774306	0.99720864028381	DT	DECISION TREE
0.21045717908104386	0.08063501158072876	0.28233586113885 18	0.326403204769 5949	PAR	PASSIVE AGGESSIVE REGRESSOR
1134261338.8801868	2.1968964453138 28e+18	1480304250.6064212	0.06184697235000098	SGR	STOCHASTIC GRADIENTE REGRESSOR

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.18 Se muestra los resultados obtenidos por la combinación de características, las cuales son 1024 características Token de inicio (SEP) más 1024 características

Token de fin que es propio del modelo Roberta Large BNE adicionalmente se le incremento 23 características manuales(Hand Crafted Features) que fueron generadas por un conjunto de estudiantes de la Facultad de Ciencias Matemáticas y Física en la cual cada estudiante manualmente agrego según su criterio la complejidad de cada palabra.

Modelo Ajustado Fine Tuning

Para el alcance del tercer objetivo planteado se realizaron las ejecuciones del *Transformes* en español Roberta-Large-BNE generando tablas en las que se puede visualizar los entrenamientos ejecutados con los diferentes algoritmos y combinaciones de características, lo que permite la exploración de dichos Transformes y su comportamiento.

Procediendo en la creación de los Datasets, el transformes utilizado se fue ajustado a un modelo entrenado Fine Tuning presentado los siguientes resultados con sus diferentes algoritmos de Machine Learning.

Tabla 19: RoBERTA-Large-BNE Token modelo Fine Tuning

M.A.E	M.S.E	R.M.S.E	PEARSON	ALGORITMOS	NOMBRE
0.17081253226019685	0.05078859053165541	0.2224577566577911	0.03906610088209448	SVR	SUPER VECTOR REGRESSOR
0.16969019331649968	0.04474901293440653	0.21014110701099772	0.6400213843588591	GBR	GRADIENT BOOSTING REGRESSOR
0.16934278071139 64	0.05633494065287983	0.23607226838328743	0.9948911364946255	ABR	ADABOOST REGRESSOR
0.18116329636877582	0.05400243531202436	0.23132144893864917	0.5014308173040313	KNR	KNEIGHBORS REGRESSOR
0.17304701282463694	0.04476278710502323	0.21048419811662136	0.980839500848577	RFR	RANDOM FOREST REGRESSOR
0.20987714720591433	0.08921667753859534	0.2983381632624881 3	0.9972086402838096	DT	DECISION TREE
0.30655812911614605	0.14647742066354677	0.37514587632235 26	0.08786181403812 26	PAR	PASSIVE AGGESSIVE REGRESSOR
10598475173.205187	2.932996389768112 7e+2	16651912076.740622	0.03764748576577050	SGR	STOCHASTIC GRADIENTE REGRESSOR

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.19 Se muestra los resultados obtenidos por los diferentes algoritmos utilizados en base al entrenamiento de los 18.397 registros, obteniendo el Error Absoluto Medio (M.A.E) que representa la referencia de la precisión de cada algoritmo en contraste con los datos reales, también el valor cuadrático medio (M.S.E) y el error cuadrático medio (R.M.S.E), también se obtiene el valor de Pearson que representa el coeficiente de correlación y por último se distingue los 9 algoritmos de Machine Learning utilizados. Su característica principal que esta tabla fue creada a partir de 1024 características Token de fin que es propio del modelo Roberta Large BNE y posteriormente ajustado con el modelo Fine Tuning.

Tabla 20: RoBERTA-Large-BNE SEP modelo Fine Tuning

M.A.E	M.S.E	R.M.S.E	PEARSON	ALGORITMOS	NOMBRE
0.17083091185101076	0.050796576941388304	0.22247358858082814	0.005245068572855 48	SVR	SUPER VECTOR REGRESSOR
0.1829192369415027	0.051164966838342964	0.22514703365505503	0.464662697854058	GBR	GRADIENT BOOSTING REGRESSOR
0.20026518426885284	0.066196477507223	0.2571170639905019	0.5807382631166733	ABR	ADABOOST REGRESSOR
0.18945857795172863	0.05824483583387693	0.24054895056269512	0.4465705049355717	KNR	KNEIGHBORS REGRESSOR
0.1988245323550185	0.06347648528489903	0.2476593686325234	0.6814584553722931	RFR	RANDOM FOREST REGRESSOR
0.2089718434304214	0.08187316914420088	0.2860010332050205	0.6539392638477887	DT	DECISION TREE
0.3104297740884098	0.16603474838608664	0.39676221556216656	0.07088597440603926	PAR	PASSIVE AGGESSIVE REGRESSOR
34242472588.902386	3.7634014889802135e+2	40654570672.74633	0.022631737759589 71	SGR	STOCHASTIC GRADIENTE REGRESSOR

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.20 Se muestra los resultados obtenidos por los diferentes algoritmos utilizados en base al entrenamiento de los 18.397 registros, obteniendo el Error Absoluto Medio (M.A.E) que representa la referencia de la precisión de cada algoritmo en contraste con los datos

reales, también el valor cuadrático medio (M.S.E) y el error cuadrático medio (R.M.S.E), también se obtiene el valor de Pearson que representa el coeficiente de correlación y por último se distingue los 9 algoritmos de Machine Learning utilizados. Su característica principal que esta tabla fue creada a partir de 1024 características Token de inicio (SEP) que es propio del modelo RoBERTA Large BNE y posteriormente ajustado con el modelo Fine Tuning.

Tabla 21: RoBERTA-Large_BNE SEP + Token modelo Fine Tuning

M.A.E	M.S.E	R.M.S.E	PEARSON	ALGORITMOS	NOMBRE
0.170846810239758	0.050785898904414 4	0.222448507636258 79	0.076411056929467 5	SVR	SUPER VECTOR REGRESSOR
0.169805852185796 55	0.044783157291380 07	0.210247175152293 1	0.645070549763110 7	GBR	GRADIENT BOOSTING REGRESSOR
0.167568529797885 58	0.055661741350336 34	0.234728815109242 07	0.995265927884492 3	ABR	ADABOOST REGRESSOR
0.186333985649054 15	0.056496368775820 836	0.236832081684933 4	0.476080846990538 15	KNR	KNEIGHBORS REGRESSOR
0.172785659711253 3	0.044716289083878 68	0.210366647877885 64	0.981511078617922 5	RFR	RANDOM FOREST REGRESSOR
0.210659926070885 05	0.089335181561208 97	0.298513718932670 37	0.997208640283809 6	DT	DECISION TREE
0.229694647485107 4	0.096123327313511 68	0.308440027515030 9	0.142875934869409 22	PAR	PASSIVE AGGESSIVE REGRESSOR
8558912674 2.24828	2.25378935636273 8e+22	140974120182.71094	-0.022097410353941 2	SGR	STOCHASTIC GRADIENTE REGRESSOR

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.21 Se muestra los resultados obtenidos por la combinación de características, las cuales son 1024 características Token de inicio (SEP) más 1024 características Token fin de la oración de que es propio del modelo RoBERTA Large BNE y posteriormente ajustado con el modelo Fine Tuning.

Tabla 22: RoBERTA-Large-BNE SEP + Token + 23 (Hand Crafted Features) modelo Fine Tuning

M.A.E	M.S.E	R.M.S.E	PEARSON	ALGORITMOS	NOMBRE
0.17080084775760893	0.05079357033353304	0.22246643872034758	0.23634454488889534	SVR	SUPER V ECTOR REGRESSOR
0.16916799191835538	0.043489261818887216	0.20733659253539 44	0.6607873136850142	GBR	GRADIENT BOOSTING REGRESSOR
0.16707162008543186	0.05480160039066 7234	0.23281046274146186	0.9949675040754256	ABR	ADABOOST REGRESSOR
0.18463579038921507	0.05587753859534681	0.23565575479954 16	0.5096177810168734	KNR	KNEIGHBORS REGRESSOR
0.17499120235196294	0.04473809313022503	0.21044673389053858	0.9799498075633986	RFR	RANDOM FOREST REGRESSOR
0.21299739073711677	0.0903886714503153	0.2998186643514499	0.9972086402838088	DT	DECISION TREE
0.2470766717397 6192	0.11051361289035853	0.32867922951743 46	0.1738764927316585	PAR	PASSIVE AGGESSIVE REGRESSOR
137195714065.10222	5.4038281519386 63e+22	222544257711.18887	0.006982145556457829	SGR	STOCHASTIC GRADIENTE REGRESSOR

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.22 Se muestra los resultados obtenidos por la combinación de características, las cuales son 1024 características Token de inicio (SEP) más 1024 características Token de fin que es propio del modelo RoBERTA-Large-BNE adicionalmente se le incremento 23 características manuales(Hand Crafted Features) que fueron generadas por un conjunto de estudiantes de la Facultad de Ciencias Matemáticas y Física en la cual cada estudiante manualmente agrego según su criterio la complejidad de cada palabra. y posteriormente ajustado con el modelo Fine Tuning

Metodología de desarrollo del Prototipo

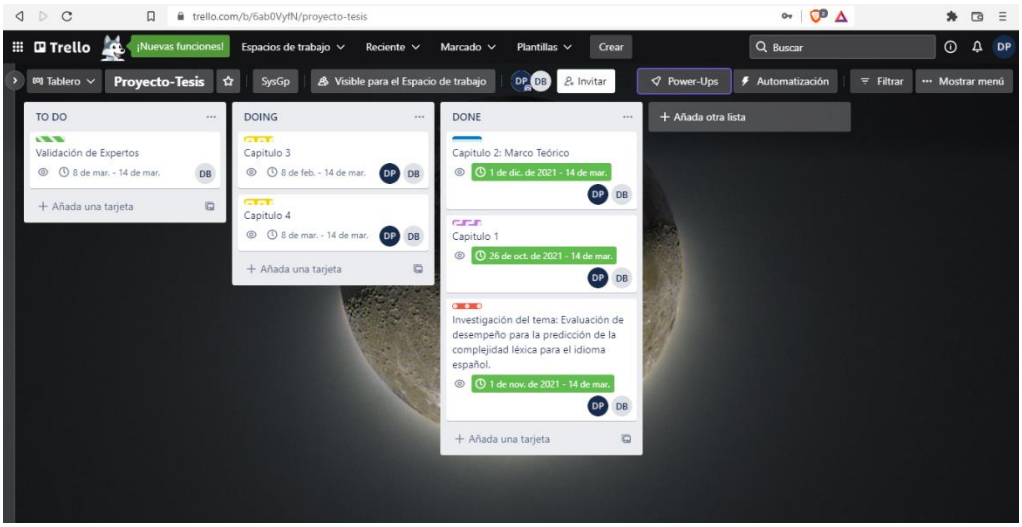
Metodología Kanban

Esta metodología es muy sencilla, se puede actualizar y los trabajos pueden admitir sin problema. Al ser un método visual permite que con una simple observación se conozca el estado de los proyectos y puedan realizarse tareas nuevas de manera simple. Para aplicarlo, es necesario un contador de tareas con el cual se logre mejorar el trabajo. (Castellano Lendínez, 2019)

Existen 3 procesos de metodología:

- To Do: Agrupar las actividades que se realizaran para la elaboración del proyecto.
- Doing: Actividades que se están realizando.
- Done: Actividades realizadas en el transcurso del proyecto.

Ilustración 13: Metodología Kanban



Nota: El proceso de la metodología Kanban que se realizó durante el transcurso de la investigación.

Beneficiarios directos e indirectos del proyecto

Beneficiario directo: Son beneficiarios directos los alumnos de la carrera de ingeniería en sistemas computacionales y software de la Universidad Estatal de Guayaquil.

Beneficiario indirecto: Son beneficiarios indirectos los profesores de la materia de literatura y lengua extranjera en el cual ayudará a explicar de una manera sencilla a los alumnos la estructura de una palabra u oración dependiendo la complejidad

Entregables del proyecto

• Documentación en Word.

• Artículo científico.

• Fuente del prototipo.

• Manual de usuario y técnico.

Propuesta

En este proyecto de investigación se presenta una evaluación de desempeño de transformadores para la predicción de la complejidad léxica para el idioma español que se encuentra en los corpus en el idioma español que está compuesto por varios textos académicos de la carrera de Ingeniería en Sistemas Computacionales y de Software, Facultad de Ciencias Matemáticas y Física de la Universidad Estatal de Guayaquil.

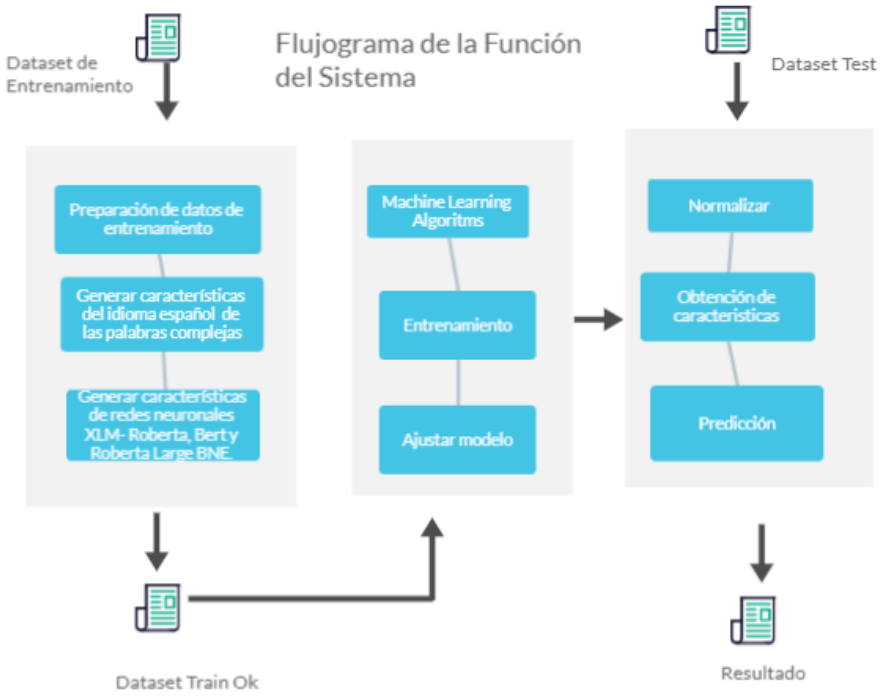
Se calcularon las características de las palabras y otras características resultantes de la ejecución de técnicas de redes neurales BERT y XLM-RoBERTa para la generación de características actualizadas con el propósito de lograr un mejor entrenamiento que el algoritmo de Machine Learning.

Flujograma de la Aplicación

La ilustración Nº 14 nos permite observar el flujograma con la que trabaja el sistema propuesto e interacción entre ellos. Este sistema obtiene un documento con un formato especifico a los cuales se extrae propiedades del lenguaje español a cada palabra, las redes neuronales de BERT ayudan a sacar propiedades de las palabras en su entorno para el lenguaje español y la red

neuronal XLM-RoBERTa que nos ayuda a la obtención de las propiedades en el lenguaje español habiendo extraído estas propiedades se las envía al algoritmo de Random Forest Regressor a hacer el entrenamiento de la predicción train.

Ilustración 14: Flujo del Sistema



Corpus

Para el entrenamiento del algoritmo en el lenguaje español, profesores de la Universidad de Guayaquil dieron el material de las grabaciones de las clases online de donde los alumnos de las respectivas de las carreras de Ingeniería en sistemas Computacionales y De Software etiquetaron generando el estudio que corresponde se obtuvo un corpus de 18.397 registros, con esta información obtenida de los cursos de las carreras mencionadas.

El corpus se conforma por los siguientes campos:

Corpus: Origen de la oración.

Id: Identificador único

Token: Palabra Etiquetada

Sentence: Oración

El proceso de la dificultad o Complexity se fundamenta en la escala a partir de cero (o) al uno (1) donde cero sugiere un grado de dificultad “muy sencillo” y uno “muy complejo”.

Tabla 23: Corpus

id	corpus	sentence	token	complexity
134451	Software_2doSem_CicloI_1erParcial_Contabilidad_texto5.txt	texto 5 de ahí nosotros tenemos lo c	mercado	0.2
134452	Software_3erSem_CicloI_1erParcial_Proceso_de_Software_Texto3	Dividir el trabajo en paquetes poco	acoplados	0.2
134453	Sistemas_6toSem_CicloI_1erParcial_legislacion_informatica_texto	texto 74 ¿Cuáles son las clases de su	pasivo	0.2
134454	Software_3erSem_CicloI_1erParcial_Sistemas_Operativos_Texto28	texto 28 Hilos Un hilo en un sistema	hilos	0.2
134455	Sistemas_5toSem_CicloI_1erParcial_Bases_de_Datos_texto46.txt	texto 46 Una transacción es una prop	propagación	0.2
134458	Sistemas_5toSem_CicloI_1erParcial_Ing_Soft_O_O_texto83.txt	TEXTO 83 En los diagramas de secue	booch	0.8
134460	Software_1erSem_CicloI_1erParcial_Democracia_texto18.txt	texto 18 Ahora la soberanía de los es	jurisdicción	0.6
134461	Sistemas_6toSem_microprocesadores_texto98.txt	En este artículo mencionaremos tres	privilegiadas	0.4
134463	Sistemas_7moSem_CicloI_1erParcial_Seguridad_Informatica_texto	texto 63 No descargar desde paginas	descargar	0.2
134464	Software_3erSem_CicloI_1erParcial_Proceso_de_Software_Texto2	texto 25 Modelado de gestión Mode	generación	0.2

Generar las características de corpus

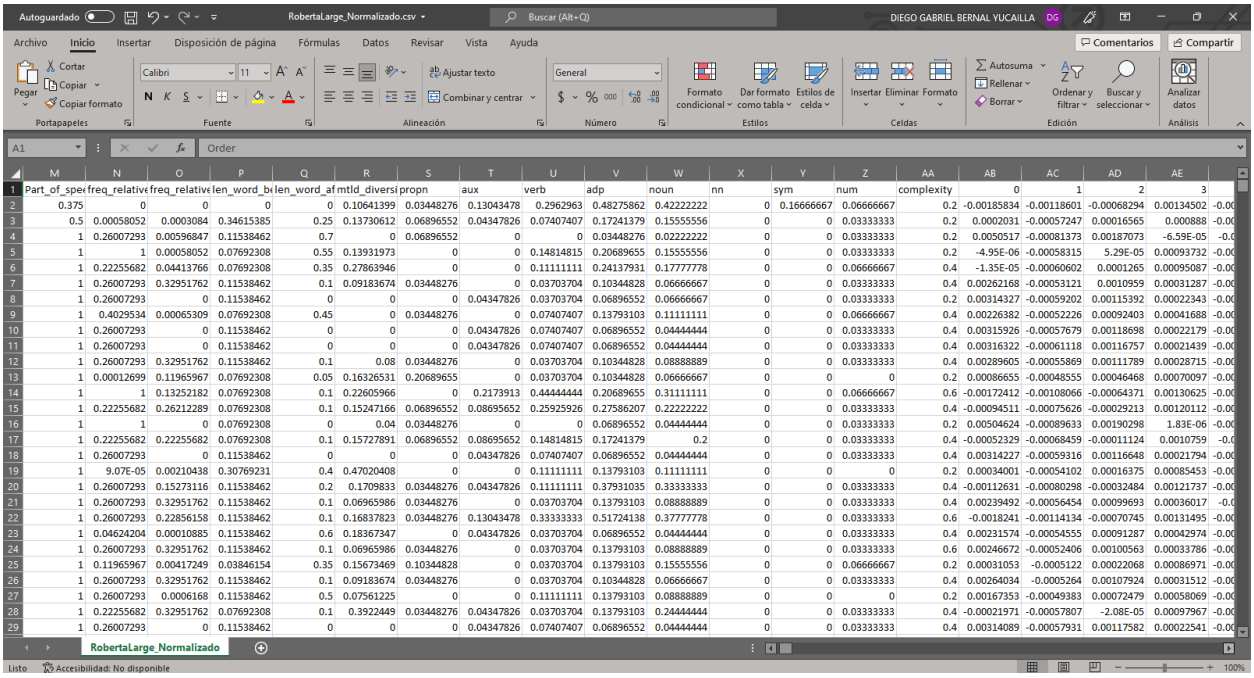
La primera parte de esta investigación se obtiene 23 características para el corpus español que se incluye una característica que se obtuvo de la ejecución de la red neuronal XLM-RoBERTa.

Entrenamiento

Para el alcance del segundo y tercer objetivo planteado se realizaron las ejecuciones del *Transformes* en español BERT, XML-RoBERTA y RoBERTA-Large-BNE generando las incrustaciones numericas, lo que nos permite evaluar el rendimiento del modelo de transformador genérico BERT, XML-RoBERTa y RoBERTa-Large-BNE para posteriormente aplicar el ajuste fino mediante la ejecución de la técnica Fine-Tuning para su ejecución sobre los modelos previamente pre-entrenados y la generación de los Dataset basados en las nuevas represntaciones numéricas.

Se elaboraron Datasets en el corpus español ClexIS^{2}

Ilustración 15: DataSets RoBERTa-Large-BNE modelo Pre-Entrenado

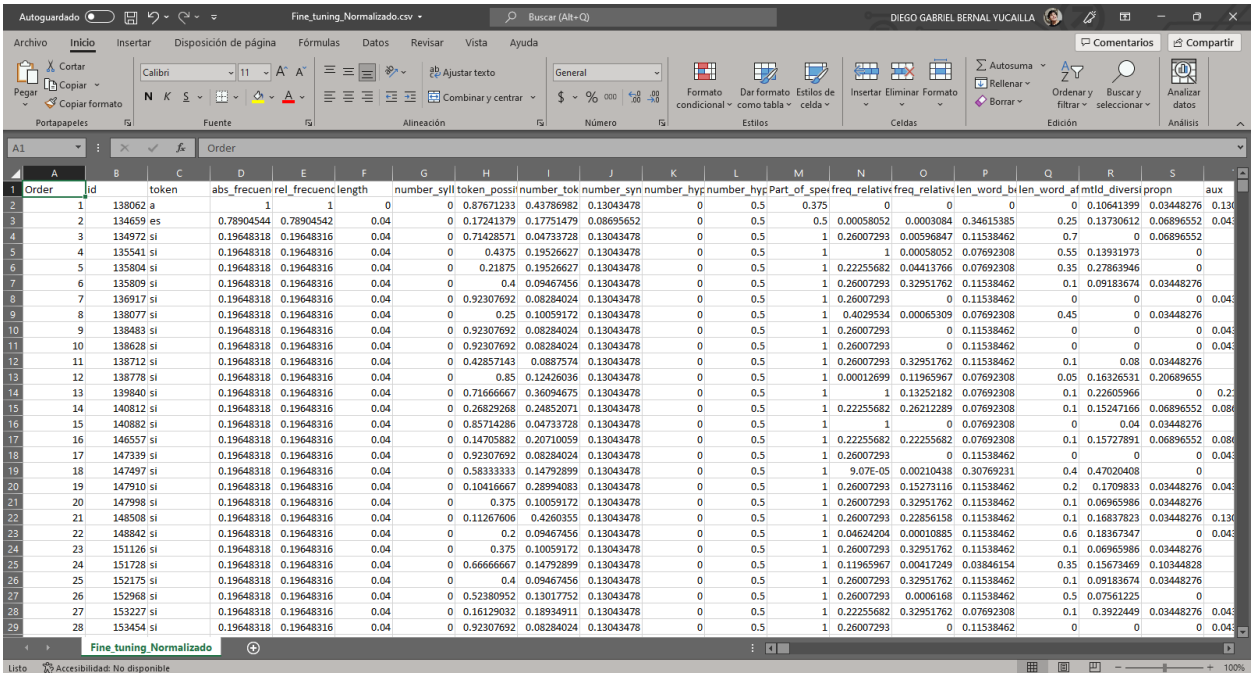


	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1	Part_of_spe	freq_relative	freq_relative	len_word_b	len_word_af	mtld_diversi	propn	aux	verb	adp	noun	nn	sym	num	complexity	0	1	2	3
2	0.375	0	0	0	0	0.10641399	0.03448276	0.13043478	0.2962963	0.48275862	0.42222222	0	0.16666667	0.06666667	0.2	-0.00185834	-0.00118601	-0.00068294	0.00134502
3	0.5	0.00058052	0.0003084	0.34615385	0.25	0.13730612	0.06896552	0.04347826	0.07407407	0.17241379	0.15555556	0	0	0.03333333	0.2	0.0002031	-0.00057247	0.00016565	0.000888
4	1	0.26007293	0.00596847	0.11538462	0.7	0	0.06896552	0	0	0.03448276	0.02222222	0	0	0.03333333	0.2	0.0050517	-0.00081373	0.00187073	-6.59E-05
5	1	1	0.00058052	0.07692308	0.55	0.13931973	0	0	0.14814815	0.20689655	0.15555556	0	0	0.03333333	0.2	-4.95E-06	-0.00058315	5.29E-05	0.00093732
6	1	0.22255682	0.04413766	0.07692308	0.35	0.27863946	0	0	0.11111111	0.24137931	0.17777778	0	0	0.06666667	0.4	-1.35E-05	-0.00060602	0.0001265	0.00095087
7	1	0.26007293	0.32951762	0.11538462	0.1	0.09183674	0.03448276	0	0.03703704	0.10344828	0.06666667	0	0	0.03333333	0.4	0.00262168	-0.00053121	0.0010959	0.00031287
8	1	0.26007293	0	0.11538462	0	0	0	0.04347826	0.03703704	0.06896552	0.06666667	0	0	0.03333333	0.2	0.00314327	-0.00059202	0.00115392	0.00022343
9	1	0.4029534	0.00065309	0.07692308	0.45	0	0.03448276	0	0.07407407	0.13793103	0.11111111	0	0	0.06666667	0.4	0.00226382	-0.00052226	0.00092403	0.00041688
10	1	0.26007293	0	0.11538462	0	0	0	0.04347826	0.07407407	0.06896552	0.04444444	0	0	0.03333333	0.4	0.00315926	-0.00057679	0.00118698	0.00022179
11	1	0.26007293	0	0.11538462	0	0	0	0.04347826	0.07407407	0.06896552	0.04444444	0	0	0.03333333	0.4	0.00316322	-0.00061118	0.00116757	0.00021439
12	1	0.26007293	0.32951762	0.11538462	0.1	0.08	0.03448276	0	0.03703704	0.10344828	0.08888889	0	0	0.03333333	0.4	0.00289605	-0.00055869	0.00111789	0.00028715
13	1	0.00012699	0.11965967	0.07692308	0.05	0.16326531	0.20689655	0	0.03703704	0.10344828	0.06666667	0	0	0	0.2	0.00086655	-0.00048555	0.00046468	0.00070097
14	1	1	0.13252182	0.07692308	0.1	0.22605966	0	0.2173913	0.44444444	0.20689655	0.31111111	0	0	0.06666667	0.6	-0.00172412	-0.00108066	-0.00064371	0.00130625
15	1	0.22255682	0.26212289	0.07692308	0.1	0.15247166	0.06896552	0.08695652	0.25925926	0.27586207	0.22222222	0	0	0.03333333	0.4	-0.00094511	-0.00075626	-0.00029213	0.00120112
16	1	1	0	0.07692308	0	0.04	0.03448276	0	0	0.06896552	0.04444444	0	0	0.03333333	0.2	0.00504624	-0.00089633	0.00190298	1.83E-06
17	1	0.22255682	0.22255682	0.07692308	0.1	0.15727891	0.06896552	0.08695652	0.14814815	0.17241379	0.2	0	0	0.03333333	0.4	-0.00052329	-0.00068459	-0.00011124	0.0010759
18	1	0.26007293	0	0.11538462	0	0	0	0.04347826	0.07407407	0.06896552	0.04444444	0	0	0.03333333	0.4	0.00314227	-0.00059316	0.00116648	0.00021794
19	1	9.07E-05	0.00210438	0.30769231	0.4	0.47020408	0	0	0.11111111	0.13793103	0.11111111	0	0	0	0.2	0.00034001	-0.00054102	0.00016375	0.00085453
20	1	0.26007293	0.15273116	0.11538462	0.2	0.1709833	0.03448276	0.04347826	0.11111111	0.37931035	0.33333333	0	0	0.03333333	0.4	-0.00112631	-0.00080298	-0.00032484	0.00121737
21	1	0.26007293	0.32951762	0.11538462	0.1	0.06965986	0.03448276	0	0.03703704	0.13793103	0.08888889	0	0	0.03333333	0.4	0.00239492	-0.00056454	0.00099693	0.00036017
22	1	0.26007293	0.22856158	0.11538462	0.1	0.16837823	0.03448276	0.13043478	0.33333333	0.51724138	0.37777778	0	0	0.03333333	0.6	-0.0018241	-0.00114134	-0.00070745	0.00131495
23	1	0.04624204	0.00010885	0.11538462	0.6	0.18367347	0	0.04347826	0.03703704	0.06896552	0.04444444	0	0	0.03333333	0.4	0.00231574	-0.00054555	0.00091287	0.00042974
24	1	0.26007293	0.32951762	0.11538462	0.1	0.06965986	0.03448276	0	0.03703704	0.13793103	0.08888889	0	0	0.03333333	0.6	0.00246672	-0.00052406	0.00100563	0.00033786
25	1	0.11965967	0.00417249	0.03846154	0.35	0.15673469	0.10344828	0	0.03703704	0.13793103	0.15555556	0	0	0.06666667	0.2	0.00031053	-0.0005122	0.00022068	0.00086971
26	1	0.26007293	0.32951762	0.11538462	0.1	0.09183674	0.03448276	0	0.03703704	0.10344828	0.06666667	0	0	0.03333333	0.4	0.00264034	-0.0005264	0.00107924	0.00031512
27	1	0.26007293	0.0006168	0.11538462	0.5	0.07561225	0	0	0.11111111	0.13793103	0.08888889	0	0	0	0.2	0.00167353	-0.00049383	0.00072479	0.00058069
28	1	0.22255682	0.32951762	0.07692308	0.1	0.3922449	0.03448276	0.04347826	0.03703704	0.13793103	0.24444444	0	0	0.03333333	0.4	-0.00021971	-0.00057807	-2.08E-05	0.00097967
29	1	0.26007293	0	0.11538462	0	0	0	0.04347826	0.07407407	0.06896552	0.04444444	0	0	0.03333333	0.4	0.00314089	-0.00057931	0.00117582	0.00022541

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la imagen No.15 Se puede observar el DataSets generado del Transformers RoBERTa-Large-BNE a través del modelo pre-entrenado propio, en las cuales se puede observar las 23 unidades manuales(Hand Crafted Features) y el conjunto de características generados 1024 SEP y 1024 Tokens

Ilustración 16: DataSets RoBERTa-Large-BNE modelo ajustado Fine-Tuning



Order	id	token	abs_frecuen	rel_frecuenc	length	number_syll	token_possi	number_tok	number_syn	number_hyp	number_hyp	Part_of_spe	freq_relative	freq_relative	len_word_b	len_word_af	mtld_diversi	propn	aux
1	138062	a	1	1	0	0	0.87671233	0.43786982	0.13043478	0	0.5	0.375	0	0	0	0	0.10641399	0.03448276	0.130
2	134659	es	0.78904544	0.78904542	0.04	0	0.17241379	0.17751479	0.08695652	0	0.5	0.5	0.00058052	0.0003084	0.34615385	0.25	0.13730612	0.06896552	0.043
3	134972	si	0.19648318	0.19648316	0.04	0	0.71428571	0.04733728	0.13043478	0	0.5	1	0.26007293	0.00596847	0.11538462	0.7	0	0.06896552	
4	135541	si	0.19648318	0.19648316	0.04	0	0.4375	0.19526627	0.13043478	0	0.5	1	1	0.00058052	0.07692308	0.55	0.13931973	0	
5	135804	si	0.19648318	0.19648316	0.04	0	0.21875	0.19526627	0.13043478	0	0.5	1	0.22255682	0.04413766	0.07692308	0.35	0.27863946	0	
6	135809	si	0.19648318	0.19648316	0.04	0	0.4	0.09467456	0.13043478	0	0.5	1	0.26007293	0.32951762	0.11538462	0.1	0.09183674	0.03448276	
7	136917	si	0.19648318	0.19648316	0.04	0	0.92307692	0.08284024	0.13043478	0	0.5	1	0.26007293	0	0.11538462	0	0	0	0.043
8	138077	si	0.19648318	0.19648316	0.04	0	0.25	0.10059172	0.13043478	0	0.5	1	0.4029534	0.00065309	0.07692308	0.45	0	0.03448276	
9	138483	si	0.19648318	0.19648316	0.04	0	0.92307692	0.08284024	0.13043478	0	0.5	1	0.26007293	0	0.11538462	0	0	0	0.043
10	138628	si	0.19648318	0.19648316	0.04	0	0.92307692	0.08284024	0.13043478	0	0.5	1	0.26007293	0	0.11538462	0	0	0	0.043
11	138712	si	0.19648318	0.19648316	0.04	0	0.42857143	0.0887574	0.13043478	0	0.5	1	0.26007293	0.32951762	0.11538462	0.1	0.08	0.03448276	
12	138778	si	0.19648318	0.19648316	0.04	0	0.85	0.12426036	0.13043478	0	0.5	1	0.00012699	0.11965967	0.07692308	0.05	0.16326531	0.20689655	
13	139840	si	0.19648318	0.19648316	0.04	0	0.71666667	0.36094675	0.13043478	0	0.5	1	1	0.13252182	0.07692308	0.1	0.22605966	0	0.21
14	140812	si	0.19648318	0.19648316	0.04	0	0.26829268	0.24852071	0.13043478	0	0.5	1	0.22255682	0.26212289	0.07692308	0.1	0.15247166	0.06896552	0.086
15	140882	si	0.19648318	0.19648316	0.04	0	0.85714286	0.04733728	0.13043478	0	0.5	1	1	0	0.07692308	0	0.04	0.03448276	
16	146557	si	0.19648318	0.19648316	0.04	0	0.14705882	0.20710059	0.13043478	0	0.5	1	0.22255682	0.22255682	0.07692308	0.1	0.15727891	0.06896552	0.086
17	147339	si	0.19648318	0.19648316	0.04	0	0.92307692	0.08284024	0.13043478	0	0.5	1	0.26007293	0	0.11538462	0	0	0	0.043
18	147497	si	0.19648318	0.19648316	0.04	0	0.58333333	0.14792899	0.13043478	0	0.5	1	9.07E-05	0.00210438	0.30769231	0.4	0.47020408	0	
19	147910	si	0.19648318	0.19648316	0.04	0	0.10416667	0.28994083	0.13043478	0	0.5	1	0.26007293	0.15273116	0.11538462	0.2	0.1709833	0.03448276	0.043
20	147998	si	0.19648318	0.19648316	0.04	0	0.375	0.10059172	0.13043478	0	0.5	1	0.26007293	0.32951762	0.11538462	0.1	0.06965986	0.03448276	
21	148508	si	0.19648318	0.19648316	0.04	0	0.11267606	0.4260355	0.13043478	0	0.5	1	0.26007293	0.22856158	0.11538462	0.1	0.16837823	0.03448276	0.130
22	148842	si	0.19648318	0.19648316	0.04	0	0.2	0.09467456	0.13043478	0	0.5	1	0.04624204	0.00010885	0.11538462	0.6	0.18367347	0	0.043
23	151126	si	0.19648318	0.19648316	0.04	0	0.375	0.10059172	0.13043478	0	0.5	1	0.26007293	0.32951762	0.11538462	0.1	0.06965986	0.03448276	
24	151728	si	0.19648318	0.19648316	0.04	0	0.66666667	0.14792899	0.13043478	0	0.5	1	0.11965967	0.00417249	0.03846154	0.35	0.15673469	0.10344828	
25	152175	si	0.19648318	0.19648316	0.04	0	0.4	0.09467456	0.13043478	0	0.5	1	0.26007293	0.32951762	0.11538462	0.1	0.09183674	0.03448276	
26	152968	si	0.19648318	0.19648316	0.04	0	0.52380952	0.13017752	0.13043478	0	0.5	1	0.26007293	0.0006168	0.11538462	0.5	0.07561225	0	
27	153227	si	0.19648318	0.19648316	0.04	0	0.16129032	0.18934911	0.13043478	0	0.5	1	0.22255682	0.32951762	0.07692308	0.1	0.3922449	0.03448276	0.043
28	153454	si	0.19648318	0.19648316	0.04	0	0.92307692	0.08284024	0.13043478	0	0.5	1	0.26007293	0	0.11538462	0	0	0	0.043

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la imagen No.16 se puede observar el DataSets generado del Transformers RoBERTa-Large-BNE a través del modelo ajustado Fine-Tuning propio, en las cuales se puede observar las 23 unidades manuales(Hand Crafted Features) y el conjunto de características generados 1024 SEP y 1024 Tokens

Criterios de validación de la propuesta

Juicios de expertos

Los profesionales que nos colaboraron en la validación para verificar la confiabilidad de la investigación son expertos con extensa trayectoria en la zona de Gramática y Zona de desarrollo.

Por lo antes mencionado, la primera practica de validación tiene como propósito la evaluación y certificación del correcto funcionamiento de la evaluación de desempeño de transformadores para la predicción de la complejidad léxica para el idioma español de la carrera de Ingeniería en Sistemas Computacionales y Software de la Universidad de Guayaquil, por lo que su juicio certifica el correcto funcionamiento del sistema.

Los detalles a continuación:

Ing. González Silva Demetrio Oswaldo - Ingeniero en Sistemas Computacionales.

Ing. Arroba Salinas Sandra Yanel. - Ingeniera en Sistemas Computacionales.

Lic. Yépez Paladines Juan Carlos – Licenciado en Sistema de Información

Las validaciones se han realizado vía conferencia por medio de la utilización de la herramienta Teams concedida por la Universidad de Guayaquil, por medio de la cual se sometió a los 3 validadores a una herramienta para la respectiva consulta que sirva para hacer la validación relacionada en cada especialización por medio del metodo de la observación y formulación de cuestiones. (Véase anexo7)

Análisis de datos

Según los datos adquiridos por los juicios de profesionales, con interacción a los criterios propuestos, evidenciaron que los recursos seleccionados han permitido validar y detallar la construcción del corpus de textos accediendo tener un contenido importante para investigaciones futuras.

Los resultados de los profesionales informáticos establecieron los criterios de la función de iniciativa, autenticación, estabilidad, diseño, métricas, reutilización, portabilidad, precisaron que

los ítems han permitido abordar el sistema de anotado han ayudado de una manera positiva al área de la identificación de palabras complejas declarando que cumplen con el objetivo del plan postulado adicional.

Las métricas que se utilizaron en los pre-entrenamientos son las siguientes:

Error Absoluto Medio (M.A.E): Es utilizado principalmente en evaluación de regresión de tareas, ya que nos determina que tan cerca están las etiquetas pronosticadas de las etiquetas doradas para nuestra tarea.

Error Cuadrático Medio (M.S.E): La diferencia en el cálculo del M.S.E y M.A.E es muy poco.

Raiz Error Cuadrático Medio (R.M.S.E): Es la relacion de la varianza de las etiquetas originales capturadas por las etiquetas pronosticadas.

Resultados

Para el alcance del cuarto y quinto objetivo planteado se realizaron las ejecuciones del *Transformes* en español BERT, XML-RoBERTA y RoBERTA-Large-BNE generando tablas en las que se puede visualizar los entrenamientos ejecutados con los diferentes algoritmos y combinaciones de características, lo que nos permite evaluar el rendimiento del modelo de transformador genérico RoBERTa Large BNE versus los modelos de *Transformers* BERT y XLM-RoBERTa.

A continuación, se presenta las comparaciones de los Transformers RoBERTA-Large-BNE vs BERT y XML-RoBERTA a través de sus mejores resultados aplicando el algoritmo pre-entrenado

genérico de código abierto de cada *Transformers* y posteriormente con el modelo ajustado Fine Tuning.

Tabla 24: RoBERTA-Large-BNE vs BERT sin Fine Tuning

M.A.E	M.S.E	R.M.S.E	PEARSON	M.A.E	M.S.E	R.M.S.E	PEARSON	Diferencia	Porcentual
0.1671	0.0548	0.2328	0.9950	0.1632	0.0502	0.2343	0.9999	0.0038	2.35%
0.1676	0.0557	0.2347	0.9953	0.1634	0.0433	0.2074	0.9023	0.0041	2.54%
0.1692	0.0435	0.2073	0.6608	0.1644	0.0512	0.2333	0.9947	0.0048	2.93%
0.1693	0.0563	0.2361	0.9949	0.1654	0.0495	0.0448	0.6977	0.0039	2.39%
0.1697	0.0447	0.2101	0.6400	0.1656	0.0455	0.2088	0.704	0.0041	2.48%

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.24 Se puede observar los mejores 5 valores de cada Transformes en la que se puede observar que los valores de (M.A.E) de RoBERTA-Large-BNE de 0.1671 siendo su mejor resultado de las diferentes combinaciones y algoritmos en comparación de BERT cuyo valor es de 0.1632 de precisión, es decir una diferencia de 0.0038 siendo un 2.35% de margen de predicción, lo que demuestra que el modelo Transformers BERT aplicado con su modelo pre-entrenado genérico de código abierto tiene una mayor exactitud de predicción en comparación a su rival.

Tabla 25: RoBERTA-Large-BNE VS BERT Con Fine Tuning

M.A.E	M.S.E	R.M.S.E	PEARSON	M.A.E	M.S.E	R.M.S.E	PEARSON	Diferencia	Porcentual
0.1609	0.04213599	0.20472927	0.67549011	0.1592	0.0379019	0.19823918	0.98830122	0.0017	1.07%
0.1631	0.04355718	0.20704193	0.4883315	0.1600	0.03670239	0.19790116	0.82023918	0.0031	1.92%
0.1667	0.04661206	0.21369143	0.4220059	0.1610	0.04018236	0.19889217	0.99876732	0.0057	3.52%
0.1677	0.05512097	0.23366249	0.99529037	0.1611	0.05061927	0.22428396	0.9998712	0.0066	4.11%
0.1684	0.04726695	0.21529056	0.4425521	0.1621	0.04873001	0.23001824	0.99999899	0.0063	3.89%

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.25 Se puede observar los mejores 5 valores de cada Transformes en la que se puede observar que los valores de (M.A.E) de RoBERTA-Large-BNE de 0.1609 siendo su mejor resultado de las diferentes combinaciones y algoritmos en comparación de BERT cuyo valor es de 0.1592 de precisión, es decir una diferencia de 0.0017 siendo un 1.07% de margen de predicción, lo que demuestra que el modelo Transformers BERT aplicado con el modelo ajustado Fine Tuning tiene una mayor exactitud de predicción en comparación a su rival.

Tabla 26: Roberta vs XML Roberta sin Fine Tuning

M.A.E	M.S.E	R.M.S.E	PEARSON	M.A.E	M.S.E	R.M.S.E	PEARSON	Diferencia	Porcentual
0.1671	0.0548	0.2328	0.9950	0.1623	0.0528	0.2270	0.99734	0.0047	2.92%
0.1676	0.0557	0.2347	0.9953	0.1623	0.0513	0.2274	0.99735	0.0052	3.21%
0.1692	0.0435	0.2073	0.6608	0.1630	0.0525	0.2293	0.99735	0.0062	3.78%
0.1693	0.0563	0.2361	0.9949	0.1654	0.0434	0.2073	0.48488	0.0039	2.38%
0.1697	0.0447	0.2101	0.6400	0.1659	0.0435	0.2074	0.48745	0.0038	2.30%

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.26 Se puede observar los mejores 5 valores de cada Transformes en la que se puede observar que los valores de (M.A.E) de RoBERTA-Large-BNE de 0.1671 siendo su mejor resultado de las diferentes combinaciones y algoritmos en comparación de XML-RoBERTA cuyo valor es de 0.1623 de precisión, es decir una diferencia de 0.0047 siendo un 2.92% de margen de predicción, lo que demuestra que el modelo Transformers XML-RoBERTA aplicado con su modelo pre-entrenado genérico de código abierto tiene una mayor exactitud de predicción en comparación a su rival.

Tabla 27: Roberta vs XML Roberta Con Fine Tuning

M.A.E	M.S.E	R.M.S.E	PEARSON	M.A.E	M.S.E	R.M.S.E	PEARSON	Diferencia	Porcentual
0.1609	59879190553	29268072132	90106625494	0.16013259	0.05013729	0.2251929	0.99871338	0.0008	0.50%
0.1631	77553162636	93155948306	15006602684	0.16200292	0.05260393	0.22682232	0.998704	0.0011	0.67%
0.1667	05929503764	43173083813	59047865802	0.16200761	0.05192832	0.22872911	0.99793002	0.0047	2.88%
0.1677	68586853725	48784228572	03721323743	0.16302916	0.04209288	0.20621809	0.47901242	0.0047	2.88%
0.1684	94913788235	56287897833	09981392946	0.1638291	0.04297019	0.20348129	0.48003123	0.0046	2.80%

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.27 Se puede observar los mejores 5 valores de cada Transformes en la que se puede observar que los valores de (M.A.E) de RoBERTA-Large-BNE de 0.1609 siendo su mejor resultado de las diferentes combinaciones y algoritmos en comparación de XML-RoBERTA cuyo valor es de 0.1601 de precisión, es decir una diferencia de 0.0008 siendo un 0.50% de margen de predicción, lo que demuestra que el modelo Transformers XML-RoBERTA aplicado con el modelo ajustado Fine Tuning tiene una mayor exactitud de predicción en comparación a su rival.

Mejores Resultados

Tabla 28: Transformers modelos pre-entrenados

Transformers	M.A.E	M.S.E	R.M.S.E	PEARSON	ALGORITMO
XLM-RoBERTa	0.1623	0.0528	0.2270	0.99734	ABR
BERT	0.1632	0.0502	0.2343	0.9999	ABR
Roberta Large Bne	0.1671	0.0548	0.2328	0.9950	ABR

Nota. Características y resultados utilizados. Elaboración y fuente propia.

En la tabla No.28 Se puede observar que los Transformers con mayor exactitud aplicando el modelo pre-entrenado genérico de código abierto de cada modelo son los siguientes XML-RoBERTA con un 0.1623 de exactitud de predicción seguido de BERT con un 0.1632 y por último RoBERTA-Large-BNE con un 0.1671 de exactitud de predicción y estos resultados obtenidos con el algoritmo (ABR) Adaboost Regressor.

Tabla 29: Transformers modelos Fine Tuning

Transformers	M.A.E	M.S.E	R.M.S.E	PEARSON	ALGORITMO
BERT	0.1592	0.0379	0.1982	0.98830122	RFR
XLM-RoBERTa	0.1601	0.0501	0.2252	0.99871338	ABR
Roberta Large Bne	0.1609	0.0421	0.2047	0.67549011	GBR

Nota. Características y resultados utilizados. Elaboración y fuente propia

En la tabla No.29 Se puede observar que los Transformers con mayor exactitud aplicando el modelo ajustado Fine Tuning son los siguientes BERT con un 0.1592 de exactitud de predicción con el algoritmo (RFR) Random Forest Regressor, seguido de XLM-RoBERTa con un 0.1601 con el algoritmo (ABR) Adaboost Regressor y por último RoBERTA-Large-BNE con un 0.1609 de exactitud de predicción con el algoritmo (GBR) Gradient Boosting Regressor.

CAPÍTULO IV

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

- Se logro explorar y evaluar los diferentes modelos de Transformers obteniendo de ellos los resultados de la predicción de la complejidad Léxica, obtenidos de las diferentes combinaciones de los modelos pre-entrenado y modelo ajustado Fine-Tuning, lo que permitirá a futuras investigaciones aportar datos relevantes para la complexión léxica de textos académicos del idioma español.

- Se logro realizar la comparación de los resultados obtenidos de los diferentes algoritmos de Machine Learning de los modelos *Transformers*, logrando obtener una tendencia de algoritmos que presentan mejores resultados siendo estos (RFR) Random Forest Regressor, (ABR) Adaboost Regressor y (GBR) Gradient Boosting Regressor todos obtenidos del modelo ajustado Fine-Tuning en comparación al modelo pre-entrenado, demostrando una mejora en la predicción léxica de las palabras complejas en español.

- Se logro determinar que el modelo *Transformers* BERT aplicando el modelo ajustado de Fine-Tuning obtuvo el mejor desempeño en comparación a XML-RoBERTA y RoBERTa-Large-BNE tanto en los modelos pre-entrenados como los modelos ajustados aplicados con Fine-tuning.

Recomendaciones

- Se recomienda adaptar esta investigación en ser utilizada en otros modelos Transformers para así lograr una mejor comprensión lectora acerca de los distintos textos educativos en el idioma español.
- Se sugiere explorar las investigaciones actuales orientadas a la predicción de la complejidad léxica para obtener diferentes perspectivas, conocimientos relevantes que aporten a la solución de un mismo objetivo.
- Se aconseja adaptar este proyecto a las diferentes ramas educativas de la Universidad de Guayaquil que permitirá conocer la dificultad léxico en cada área estudiantil que contribuirá al planteamiento de soluciones académicas.
- Se propone ampliar el corpus de textos en idioma español con la integración en su totalidad de los textos académicos de las materias que constituyen en la Carrera de Ingeniera en Sistema y Software.

Trabajos futuros

- Se puede proponer la investigación de diferentes modelos Transformers y adecuarlos a la necesidad presente de los textos académicos.
- Se puede desarrollar un código que permita la identificación del modelo Transformers mas adecuado para la predicción de complejidad de la palabra compleja
- Se puede desarrollar la creación de un sistema que permita el ingreso de un texto académico y permita establecer las palabras complejas manualmente a través de una interfaz y elegir el modelo Transformers para obtener su complejidad léxica.

Bibliografía

1. Alarcon, R., & Moreno, L. &. (2020). Language Learning Disabilities Applied To University Educational Texts. *Task-CWI*.
2. Alban, G. P., Arguello, A. E., & & Molina, N. E. (2020). Metodologías de investigación educativa (descriptivas, experimentales, participativas, y de investigación-acción). *Recimundo*.
3. Aristu Ollero, A., & Torres Ríos, l. (2021). El mapeado de textos: un recurso para el aprendizaje lexico. *Revista Didáctica Español Lengua Extranjera*.
4. Arroba, S., & Pozo, L. (2021). Sistema de prediccion de la complejidad lexica implementando Machine Learning y redes neuronales para reducir barreras de la compresion lectora en los estudiantes universitarios.
5. Asanza, W. R., Olivo, B. M., & Peñafiel, E. M. (2018). Redes Neuronales artificiales aplicadas al reconocimiento de patrones . *UTMACH*.
6. Beltran, N. C., & Mojica, E. C. (2021). Procesamiento del Lenguaje Natural (PLN) - GPT-3 y su aplicacion de la Ingenieria de Software. *Revista TIA, Tecnologia, Investigacion y Academia* .
7. Borbor Merejildo, D. F., & Labre Hidalgo, K. F. (2021). Creación de un corpus de textos universitarios en español para la identificación de palabras complejas en el área de la simplificación léxica.
8. Brenes Jiménez, A. (2020). Predicción del caudal promedio horario de la estación hidrológica Palmar, utilizando modelos de Machine Learning basados en Árboles de decisión.
9. Castellano Lendínez, L. (2019). KANBAN. METODOLOGÍA PARA AUMENTAR LA EFICIENCIA DE LOS PROCESOS.
10. Castillo, K. N. L., Mendoza, O. R. G., & Perdomo, B. C. C. (2019). VYTEDU-CW: Difficult Words as a Barrier in the Reading Comprehension of University Students. *Advances in Emerging Trends and Technologies: Volume 1*.
11. Chulilla Alcalde, J. (2021). Análisis de sentimiento de textos basado en opiniones de películas usando algoritmos de aprendizaje computacional.
12. Collarte Gonzale, L. (2020). Procesamiento del lenguaje natural con BERT: Análisis de sentimientos en tuits.
13. Concepcion-Toledo, D. N. (2019). Metodología de la investigación: Origen y construcción de una tesis doctoral. *Revista Científica de la UCSA*, 76-87.
14. Cornelio, O. M. (2019). Algoritmo para determinar y eliminar nodos neutrales en Mapa Cognitivo Neutrosófico. *Neutrosophics Computing and Machine Learning*.
15. Corrales Beltrán, S. H. (2020). Métodos para el análisis de la información en corpus de artículos científicos con algoritmos de clasificación y librerías NLTK en la Plataforma Científica. *ECUCIENCIA*.
16. Erick Quezada, M. (2020). Plataforma de apoyo para personas con trastornos de Memoria.
17. Florian, A. &. (2021). 4 ANÁLISIS DE DESEMPEÑO DE ALGORITMOS DE REGRESIÓN USANDO SCIKIT-LEARN. MODUM. *Divulgativa Multidisciplinar de Ciencia, Tecnología e Innovación*.

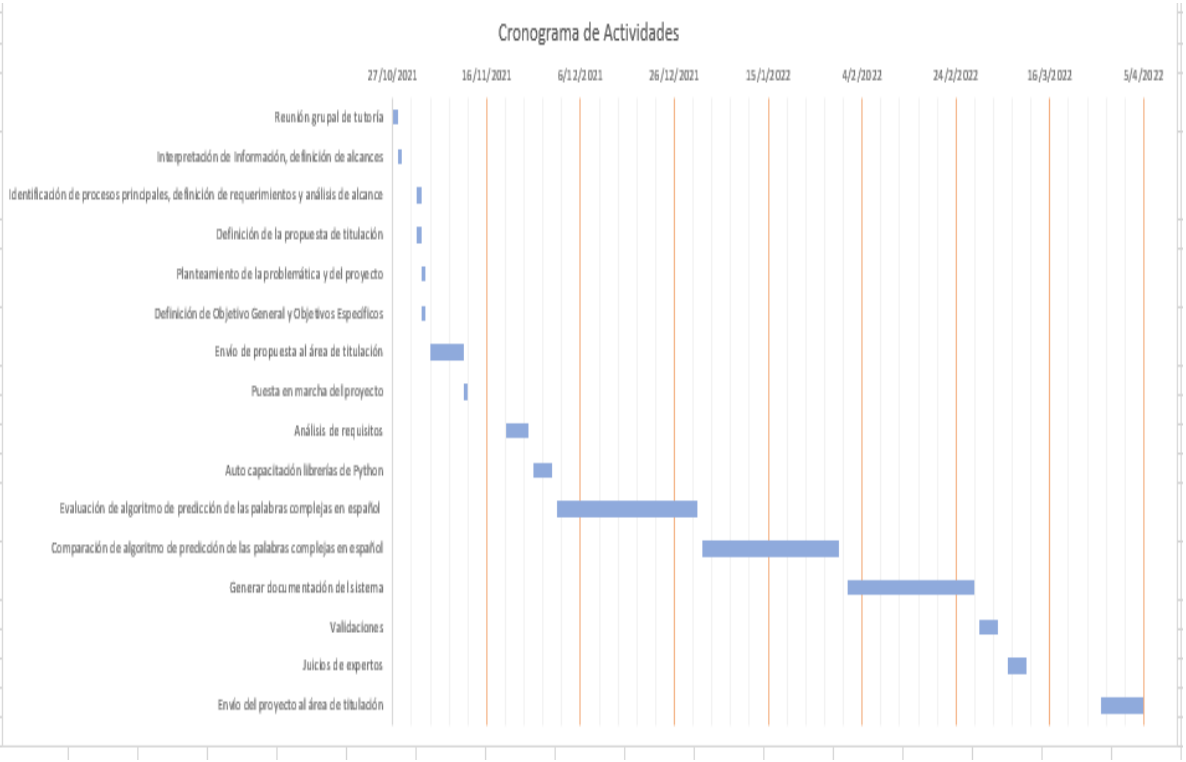
18. Garcia Garcia, M. A., Arevalo Duarte, M. A., & Hernandez Suarez, C. A. (2018). La Compresion Lectora y el rendimiento escolar . Cuadernos de Linguistica Hispanica .
19. Gonzalo Fuentes, P. E. (2019). Aplicación de algoritmos de Machine Learning para la predicción del beneficio por cliente a partir de métricas de Google Analytics.
20. Humeau, S., Shuster, K., Lachaux, M. A., & Weston, J. (2019). Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring.
21. Jara, C. A. (2021). Desarrollo de algoritmos para la clasificacion de textos biomedicos utilizando expresions regulares y aprendizaje activo .
22. Kevin Gaspar, T. P. (2021). SISTEMA DE PREDICCIÓN Y EVALUACIÓN DE LA COMPLEJIDAD LÉXICA.
23. Mahabadi, R. K., Ruder, S. D., & Henderson, J. (2021). Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks.
24. Mendizábal, M. A., Palma, I. A., Ramírez, E. D., Johany, Sánchez, A. S., & Álvarez, M. N. (2021). Implementación de inteligencia artificial y tecnología blockchain que permita optimizar sistemas productivos de la Pymes. *Congreso Internacional de Investigación Academia Journals Morelia* .
25. Messina Valverde, C. (2018). Librería de métodos de poda en conjuntos de clasificadores para Scikit-Learn. *Bachelor's thesis*.
26. Meza Rodríguez, A. R. (2018). Predicción de fuga de clientes en una empresa de telefonía utilizando el algoritmo Adaboost desbalanceado y la regresión logística asimétrica.
27. Mite-Baidal, K., Delgado-Vera, C., Solís-Avilés, E., Espinoza, A. H., Ortiz-Zambrano, J., & Varela-Tapia, E. (2018). Sentiment analysis in education domain: A systematic literature review. *International Conference on Technologies and Innovation*, 285-297.
28. Ortiz Zambrano, J. A., & Montejo Ráez, A. (2017). VYTEDU: Un corpus de vídeos y sus transcripciones para investigación en el ámbito educativo.
29. Ortiz Zambrano, J., MontejoRáez, A., Lino Castillo, K. N., Gonzalez Mendoza, O. R., & Cañizales Perdomo, B. C. (2019). VYTEDU-CW: Difficult Words as a Barrier in the Reading Comprehension of University Students. *The International Conference on Advances in Emerging Trends and Technologies*, 167-176.
30. Ortiz-Zambranoa, J. A., & Montejo-Ráezb, A. (2020). Overview of alexs 2020: First workshop on lexical analysis at sepln. *IberLEF 2020*.
31. Özçift, A., Akarsu, K., Yumuk, F., & Söylemez, C. (2021). Advancing natural language processing (NLP) applications of morphologically rich languages with bidirectional encoder representations from transformers (BERT). 226-238.
32. Palacios Villalobos, F. V. (2020). Plataformas y herramientas e–Learning para evaluar la velocidad y comprensión lectora del estudiante en la EBR.
33. Patil, S. P. (2018). Life Prediction of Bearing by Using Adaboost Regressor. In Proceedings of TRIBOINDIA-2018. *International Conference on Tribology*.
34. Penadillo Palomino, C. T. (2021). Aplicación de técnicas de análisis de regresión y aprendizaje automático para la estimación de sobre dilución en el método de Sub Level Stoping-Compañía Minera Condestable. *Condestable*.
35. Reategui, L., & Suarez, A. (2021). Sistema de prediccion de la complejidad lexica para contribuir a la reduccion de las barreras de la compresion lectora.
36. Sánchez, J. &. (2021). Red neuronal artificial para detección de armas de fuego y armas blancas en video vigilancia. *Iniciación Científica,*, 83-88.

37. Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*.
38. Shardlow, M., Evans, R., & paetzoold, G. &. (2021). Lexical Complexity prediction. *Semeval* .
39. Studio, V. (2019). Visual Studio, 2018.
40. Ubeda, P. L. (2021). Doctoral thesis Biomedical entities recognition in Spanish combining word embeddings.
41. Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*.
42. Ufarte Ruiz, M. J., & Manfredi Sánchez, J. L. (2019). Algoritmos y bots aplicados al periodismo. El caso de Narrativa Inteligencia Artificial: estructura, producción y calidad informativa.
43. Van den Broek, P. (2010). Using Texts in Science Education: Cognitive Processes and Knowledge Representation. *Science*.
44. Vidal-Silva, C. L.-O. (2021). Experiencia académica en desarrollo rápido de sistemas de información web con Python y Django.
45. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing.
46. Zagorulko, Y. A. (2021). Approach to automatic population of ontologies of scientific subject domain using lexico-syntactic patterns. *Journal of Physics: Conference Series* .
47. Zambrano, J. &. (2021). ClexLS2: A New Corpus for Complex Wprd Identification Research in computing Studies. *RANLP*.
48. Zambrano, J. A. O., & Montejo-Ráez, A. (2021). CLexIS2: A New Corpus for Complex Word Identification Research in Computing Studies. *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 1075-1083.
49. Zapata Garcia, A. (2021). Análisis de textos mediante técnicas NLP para la categorización de usuarios.

ANEXOS

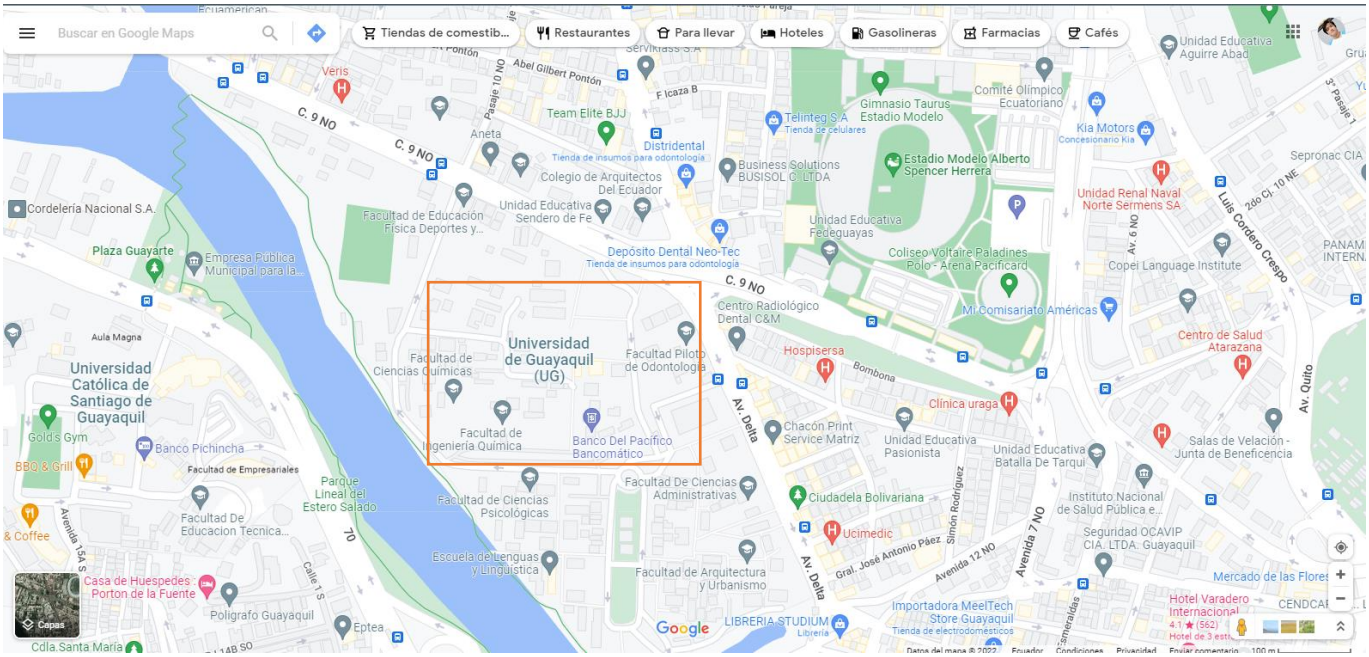
Anexo 1. Planificación de actividades del proyecto

Nombre de Actividad	Fecha de Inicio	Duracion en días	Fecha Fin
Reunión grupal de tutoría	27/10/2021	1	27/10/2021
Interpretación de Información, definición de alcances	28/10/2021	1	28/10/2021
Identificación de procesos principales, definición de requerimientos y análisis de alcance	1/11/2021	1	1/11/2021
Definición de la propuesta de titulación	1/11/2021	1	1/11/2021
Planteamiento de la problemática y del proyecto	2/11/2021	1	2/11/2021
Definición de Objetivo General y Objetivos Específicos	2/11/2021	1	2/11/2021
Envío de propuesta al área de titulación	4/11/2021	7	11/11/2021
Puesta en marcha del proyecto	11/11/2021	1	11/11/2021
Análisis de requisitos	20/11/2021	5	25/11/2021
Auto capacitación librerías de Python	26/11/2021	4	30/11/2021
Evaluación de algoritmo de predicción de las palabras complejas en español	1/12/2021	30	31/12/2021
Comparación de algoritmo de predicción de las palabras complejas en español	1/1/2022	29	30/1/2022
Generar documentación del sistema	1/2/2022	27	28/2/2022
Validaciones	1/3/2022	4	5/3/2022
Juicios de expertos	7/3/2022	4	11/3/2022
Envío del proyecto al área de titulación	27/3/2022	9	5/4/2022



Elaborado por: Diana Aroca y Diego Bernal
Fuente: Propia

Anexo 2. Geo-localización del problema



Elaborado por: Diana Aroca y Diego Bernal
Fuente: Google Maps

Anexo 4. Fundamentación Legal

El presente proyecto de titulación se fundamenta en la constitución, leyes y normas como se detalla a continuación.

Artículo de la LOES	**Contexto**
ART. 1 ÁMBITO	Esta Ley regula el sistema de educación superior en el país, a los organismos e instituciones que lo integran; determina derechos, deberes y obligaciones de las personas naturales y jurídicas, y establece las respectivas sanciones por el incumplimiento de las disposiciones contenidas en la Constitución y la presente Ley ARTICULO 1

Art. 2.- Objeto	Esta Ley tiene como objeto definir sus principios, garantizar el derecho a la educación superior de calidad que propenda a la excelencia interculturalidad, al acceso universal, permanencia, movilidad y egreso sin discriminación alguna y con gratuidad en el ámbito público hasta el tercer nivel.
Art. 4.- Derecho a la Educación Superior	El derecho a la educación superior consiste en el ejercicio efectivo de la igualdad de oportunidades, en función de los méritos respectivos, a fin de acceder a una formación académica y profesional con producción de conocimiento pertinente y de excelencia.
Art. 8.- Fines de la Educación Superior	Formar académicos y profesionales responsables, en todos los campos del conocimiento, con conciencia ética y solidaria, capaces de contribuir al desarrollo de las instituciones de la República, a la vigencia del orden democrático, y a estimular la participación social
Art. 144.- Principios	Art. 144.- Tesis Digitalizadas. - Todas las instituciones de educación superior estarán obligadas a entregar las tesis que se elaboren para la obtención de títulos académicos de grado y posgrado en formato digital para ser integradas al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública respetando los derechos de autor.
Artículo. 26	La educación es un derecho de las personas a lo largo de su vida y un deber ineludible e inexcusable del Estado. Constituye un área prioritaria de la política pública y de la inversión estatal, garantía de la igualdad e inclusión social y condición indispensable para el buen vivir
Artículo 28	La educación responderá al interés público y no estará al servicio de intereses individuales y corporativos. Se garantizará el acceso universal, permanencia, movilidad y egreso sin discriminación alguna
ARTÍCULO 340	EI sistema nacional de inclusión y equidad social es el conjunto articulado y coordinado de sistemas, instituciones,

	políticas, normas, programas y servicios que aseguran el ejercicio, garantía y exigibilidad de los derechos reconocidos en la Constitución y el cumplimiento de los objetivos del régimen de desarrollo. El sistema se articulará al Plan Nacional de Desarrollo y al sistema nacional descentralizado de planificación participativa; se guiará por los principios de universalidad, igualdad, equidad, progresividad, interculturalidad, solidaridad y no discriminación; y funcionará bajo los criterios de calidad, eficiencia, eficacia, transparencia, responsabilidad y participación. El sistema se compone de los ámbitos de la educación, salud, seguridad social, gestión de riesgos, cultura física y deporte, hábitat y vivienda, cultura, comunicación e información, disfrute del tiempo libre, ciencia y tecnología, población, seguridad humana y transporte.
ARTÍCULO 341	El Estado generará las condiciones para la protección integral de sus habitantes a lo largo de sus vidas, que aseguren los derechos y principios reconocidos en la Constitución, en particular la igualdad en la diversidad y la no discriminación, y priorizará su acción hacia aquellos grupos que requieran consideración especial por la persistencia de desigualdades, exclusión, discriminación o violencia, o en virtud de su condición etaria, de salud o de discapacidad. La protección integral funcionará a través de sistemas especializados, de acuerdo con la ley. Los sistemas especializados se guiarán por sus principios específicos y los del sistema nacional de inclusión y equidad social. El sistema nacional descentralizado de protección integral de la niñez y la adolescencia será el encargado de asegurar el ejercicio de los derechos de niñas, niños y adolescentes. Serán parte del sistema las instituciones públicas, privadas y comunitarias.
ARTÍCULO 348	La educación pública será gratuita y el Estado la financiará de manera oportuna, regular y suficiente. La distribución de

	los recursos destinados a la educación se regirá por criterios de equidad social, poblacional y territorial, entre otros. El Estado financiará la educación especial y podrá apoyar financieramente a la educación fiscomisional, artesanal y comunitaria, siempre que cumplan con los principios de gratuidad, obligatoriedad e igualdad de oportunidades, rindan cuentas de sus resultados educativos y del manejo de los recursos públicos, y estén debidamente calificadas, de acuerdo con la ley. Las instituciones educativas que reciban financiamiento público no tendrán fines de lucro. La falta de transferencia de recursos en las condiciones señaladas será sancionada con la destitución de la autoridad y de las servidoras y servidores públicos remisos de su obligación.
Artículo 350	El sistema de educación superior tiene como finalidad la formación académica y profesional con visión científica y humanista; la investigación científica y tecnológica; la innovación, promoción, desarrollo y difusión de los saberes y las culturas; la construcción de soluciones para los problemas del país, en relación con los objetivos del régimen de desarrollo.

Elaborado por: Diana Aroca y Diego Bernal
Fuente: Ley Orgánica de Educación Superior

Decreto N.1014 de Software libre	**Contexto**
Artículo 1	Establecer como política pública para las Entidades de Administración Pública Central la utilización del Software Libre en sus sistemas y equipamientos informáticos.
Artículo 2	Se entiende por software libre, a los programas de computación que se pueden utilizar y distribuir sin restricción alguna, que permitan su acceso a los códigos fuentes y que sus aplicaciones puedan ser mejoradas
Artículo 3	Las entidades de la Administración Pública Central previa a la instalación del software libre en sus equipos, deberán verificar la existencia de capacidad

	técnica que brinde el soporte necesario para el uso de este tipo de software.
Artículo 4	Se faculta la utilización de software propietario (no libre) únicamente cuando no exista una solución de software libre que supla las necesidades requeridas, o cuando esté en riesgo la seguridad nacional, o cuando el proyecto informático se encuentre en un punto de no retorno

Elaborado por: Diana Aroca y Diego Bernal
Fuente: Estrategia para la Implementación de Software Libre en la Administración Pública Central

Anexo 5. Criterios éticos para utilizarse en el desarrollo del proyecto

Criterios	**Características del criterio**	**Procedimiento**
Credibilidad	Aproximación de los resultados de una investigación frente al fenómeno observado	Hallazgos reales sobre la problemática de la investigación. Relevancia del estudio.
Transferibilidad	Conocimiento sobre el contexto que permite transferir las conclusiones a contextos similares	Proporcionar información detallada del contexto. Muestro teórico.
Dependencia	Estabilidad relativa y variabilidad de los datos	Proceso de recolección análisis e interpretación de los datos
Confirmabilidad	Refleja la veracidad de los resultados y la investigación realizada	La información está respaldada de fuentes confiables y científicas Resultados evaluados y confirmados por personas externas a la investigación

Elaborado por: Diana Aroca y Diego Bernal
Fuente: Datos de la investigación

Anexo 7. Validación de expertos

Apellidos Y Nombres del Experto	**Título Profesional del Experto**	**Autores**				
Arroba Salinas Sandra Yanel	Ingeniera en sistemas computacionales	Aroca Pincay Diana Geovanna		Bernal Yucailla Diego Gabriel		
Título del Proyecto	EVALUACIÓN DE DESEMPEÑO DE TRANSFORMADORES PARA LA PREDICCIÓN LÉXICA PARA EL IDIOMA ESPAÑOL.					
Indicador	**Criterio**	**Deficiente**	**Regular**	**Buena**	**Muy Buena**	**Excelente**
Actualidad	El Sistema es determinado a la investigación elaborada con base científica-teórica				X	
Efectividad	Los resultados de predicción se aproximan a los valores reales.					X
Capacidad de Solución	El Sistema refleja conclusiones de predicción a niveles de corpus en el idioma español					X
Usabilidad	Interfaz adaptable y sutil.				X	
Consistencia	El sistema ofrece una respuesta a la dificultad de la investigación				X	
Portabilidad	Es muy fácil de instalar y ejecutar.					X
Freeware	La utilización del Software es libre.					X
Reutilización	El Sistema aplica métodos correspondientes al lenguaje de programación					X
Flexibilidad	El lenguaje de Programación es flexible a la innovación del desarrollador.					X
Aplicabilidad	Utilización de Ensemble Methods y Machine Learning.				X	

Constancia de Juicio de experto

Estimada Ingeniera

M.Sc Jenny Ortiz Zambrano

DOCENTE TUTOR DEL TRABAJO DE TITULACION

Guayaquil, 12 de marzo de 2022

El presente instrumento certifica que se realizó la revisión del proyecto de titulación "EVALUACIÓN DE DESEMPEÑO DE TRANSFORMADORES PARA LA PREDICCIÓN LÉXICA PARA EL IDIOMA ESPAÑOL." cuyos criterios e indicaciones empleados permitieron articular el trabajo según se muestra en el Anexo 7, por tanto, Aroca Pincay Diana Geovanna y Bernal Yucailla Diego Gabriel estudiantes no titulados de la Carrera de Ingeniería en sistemas computacionales de la Universidad de Guayaquil, pueden continuar con el proceso de titulación en vista que no existen observaciones.

Por lo actuado en el Anexo 7, se procede a validar el trabajo de titulación

Sin otro particular

Sandra Arroba Salinas

Arroba Salinas Sandra Yanel

C.I. N° 0955465091

Apellidos Y Nombres del Experto	Título Profesional del Experto	Autores				
Demetrio Oswaldo González Silva	Ingeniero en sistemas computacionales & analista de sistemas	Aroca Pincay Diana Geovanna		Bernal Yucailla Diego Gabriel		
Título del Proyecto	EVALUACIÓN DE DESEMPEÑO DE TRANSFORMADORES PARA LA PREDICCIÓN LÉXICA PARA EL IDIOMA ESPAÑOL.					
Indicador	Criterio	Deficiente	Regular	Buena	Muy Buena	Excelente
Usabilidad	El Sistema muestra una interfaz sencilla y es intuitivo de fácil uso					x
Capacidad de Respuesta	El sistema genera como resultado el nivel de predicción de la palabra compleja contenida en los corpus que están en los idiomas inglés y español				x	
Efectividad	Los resultados predictivos se asemejan al resultado real				x	
Flexibilidad	El sistema emplea lenguaje de programación de vanguardia que puede ser modificado por los desarrolladores					x
Reutilización	El sistema hace uso de recursos propios del lenguaje				x	
Freeware	El desarrollo del sistema emplea aplicaciones de software libre					x
Actualidad	Es adecuado al contexto de la investigación ya que se basa en aspectos teóricos y científicos					x
Aplicabilidad	El sistema utiliza algoritmos de machine learning y redes neuronales					x
Consistencia	El instrumento responde al problema formulado de investigación					x
Portabilidad	Facilidad de instalación, ajuste					x

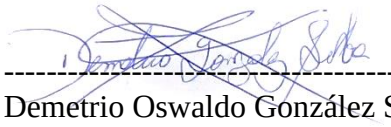
Validación de expertos

Constancia de juicio de experto

Estimada Ingeniera
M.Sc Jenny Ortiz Zambrano

DOCENTE TUTOR DEL TRABAJO DE TITULACION
Guayaquil, 12 de marzo de 2022
El presente instrumento certifica que se realizó la revisión del proyecto de titulación "EVALUACIÓN DE DESEMPEÑO DE TRANSFORMADORES PARA LA PREDICCIÓN LÉXICA PARA EL IDIOMA ESPAÑOL." cuyos criterios e indicaciones empleados permitieron articular el trabajo según se muestra en el Anexo 7, por tanto, Aroca Pincay Diana Geovanna y Bernal Yucailla Diego Gabriel estudiantes no titulados de la Carrera de Ingeniería en sistemas computacionales de la Universidad de Guayaquil, pueden continuar con el proceso de titulación en vista que no existen observaciones

Por lo actuado en el Anexo 7, se procede a validar el trabajo de titulación
Son otro particular



Demetrio Oswaldo González Silva
C.I. N° 0929373488

Apellidos Y Nombres del Experto	**Título Profesional del Experto**	**Autores**	
Yépez Paladines Juan Carlos	Licenciado en Sistema de Información	Aroca Pincay Diana Geovanna	Bernal Yucailla Diego Gabriel
Título del Proyecto	EVALUACIÓN DE DESEMPEÑO DE TRANSFORMADORES PARA LA PREDICCIÓN LÉXICA PARA EL IDIOMA ESPAÑOL.		

Indicador	**Criterio**	**Deficiente**	**Regular**	**Buena**	**Muy Buena**	**Excelente**
Actualidad	El Sistema es determinado a la investigación elaborada con base científica-teórica					X
Efectividad	Los resultados de predicción se aproximan a los valores reales.					X
Capacidad de Solución	El Sistema refleja conclusiones de predicción a niveles de corpus en el idioma español					X
Usabilidad	Interfaz adaptable y sutil.					X
Consistencia	El sistema ofrece una respuesta a la dificultad de la investigación				X	
Portabilidad	Es muy fácil de instalar y ejecutar.					X
Freeware	La utilización del Software es libre.					x
Reutilización	El Sistema aplica métodos correspondientes al lenguaje de programación					X
Flexibilidad	El lenguaje de Programación es flexible a la innovación del desarrollador.				X	
Aplicabilidad	Utilización de Ensemble Methods y Machine Learning.					X

Validación de expertos

Constancia de juicio de experto

Estimada Ingeniera

M.Sc Jenny Ortiz Zambrano

DOCENTE TUTOR DEL TRABAJO DE TITULACION

Guayaquil, 12 de marzo de 2022

El presente instrumento certifica que se realizó la revisión del proyecto de titulación "EVALUACIÓN DE DESEMPEÑO DE TRANSFORMADORES PARA LA PREDICCIÓN LÉXICA PARA EL IDIOMA ESPAÑOL." cuyos criterios e indicaciones empleados permitieron articular el trabajo según se muestra en el Anexo 7, por tanto, Aroca Pincay Diana Geovanna y Bernal Yucailla Diego Gabriel estudiantes no titulados de la Carrera de Ingeniería en sistemas computacionales de la Universidad de Guayaquil, pueden continuar con el proceso de titulación en vista que no existen observaciones.

Por lo actuado en el Anexo 7, se procede a validar el trabajo de titulación

Sin otro particular

Yépez Paladines Juan Carlos

C.I. N° 0913238572

Anexo 10. Acta de Entrega y recepción definitivo

En la ciudad de Guayaquil, a 26 días del mes de marzo del 2022

Por el presente documento

Los estudiantes no titulados de la Carrera de Ingeniería en Sistemas Computacionales Aroca Pincay Diana Geovanna con cédula de identidad N° 0954536041 y Bernal Yucailla Diego Gabriel con cédula de identidad N° 0952491025 hacemos la entrega del código fuente del proyecto de titulación a la Dirección de la Carrera de Ingeniería en Sistemas Computacionales en un medio magnético.

Los códigos del programa/producto que se encargaron por compromiso al estar inserto en el proceso de titulación desde fecha 29 de noviembre de 2021.

Para efectos de dar cumplimiento a la entrega del código fuente, cedo todos los derechos de explotación sobre el programa y, en concreto, los de transformación, comunicación pública, distribución y reproducción, de forma exclusiva, con un ámbito territorial nacional.

Aroca Pincay Diana Geovanna — 0954536041 cédula de identidad

Bernal Yucailla Diego Gabriel — 0952491025 cédula de identidad

Anexo 13. Manual Técnico

UNIVERSIDAD DE GUAYAQUIL
FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

MANUAL TECNICO
Evaluación de Desempeño de los modelos transformadores para la predicción de la complejidad léxica para el idioma español

AUTORES:
Aroca Pincay Diana Geovanna
Bernal Yucailla Diego Gabriel

TUTORA:
MSc. JENNY ORTIZ ZAMBRANO

GUAYAQUIL – ECUADOR
2022

Preparación del entorno de desarrollo

Requerimientos recomendados

- 2 GB de VRAM
- Core i5-8265U 1,60 GHz
- 8 GB de RAM
- Windows 10 de 64 bits

Requerimientos mínimos

- 4 GB de RAM
- Core i3-3225 3,3 GHz
- Windows 7 de 64 bits o versión de macOS

Herramientas	Versiones y Tipo
Lenguaje de programación	Python 3.8.10
Entorno de desarrollo	Visual Studio Code
Interfaz Grafica	Tkinter
Librerías	SpaCy 2.3.5, NLTK 3.5, scikit-learn 0.23.2, matploib 3.3.2, syllabus 0.1.0, Pandas 1.1.3, Numpy 1.19.2, RE, Threading, lexical_diversity 0.0.1,os, stadistics, seaborn 0.11.0, statmodels 0.12.O

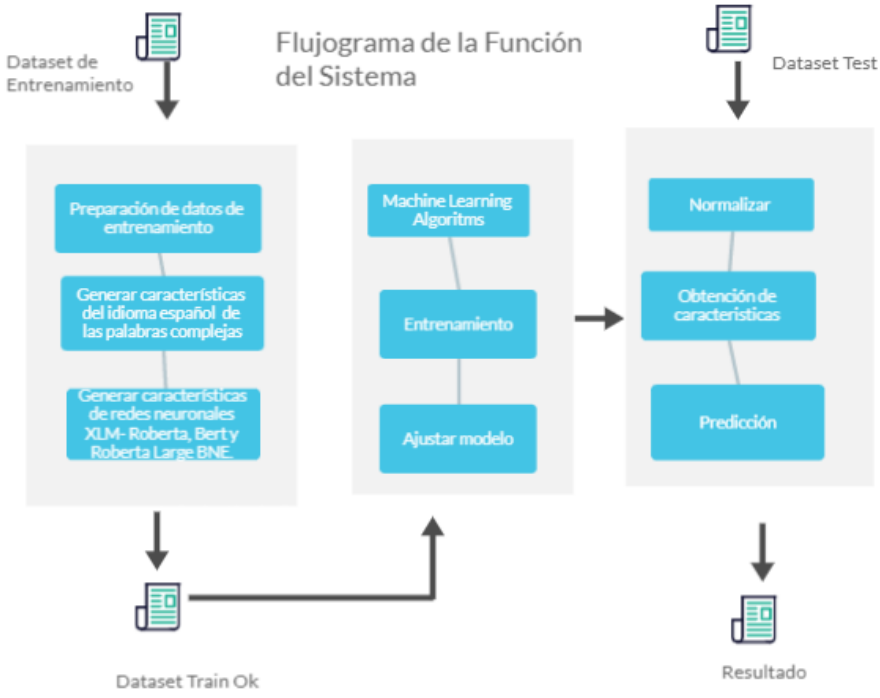
Elaborado por: Diana Aroca y Diego Bernal
Fuente: Propia

Para el Sistema de este proyecto se utilizó el entorno de Visual Studio Code por lo cual se procedió instalar la mayoría de las librerías y configuraciones de este, además se utilizó TKINTER para la interfaz gráfica por su facilidad integración con Python.

Elaborado por: Diana Aroca y Diego Bernal
Fuente: Propia

Librerías usadas para la configuración del entorno de trabajo en Visual Studio Code

Flujograma de la Función del Sistema

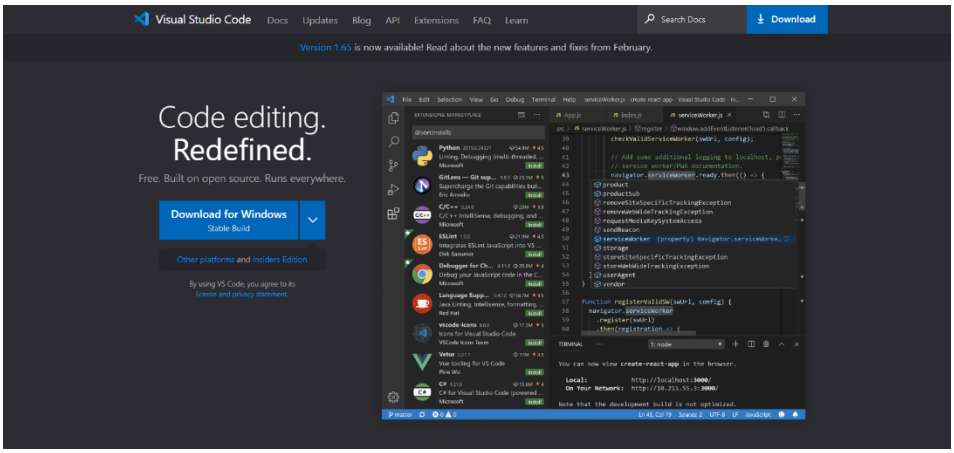


Elaborado por: Diana Aroca y Diego Bernal
Fuente: Propia

Instalación de las Herramientas Utilizadas

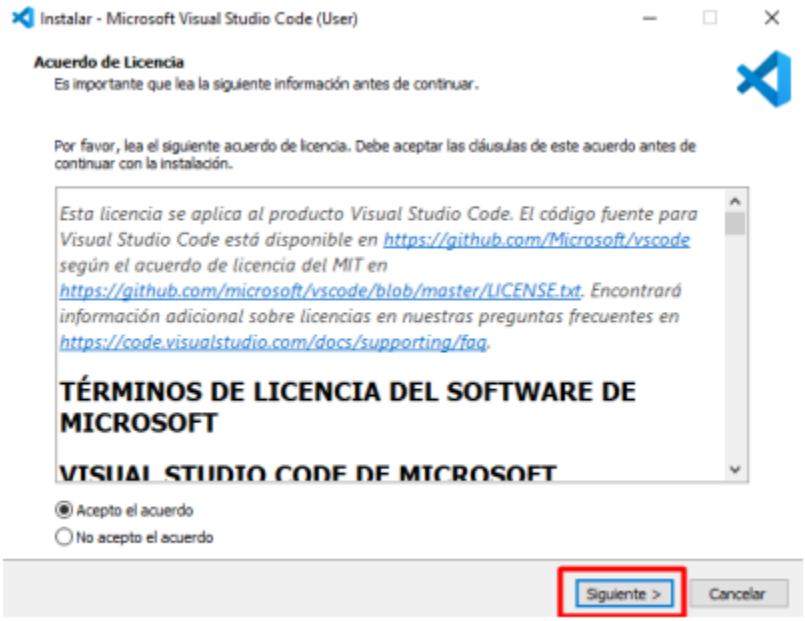
Instalación de Visual Studio Code

Ingresar a la página oficial de Visual Studio Code en el siguiente enlace https://code.visualstudio.com/ luego seleccionar la versión de acuerdo con su sistema operativo.



Elaborado por: Diana Aroca y Diego Bernal
Fuente: Propia

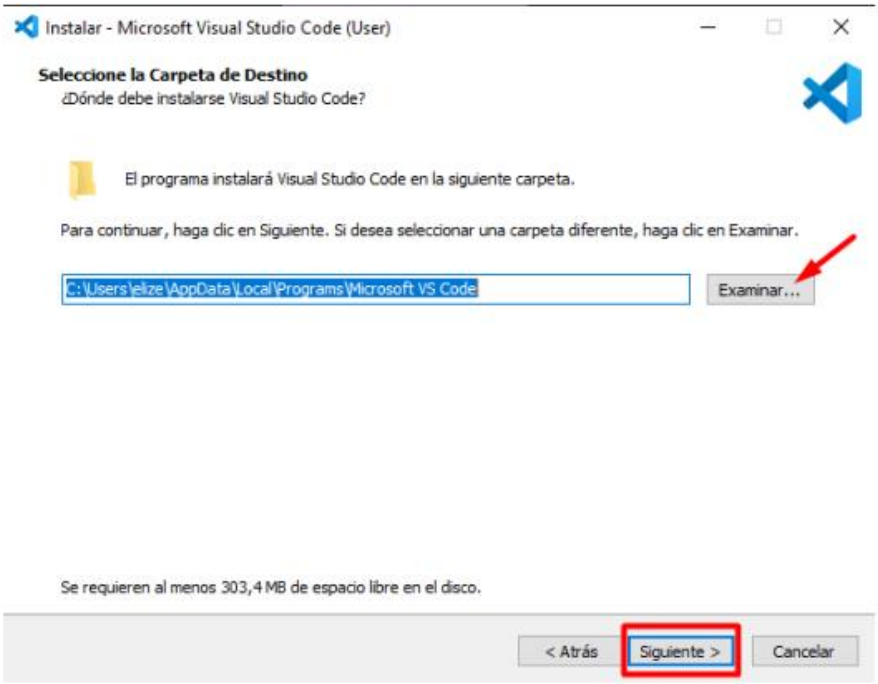
Una vez ejecutada la aplicación como administrador, aparecerá la siguiente ventana donde debe aceptar los términos y condiciones, dar clic en el botón siguiente.



Elaborado por: Diana Aroca y Diego Bernal
Fuente: Propia

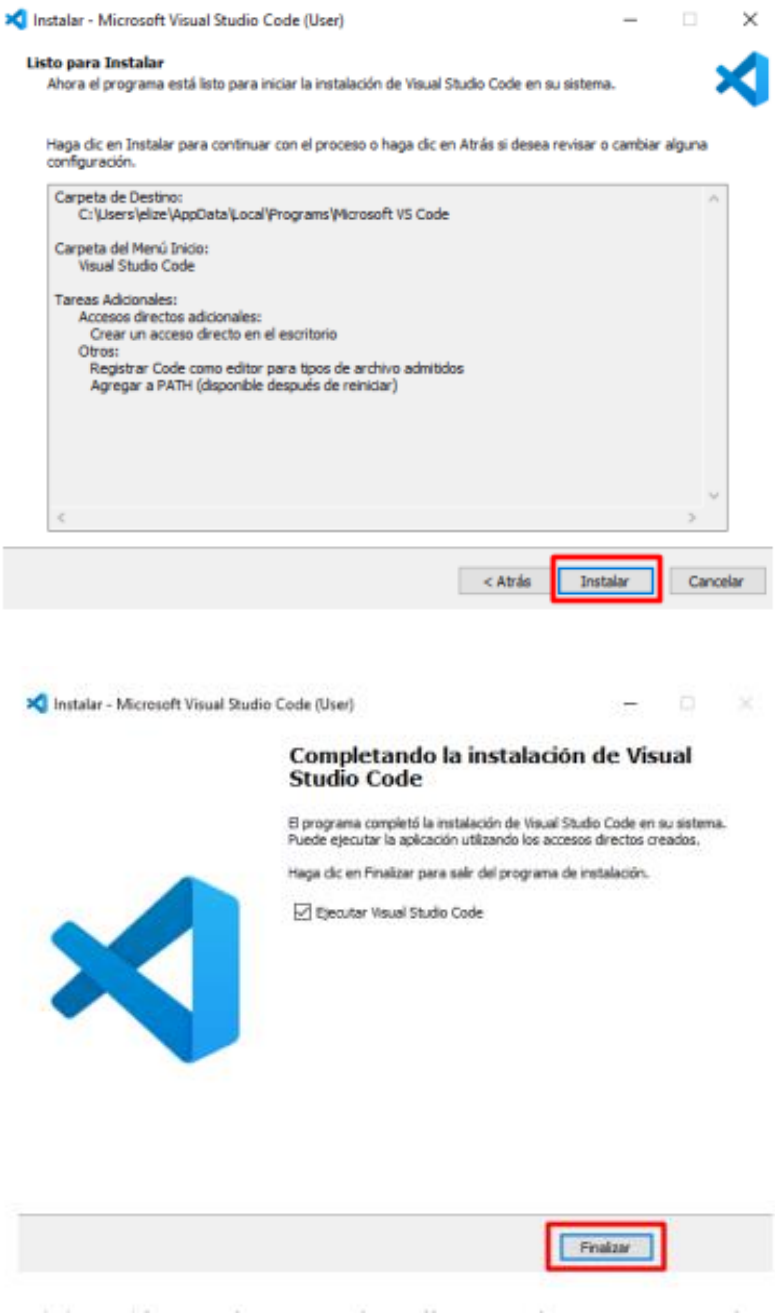
Aceptando los términos y políticas se procederá con las siguientes ventanas emergentes de instalación que por defecto(recomendado) se procederá la instalación de todos los recursos

necesarios y configuraciones o para gusto del usuario puede llegar a realizar alguna modificación.



Elaborado por: Diana Aroca y Diego Bernal
Fuente: Propia

Ya aceptando los términos y configuraciones recomendados se instalara por completo el entorno de desarrollo de Visual Studio Code para su uso.

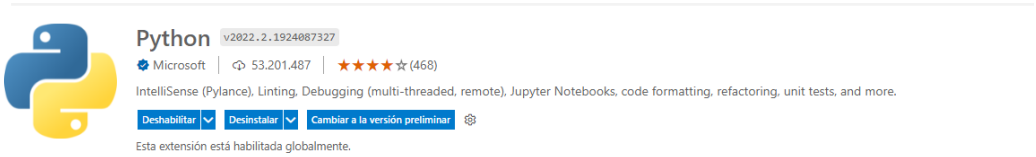


Elaborado por: Diana Aroca y Diego Bernal
Fuente: Propia

Instalación de Python

Ingresar a la página oficial de Python https://www.python.org/downloads/Python y buscar la versión deseada 3.8.10, luego dar clic en el botón download para iniciar la descarga y posteriormente para iniciar la instalación

Una vez instalado el aplicativo deberá ingresar al Visual Studio Code y buscar la extensión de Python y dar clic en Install.



Elaborado por: Diana Aroca y Diego Bernal
Fuente: Propia

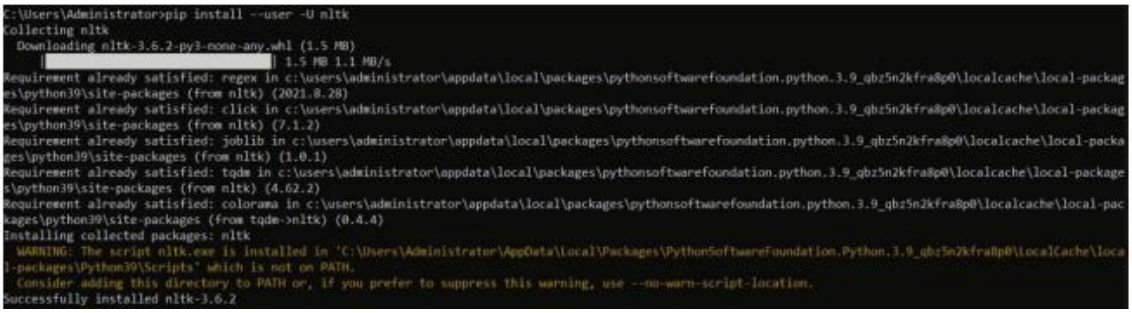
Instalación de las Librerías Utilizadas

Instalación de NLTK

NLTK es un conjunto de bibliotecas y programas para el procesamiento del lenguaje natural simbólico y estadísticos para el lenguaje de programación Python. En el método de inicialización de la aplicación esta la siguiente línea de código que permite bajar de forma local en el repositorio de anaconda los corpus con el cual trabajo nltk para hacer uso de sus funciones se procede a indicar de esta forma para las peticiones que se utilizarán:

- onltk.download('wordnet')
- nltk.download('omw')

En este sistema se utiliza wordnet para sacar varias características como los sinónimos de las palabras, hiperónimos, hipónimos y OMW (Open Multilingual Wordnet) que se requiere de especificaciones como el 'lang=…', esté parámetro es una adición realizada que vincula wordnet de diferentes idiomas a Princeton WordNet versión 3.0. El comando para la instalación es: pip install --user -U nltk



Elaborado por: Diana Aroca y Diego Bernal
Fuente: Propia

De la misma manera se procede instalar manualmente las librerías necesarias para la ejecución del sistema a través del comando pip install, ejemplos:

- pip install aiohttp==3.7.4.post0
- pip install anyio==3.0.1
- pip install apache-beam==2.30.0
- pip install argon2-cffi==20.1.0

Instalación de Transformers

Transformers proporciona miles de modelos previamente entrenados para hacer trabajos en textos como categorización, sustracción de información, esto y más en varios lenguajes. La librería Transformers está respaldada por bibliotecas de aprendizaje profundo como PyTorch y TensorFlow. Para la instalación se ejecuta el comando: pip install Transformers.

```
Microsoft Windows [Versión 10.0.19044.1586]
(c) Microsoft Corporation. Todos los derechos reservados.

C:\WINDOWS\system32>pip install Transformers
```

Elaborado por: Diana Aroca y Diego Bernal
Fuente: Propia

Instalación de PyTorch

Es un paquete de Python diseñado para realizar cálculos numéricos haciendo uso de la programación de tensores, pytorch.tensor es una matriz multidimensional que contiene elementos de un solo tipo de datos. El comando para la instalación es:

- **conda install pytorch torchvision -c pytorch**

```
Collecting package metadata (current_repodata.json): done
Solving environment: done

## Package Plan ##

  environment location: C:\ProgramData\Anaconda3

  added / updated specs:
    - pytorch
    - torchvision


The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    cudatoolkit-10.1.243       |       h74a9793_0       300.3 MB
    libuv-1.40.0               |       he774522_0         255 KB
    ninja-1.10.2               |       h6d14046_1         250 KB
    pytorch-1.8.1              |py3.8_cuda10.1_cudnn7_0       835.7 MB  pytorch
    torchvision-0.9.1          |       py38_cu101         7.2 MB  pytorch
    ------------------------------------------------------------
                                           Total:        1.12 GB

The following NEW packages will be INSTALLED:

  cudatoolkit        pkgs/main/win-64::cudatoolkit-10.1.243-h74a9793_0
  libuv              pkgs/main/win-64::libuv-1.40.0-he774522_0
  ninja              pkgs/main/win-64::ninja-1.10.2-h6d14046_1
  pytorch            pytorch/win-64::pytorch-1.8.1-py3.8_cuda10.1_cudnn7_0
  torchvision        pytorch/win-64::torchvision-0.9.1-py38_cu101


Proceed ([y]/n)? y


Downloading and Extracting Packages
ninja-1.10.2         | 250 KB    | ##################################################
```

Elaborado por: Diana Aroca y Diego Bernal
Fuente: Propia

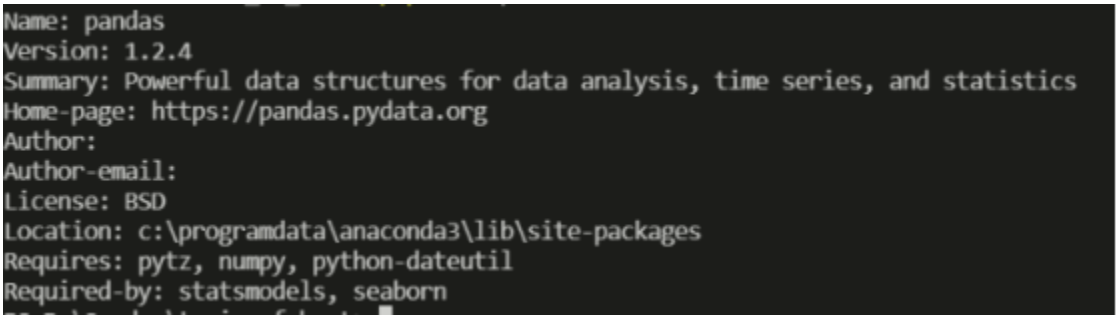
Scikit-learn

Está librería facilita trabajar con algoritmos de aprendizajes. Es reconocida por ser una librería de código abierto que utiliza Random Forest Regressor; encapsula preprocesamiento de datos, selección de características, métricas de rendimiento.

- from sklearn.esemble import RandomForestRegresor
- from sklearn.model_selection import train_test_split
- from sklearn import metrics, preprocessing

Instalación de Pandas

Esta librería se utiliza para la lectura y escritura de los archivos de Excel o CSV, además que posibilita el trabajo con los DataFrame.



Elaborado por: Diana Aroca y Diego Bernal
Fuente: Propia

Anexo 14. Manual de Usuario

UNIVERSIDAD DE GUAYAQUIL
FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

MANUAL USUARIO
Evaluación de Desempeño de los modelos transformadores para la predicción de la complejidad léxica para el idioma español

AUTORES:
Aroca Pincay Diana Geovanna
Bernal Yucailla Diego Gabriel

TUTORA:
MSc. JENNY ORTIZ ZAMBRANO

GUAYAQUIL – ECUADOR
2022

REQUERIMIENTOS

Requerimientos Recomendados

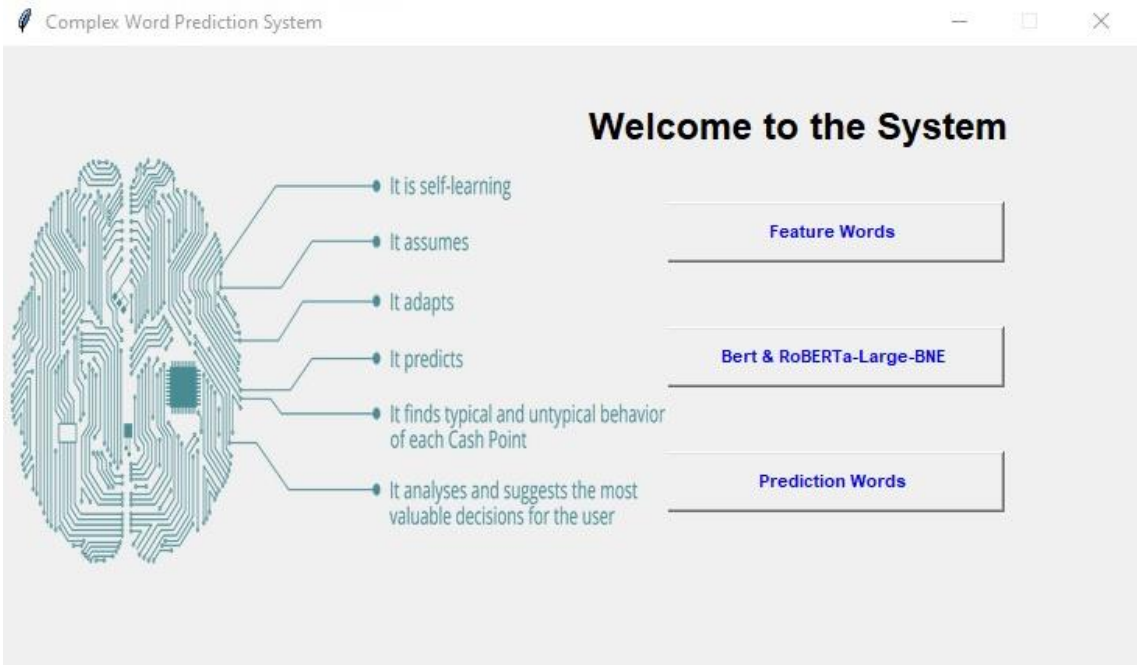
- 2 GB de VRAM
- Core i5-8265U 1,60 GHz
- 8GB de RAM
- Windows 10 de 64 bits

Requerimientos Mínimos

- 4 GB de RAM
- Core i3-3225 3,3 GHz
- Windows 7 de 64 bits o versión de macOS

PRESENTACIÓN

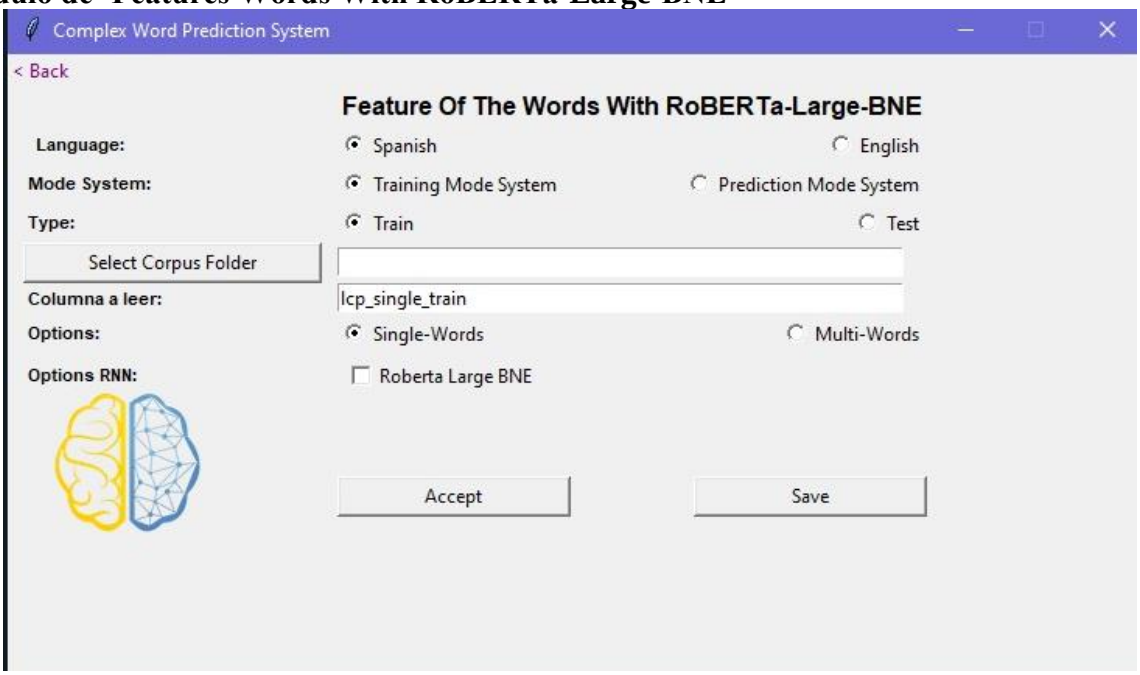
VENTANA DE INICIO



Elaborado por: Diana Aroca y Diego Bernal
Fuente: Propia

En la pantalla principal aparecerán 3 botones: ‘Feature words’ (lleva al módulo donde se extraen las características de las palabras),’Bert & Roberta-Large-BNE’ (saca características de las palabras con redes neuronales) y ’Prediction Words’ (los llevara al módulo de predecir nivel de complejidad)

Módulo de 'Features Words With RoBERTa-Large-BNE'

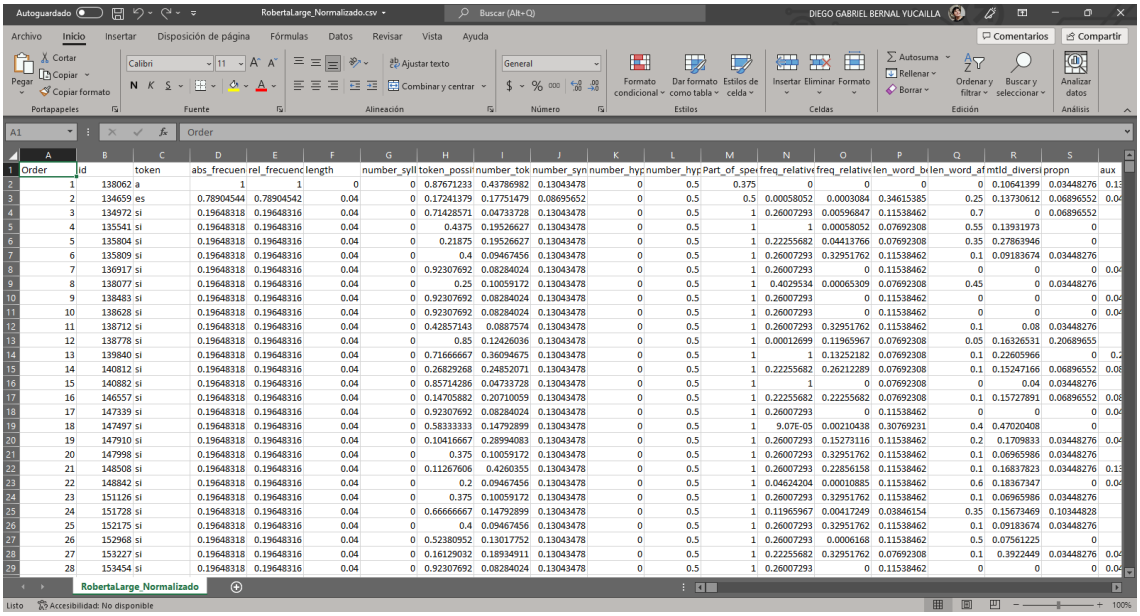


Elaborado por: Diana Aroca y Diego Bernal
Fuente: Propia

1. **Language:** Seleccionar el lenguaje
2. **Mode System:** El modo del Sistema se refiere a que el data Set que se va a analizar si tiene el nivel de complejidad tanto para entrenamiento como para prueba.
3. **Type:** El tipo viene a ser si es de entrenamiento o prueba (Train-Test).
4. **Select Corpus Folder**: Es para seleccionar el archivo del corpus a generar las características.
5. **Columna a leer:** Se presenta el nombre del archivo que debe ir al archivo a seleccionar.
6. **Options:** Para seleccionar Palabras simples (Single-Words), es decir de una sola palabra compleja o Palabras múltiples (Multi-Words), que serían para dos palabras.
7. **Options RNN:** Selecciona las opción del modelo Transformers que se quiera utilizar(RoBERTa-Large-BNE)
8. **Accept:** El botón Accept, comienza a generar el archivo resultante con las características.
9. **Save:** El botón Save, comienza a guardar el archivo con los resultados.

EJEMPLO

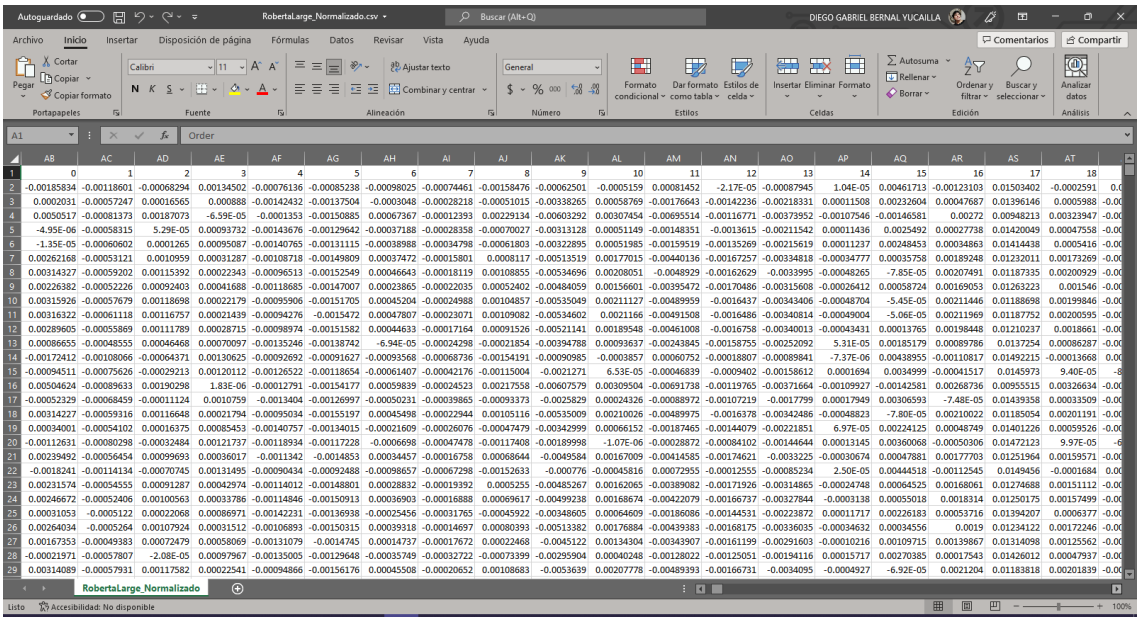
Al ejecutar el modelo Transformers deseado se procederá a crearse el Datasets con todas las características correspondientes del modelo.



Order	id	token	abs_frecuen	rel_frecuenc	length	number_syll	token_possi	number_tok	number_syn	number_hyp	number_hyp	Part_of_spe	freq_relativ	freq_relativ	len_word_b	len_word_af	mtld_diversi	propn	aux
1	138062	a	1	1	0	0	0.87671233	0.43786982	0.13043478	0	0.5	0.375	0	0	0	0	0.10641399	0.03448276	0.13
2	134659	es	0.78904544	0.78904542	0.04	0	0.17241379	0.17751479	0.08695652	0	0.5	0.5	0.00058052	0.0003084	0.34615385	0.25	0.13730612	0.06896552	0.04
3	134972	si	0.19648318	0.19648316	0.04	0	0.71428571	0.04733728	0.13043478	0	0.5	1	0.26007293	0.00596847	0.11538462	0.7	0	0.06896552	
4	135541	si	0.19648318	0.19648316	0.04	0	0.4375	0.19526627	0.13043478	0	0.5	1	1	0.00058052	0.07692308	0.55	0.13931973	0	
5	135804	si	0.19648318	0.19648316	0.04	0	0.21875	0.19526627	0.13043478	0	0.5	1	0.22255682	0.04413766	0.07692308	0.35	0.27863946	0	
6	135809	si	0.19648318	0.19648316	0.04	0	0.4	0.09467456	0.13043478	0	0.5	1	0.26007293	0.32951762	0.11538462	0.1	0.09183674	0.03448276	
7	136917	si	0.19648318	0.19648316	0.04	0	0.92307692	0.08284024	0.13043478	0	0.5	1	0.26007293	0	0.11538462	0	0	0	0.04
8	138077	si	0.19648318	0.19648316	0.04	0	0.25	0.10059172	0.13043478	0	0.5	1	0.4029534	0.00065309	0.07692308	0.45	0	0.03448276	
9	138483	si	0.19648318	0.19648316	0.04	0	0.92307692	0.08284024	0.13043478	0	0.5	1	0.26007293	0	0.11538462	0	0	0	0.04
10	138628	si	0.19648318	0.19648316	0.04	0	0.92307692	0.08284024	0.13043478	0	0.5	1	0.26007293	0	0.11538462	0	0	0	0.04
11	138712	si	0.19648318	0.19648316	0.04	0	0.42857143	0.0887574	0.13043478	0	0.5	1	0.26007293	0.32951762	0.11538462	0.1	0.08	0.03448276	
12	138778	si	0.19648318	0.19648316	0.04	0	0.85	0.12426036	0.13043478	0	0.5	1	0.00012699	0.11965967	0.07692308	0.05	0.16326531	0.20689655	
13	139840	si	0.19648318	0.19648316	0.04	0	0.71666667	0.36094675	0.13043478	0	0.5	1	1	0.13252182	0.07692308	0.1	0.22605966	0	0.2
14	140812	si	0.19648318	0.19648316	0.04	0	0.26829268	0.24852071	0.13043478	0	0.5	1	0.22255682	0.26212289	0.07692308	0.1	0.15247166	0.06896552	0.08
15	140882	si	0.19648318	0.19648316	0.04	0	0.85714286	0.04733728	0.13043478	0	0.5	1	1	0	0.07692308	0	0.04	0.03448276	
16	146557	si	0.19648318	0.19648316	0.04	0	0.14705882	0.20710059	0.13043478	0	0.5	1	0.22255682	0.22255682	0.07692308	0.1	0.15727891	0.06896552	0.08
17	147339	si	0.19648318	0.19648316	0.04	0	0.92307692	0.08284024	0.13043478	0	0.5	1	0.26007293	0	0.11538462	0	0	0	0.04
18	147497	si	0.19648318	0.19648316	0.04	0	0.58333333	0.14792899	0.13043478	0	0.5	1	9.07E-05	0.00210438	0.30769231	0.4	0.47020408	0	
19	147910	si	0.19648318	0.19648316	0.04	0	0.10416667	0.28994083	0.13043478	0	0.5	1	0.26007293	0.15273116	0.11538462	0.2	0.1709833	0.03448276	0.04
20	147998	si	0.19648318	0.19648316	0.04	0	0.375	0.10059172	0.13043478	0	0.5	1	0.26007293	0.32951762	0.11538462	0.1	0.06965986	0.03448276	
21	148508	si	0.19648318	0.19648316	0.04	0	0.11267606	0.4260355	0.13043478	0	0.5	1	0.26007293	0.22856158	0.11538462	0.1	0.16837823	0.03448276	0.13
22	148842	si	0.19648318	0.19648316	0.04	0	0.2	0.09467456	0.13043478	0	0.5	1	0.04624204	0.00010885	0.11538462	0.6	0.18367347	0	0.04
23	151126	si	0.19648318	0.19648316	0.04	0	0.375	0.10059172	0.13043478	0	0.5	1	0.26007293	0.32951762	0.11538462	0.1	0.06965986	0.03448276	
24	151728	si	0.19648318	0.19648316	0.04	0	0.66666667	0.14792899	0.13043478	0	0.5	1	0.11965967	0.00417249	0.03846154	0.35	0.15673469	0.10344828	
25	152175	si	0.19648318	0.19648316	0.04	0	0.4	0.09467456	0.13043478	0	0.5	1	0.26007293	0.32951762	0.11538462	0.1	0.09183674	0.03448276	
26	152968	si	0.19648318	0.19648316	0.04	0	0.52380952	0.13017752	0.13043478	0	0.5	1	0.26007293	0.0006168	0.11538462	0.5	0.07561225	0	
27	153227	si	0.19648318	0.19648316	0.04	0	0.16129032	0.18934911	0.13043478	0	0.5	1	0.22255682	0.32951762	0.07692308	0.1	0.3922449	0.03448276	0.04
28	153454	si	0.19648318	0.19648316	0.04	0	0.92307692	0.08284024	0.13043478	0	0.5	1	0.26007293	0	0.11538462	0	0	0	0.04

Elaborado por: Diana Aroca y Diego Bernal
Fuente: Propia

Lo que se muestra son las características que empiezan desde 0 hasta 1023 y reinicia el conteo hasta generar la misma cantidad



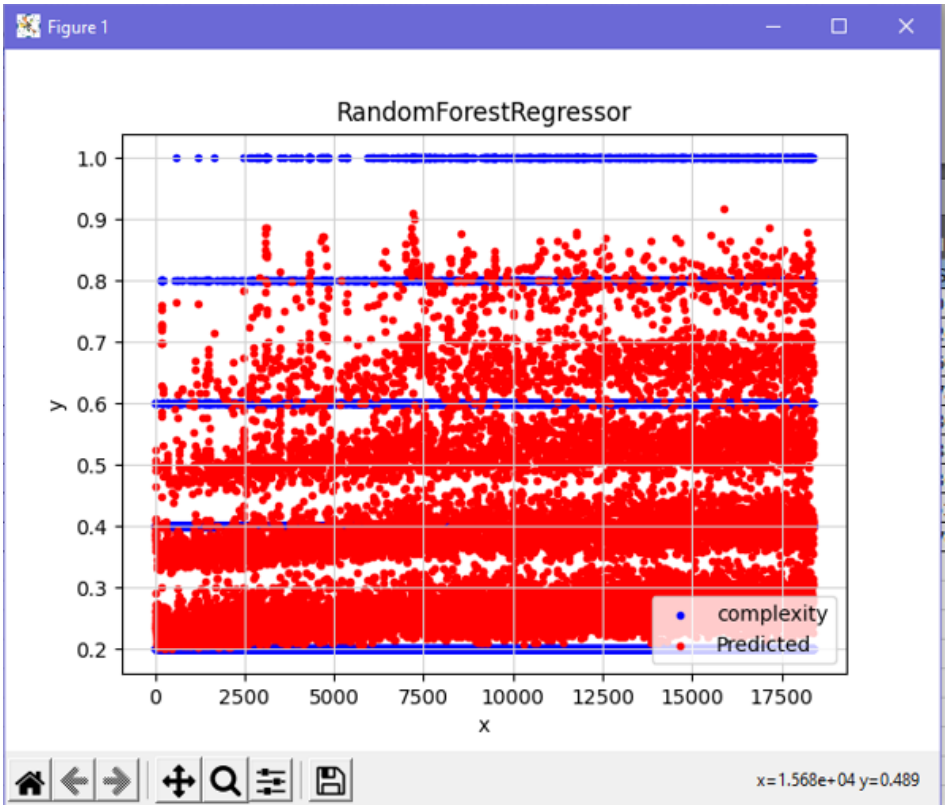
Elaborado por: Diana Aroca y Diego Bernal
Fuente: Propia

Luego procedemos a utilizar los Datasets del modelo pre-entrenado y del modelo ajustado con Fine-Tuning a pasarlos los diferentes algoritmos de Machine Learning entrenándolos y así obteniendo los niveles de predicción.

M.A.E	M.S.E	R.M.S.E	PEARSON	#CRT	ALGORITMOS	NOMBRE
0.16308555308402606	0.043557177553162636	0.20704193155948306	0.4883315006602684	23-1024-1024-Rlnb	SVR	SUPER VECTOR REGRESSOR
0.16860804094672288	0.04323588771905725	0.20672926707213105	0.6854501006254142	23-1024-1024-Rlnb	GBR	GRADIENT BOOSTING REGRESSOR
0.1689049442736826	0.05588036325652185	0.23516513294148375	0.9951151321218442	23-1024-1024-Rlnb	ABR	ADABOOST REGRESSOR
0.17717982170036964	0.0512804522722331	0.2253668315321251	0.5947089865485775	23-1024-1024-Rlnb	KNR	KNEIGHBORS REGRESSOR
0.1755826387092477	0.044910209923648504	0.21089589828270824	0.9806904917302182	23-1024-1024-Rlnb	RFR	RANDOM FOREST REGRESSOR
0.21740595781691674	0.0940226136116547	0.3060424907774306	0.99720864028381	23-1024-1024-Rlnb	DT	DECISION TREE
0.21045717908104386	0.08063501158072876	0.2823358611388518	0.3264032047695949	23-1024-1024-Rlnb	PAR	PASSIVE AGGESSIVE REGRESSOR
1134261338.8801868	2.1968964453138328e+18	1480304250.6064212	0.06184697235000098	23-1024-1024-Rlnb	SGR	STOCHASTIC GRADIENTE REGRESSOR
0.185958216218824	0.06423810146993243	0.2524529830068898	0.505022154274433	23-1024-1024-Rlnb	RLM	REGRESSOR LINEAR M

Elaborado por: Diana Aroca y Diego Bernal
Fuente: Propia

Valores clasificados por M.A.E, M.S.E, R.M.S.E y Pearson con sus respectivos algoritmos



Elaborado por: Diana Aroca y Diego Bernal
Fuente: Propia

Por último, de la ejecución de los algoritmos dando los valores de predicción de la complejidad de las palabras, se crea una representación gráfica del algoritmo como ejemplo presente (RFR)Random Forest Regresor donde se puede observar la relación que existe entre los valores reales (color azul) y los datos predichos por el sistema (color rojo) según la segunda propuesta experimental

Transformers for Lexical Complexity Prediction in Spanish Language

Transformers para la predicción de la complejidad léxica en lengua española

Resumen: En este artículo hemos presentado una contribución para predecir la complejidad de palabras simples en lengua española combinando un gran número de rasgos de diferente tipo. Después de las múltiples experimentaciones que realizamos, nos permitió conocer el rendimiento máximo para las diferentes combinaciones de los conjuntos de datos aplicando la regresión. Obtuvimos los resultados tras la ejecución de varios modelos de aprendizaje automático basados en Transformers previamente ajustados y ejecutados sobre los modelos pre-entranados BERT, XLM-RoBERTa, y RoBERTa-large-BNE en los diferentes conjuntos de datos en español. Se logró un mejor desempeño lo que llevó a un puntaje de un MAE de 0.1598 y un Pearson de 0.9883 logrado con la evaluación y entrenamiento del algoritmo Random Forest Regressor para el modelo transformador BERT afinado. Como posible propuesta alternativa para lograr un mejor rendimiento, estamos muy interesados en continuar probando el rendimiento de LCP en conjuntos de datos para España probando modelos de transformadores de última generación.
Palabras clave: Predicción de Complejidad Léxica; incrustaciones de palabras; Características lingüísticas; Transformadores

Abstract: In this article we have presented a contribution to predict the complexity of words in the Spanish language by combining a large number of features of different types. After the multiple experimentations, it allowed us to know the performance for the different combinations of the data sets applying the regression. We obtained the results after the execution of several machine learning models based on previously adjusted Transformers and executed on the pre-trained models BERT, XLM-RoBERTa, and RoBERTa-large-BNE. Better performance was achieved leading to a MAE score of 0.1598 and a Pearson score of 0.9883 achieved with the evaluation and training of the Random Forest Regressor algorithm for the tuned BERT transformer model. As a possible alternative proposal to achieve better performance, we are very interested in continuing to test the performance of LCP on datasets for Spain by testing state-of-the-art transformer models.
Keywords: Predicción de Complejidad Léxica; incrustaciones de palabras; Características lingüísticas; Transformadores

1 Introduction

A more common assumption is that people who are familiar with the vocabulary of a text can often understand the meaning of the words, even if they have difficulty with grammatical structures. Automatic lexical simplification can then become an effective method of making the text accessible to different audiences (Uluslu, 2022).

Predicting which words are considered difficult for a particular audience to understand is commonly known as complex word identification (CWI) (Shardlow, Cooper, y Zampieri, 2020). The task of detecting in the content of the documents the words that are difficult or complex to understand by the people of a given group is known as CWI - Complex Word Identification (Rico-Sulayes, 2020) and

it is a task that constitutes a fundamental step in many applications related to natural language, such as Text Simplification.

Deep learning and its revolutionary technology constitute the new state of the art in various Natural Language Processing (NLP) tasks (Singh y Mahmood, 2021), in which lexical complexity prediction (LCP) is no exception (Nandy et al., 2021). It is important to point out that, after the comparison and analysis of other approaches versus deep learning approaches, a viable path of possible solutions is generated for low-resource languages, where deep models are not always available or work as well as those of deep learning. English language. Likewise, it should be taken into account that the computational requirements for the application of deep learning models turn out to be significantly higher compared to those used in traditional approaches (Bender et al., 2021).

We propose our approach aimed at predicting the complexity score for single words in the Spanish language, since resources are scarce and are not as numerous as those available for the English language. Our model leverages the combination of advanced NLP techniques of deep learning models based on transformers: BERT (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2019), RoBERTa-large-BNE (Gutiérrez-Fandiño et al., 2021) and pre-trained word embeddings together with a set of textual complexity features made by hand (hand-crafted). For this, we use the corpus in Spanish CLexIS^2 corpus proposed by (Zambrano y Montejo-Ráez, 2021). we challenge is the application of a fine-tuned model on a previously trained model to improve the prediction of lexical complexity, for which, we follow the work done by (Rojas y Alva-Manchego, 2021).

The models used achieve a good performance shown in the results, with a MAE 0.1592 and a correlation of Person 0.9883 for the identification of a single word.

2 Related Work

In past decades, the application of very simple metrics such as calculating the number of syllables in words (Mc Laughlin, 1969) or verifying whether the word was part of a specific list to classify it as easy or complex (Dale y Chall, 1948) were the techniques that were applied in text legibility tasks.

Then the systems based on the characterization of words (using contextual, lexical and semantic characteristics) and the application of a Random Forest classifier (Breiman, 2001) to determine whether a word is complex or not are presented. A total of 45 handwritten features were computed in these systems, and each word was modeled as a feature vector. Surface functions (three functions), dependency tree functions (eight functions), Corpus-based functions (fifteen functions), WordNet functions (eleven functions), and WordNet and corpus frequency functions (four functions) were applied. The best result obtained was a precision value of 0.186, a recall of 0.673, a G-score of 0.750 99 and an F-score of 0.292.

The investigations in the last years are directed to the Identification of Complex Words - CWI. The objective of these applications is to be able to predict the complexity of words based on the construction of their features, as exposed in the work carried out by (Shardlow, Cooper, y Zampieri, 2020) presenting their approach on a set of features of word embeddings from Glove, InferSent, and various linguistic features obtained as predictive sources of lexical complexity, such as word frequency, word length, or number of syllables. Then, they trained a linear regression model using different subsets of functions, obtaining as a result an MAE of 0.0853.

A system for predicting word complexity (Ortiz-Zambrano y Montejo-Ráez,) was developed for the shared LCP task hosted on SemEval 2021 (Shardlow et al., 2021) where task organizers distributed to participants the CompLex corpus (Shardlow, Cooper, y Zampieri, 2020) but in its augmented version. The task was located on the Lexical Semantics track, which consisted of predicting the complexity value of words in context. Carrying out a machine learning approach that was based on 15 linguistic features obtained at the word level and their environment (Ortiz-Zambrano y Montejo-Ráez,) trained a supervised random forest regression algorithm on the set of features. Several runs were made with different values to observe the performance of the algorithm. The best results achieved were a MAE of 0.07347, an MSE of 0.00938 and an RMSE of 0.096871.

The field of NLP has shown incredible progress in the last two years, this is particularly due to the Transformer architecture (Vaswa-

ni et al., 2017) that takes advantage of large amounts of unlabeled text corpus (Canete et al., 2020). Deep learning models show significant improvement over shallow machine learning models with the rise of transfer learning and pretrained language models. The deep learning pretrained language models, BERT and XLM-RoBERTa, are considered state-of-the-art in many NLP tasks (Yaseen et al., 2021).

In our approach, we review the use of word embeddings from the pre-trained and fine-tuned models, and compare them to a broader list of linguistic features at the lexical level. Our objective is to provide an exhaustive evaluation that shows more clearly, the executions carried out on several different data sets in the Spanish language, how the lexical characteristics together with the deep encodings contribute to the prediction of lexical complexity.

3 Materials and Method

This section briefly details about the pretrained models and their application for the generation of encodings at both the sentence and word levels. Likewise, the data sets that have been used in the different experiments are presented. Finally, the different classification algorithms and the applied characteristics are shown.

3.1 Dataset

The **CLexIS**[2] corpus was elaborated with the transcripts of the recorded classes of the professors of the careers of Computer Systems Engineering and Software Engineering, two careers that belong to the Faculty of Mathematical Sciences of the University of Guayaquil (Ecuador) (Zambrano y Montejo-Ráez, 2021).

CLexIS[2] has become a resource of great interest and importance, due to the fact that there are few resources in Spanish available for NLP researchers, and some of them do not usually contain annotations that facilitate the development of NLP models (Davidson et al., 2020). For its construction, the collection presented in the ALexS 2020 competition at IberLEF 2020 (Ortiz-Zambranoa y Montejo-Ráezb, 2020) was taken as a reference.

Table 1 shows some statistics on different type of words present in the CLexIS[2] dataset.

3.2 Transformer-based Language Models

The models were taken from the *Transformers*[1] library.

- The pre-trained **BERT** model that we chose was the one that the Spanish community uses mostly in research work to date, which is bert-base-uncased (BETO) (Canete et al., 2020).

  BERT-base model has the number of layers L=12, the hidden size H=768, the number of self-attention heads A=12,and Total Parameters=110M.

  BERT-large model has the number of layers=24, the hidden size=1024, the number of self-attention heads=16, and Total Parameters=335M (Conneau et al., 2019).

- The **XLM-RoBERTa** model applied was xlm-roberta-base (Conneau et al., 2019).

  The **XLM-RoBERTa-base** model has the number of layers L=12, the hidden size H=768, the number of self-attention heads A=12,and Total Parameters=270M.

  XLM-RoBERTa-large model has the number of layers L=24, the hidden size H=1024, the number of self-attention heads A=16, and Total Parameters=550M (Conneau et al., 2019).

- The **RoBERTa-large-BNE** model used was PlanTL-GOB-ES/roberta-large-bne being the largest Spanish-specific model to date (Gutiérrez-Fandiño et al., 2021).

  XLM-RoBERTa-large is a transformer-based masked language model for the Spanish language. It is based on the RoBERTa large model [2].

  The **RoBERTa-large-BNE** model has the number of layers L=24, the hidden size H=1024, the number of self-attention heads A=16,and Total Parameters=335M(Conneau et al., 2019).

3.3 Experiments design

Our purpose is to demonstrate how the combination of different types of features con-

[1]https://huggingface.co/

[2]https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne

Number of	Count
Sum. of content words (verbs, adjectives and nouns)	153,885
Different content words	200,785
Rare words (low frequency in CREA corpus [31])	143,464
Sentences	9,756
Complex sentences	4,101
Total Words	300,420

Tabla 1: Volumetrics for CLexIS[2]

tribute to better performance (Zaharia, Cercel, y Dascalu, 2021), (Paetzold, 2021). We will refer to the linguistic features - *hand-crafted* (HCF) versus the Transformers encodings coming from the models: BERT in Spanish, XLM-RoBERTa multilingual, and RoBERTa-large-BNE.

Next, we applied a fitted model on top of the previously trained model to demonstrate how the fitted model run on the previously trained model contributed to more accurate LCP results.

Finally, the execution of the different supervised learning algorithms was applied on the training data set to evaluate which of them achieved the best prediction score. A three-fold cross-validation was performed to ensure that the partitions contained independent data for training and testing. See fig.1

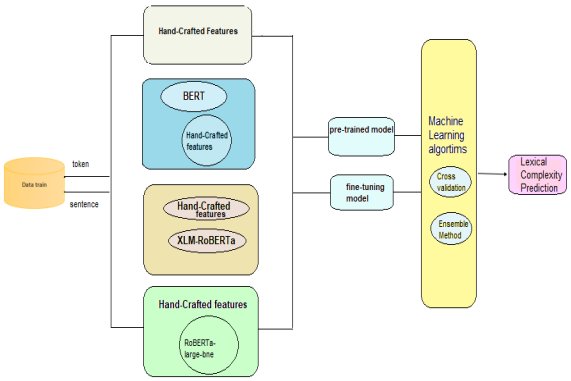


Figura 1: Representation of the workflow of the process for LCP

3.3.1 Features

- **Hand-Crafted Features - HCF**

To obtain the morphological aspects of the text, we perform several experiments applying a total of 23 linguistic features and combine them with the word and sentence embeddings of previously trained deep learning models.

we have considered the 15 HCF proposed by (Ortiz-Zambrano y Montejo-Ráez, 2021) and added a sets of features computed from POS categories counts (Vettigli y Sorgente, 2021), (Liebeskind, Elkayam, y Liebeskind, 2021), giving a total of 23 hand-crafted features. It was necessary to use the Spacy library together with the model in core web sm to extract these features. All these features were normalized with a z-score transformation before passing them to the learning algorithm.

1. *Absolute frequency*: the absolute frequency.
   The frequency of words is a measure that serves as an indicator of lexical complexity. If in common parlance a word occurs frequently, it is more likely to be recognized (Rayner y Duffy, 1986) and (Shardlow, Cooper, y Zampieri, 2020).

2. *Relative frequency*: the relative frequency of the target word.

3. *Word length*: the number of characters of the token. The length of the word was calculated in number of its characters. It is often the case that longer length words are more difficult to process and can therefore be considered *complex*. (Shardlow, 2013) (Shardlow, Cooper, y Zampieri, 2020) (Paetzold y Specia, 2016).

4. *Number of syllables*: the number of syllables. A good estimate of complexity complexity is the number of syllables contained in a word (Shardlow, 2013) (Ronzano et al., 2016) (Shardlow, Cooper, y Zampieri, 2020) (Paetzold y Specia, 2016).

5. *Target word position* (token-position): the position of the target word in the sentence. Position of the word (WordPosition) (Shardlow, 2013) (Ronzano et al., 2016).

6. *Number of words in the sentence* : number of words in sentence. Words in sentence (NumSentenceWords) (Shardlow, 2013) (Ronzano et al., 2016).

   Based on the work proposed by (Ronzano et al., 2016) in the exploring linguistic features for lexical complexity prediction.

7. *Part Of Speech (POS)*: the Part Of Speech category.
8. *Relative frequency of the previous the token*: the relative frequency of the word before the token.
9. *Relative frequency of the word after the token*: the relative frequency of the word after the token.
10. *Length of previous word*: the number of characters in the word before the token.
11. *Length of the after word*: the number of characters in the word after the token.
12. *Lexical diversity - MTDL*: the lexical diversity of the target word in the sentence.

    Additionally, the following WordNet features were also considered for each target word, as in the works carried out by (Gooding y Kochmar, 2018):

13. *Number of synonyms.*
14. *Number of hyponyms.*
15. *Number of hyperonyms.*

We follow the recommendations of (Paetzold y Specia, 2016), (Ronzano et al., 2016), (Gooding y Kochmar, 2018), (Liebeskind, Elkayam, y Liebeskind, 2021), (Desai et al., 2021) with the aim of improving results, generating 8 new features originating from the POS, which were:

1. PROPN - Number of pronouns within the sentence.
2. AUX - Number of auxiliaries within the sentence.
3. VERB - Number of verbs within the sentence.
4. ADP - Number of adverbs within the sentence.
5. NOUN - Number of nouns within the sentence.
6. NN - Number of Nouns, singular or massive.
7. SYM - Number of symbols within the sentence.
8. NUM - Number of numbers within the sentence.

- **BERT vector:** The bert-base-uncased model from the Hugging Face transformer library (Wolf et al., 2020) was applied. We take all the 768-dimensional numerical representation produced by the pre-trained BERT model (Devlin et al., 2018) in the different combinations of sentence and target word encodings, for both the pre-trained model and the model fine-tuned, reaching a total of 1559 linguistic characteristics of different nature.
- **XLM-RoBERTa vector:** As in the case of the BERT model, we take all the 768-dimensional numerical representation produced by the pre-trained RoBERTa model (Conneau et al., 2019) in the different combinations of sentence and target word encodings, for both the pre-trained model and the model fine-tuned, reaching a total of 1559 linguistic characteristics of different nature.
- **RoBERTa-large-BNE vector:** Regarding this model, we take all the 1024-dimensional numerical representation produced by the pre-trained RoBERTa-large-model model (Gutiérrez-Fandiño et al., 2021), in the same way that they were applied in the previous models, the data sets were made up of for the different combinations of sentence and target word encodings, for both the pre-trained model and the fine-tuned model, reaching a total of 2071 linguistic characteristics of different nature.

3.3.2 Machine Learning Algorithms

similar to the work done by (Zaharia, Cercel, y Dascalu, 2021) in the case of the algorithms, the training and evaluation of the different combinations of the sets was carried out with a total of eight supervised algorithms for the regression, these are:

1. AdaBoost **(AB)** (Paetzold, 2021).
2. Desicion Tree **(DT)** (Shardlow, Evans, y Zampieri, 2021).

3. Gradient Boosting **(GB)** (Vettigli y Sorgente, 2021).
4. Stochastic Gradient **(SG)** (Bottou, 2010).
5. Nearest Neighbors **(KNN)** (Liebeskind, Elkayam, y Liebeskind, 2021).
6. Support Vector Machines **(SVM)** (Liebeskind, Elkayam, y Liebeskind, 2021),
7. Passive Aggressive **(PA)** (Crammer et al., 2006).
8. Random Forest (RF) (Zaharia, Cercel, y Dascalu, 2021), (Desai et al., 2021).

Several experiments were carried out for each of the data sets, where different configurations were explored for each of the algorithms. We apply the default values for the algorithms in their configurations, except for the case for tree-based algorithms where we manage to determine the best hyperparameters with the following number of nodes:

- AdaBoost with 100 nodes.
- Random forest with 241 nodes.
- Gradient Boosting algorithm with 350 nodes.

4 Results

4.1 Features Sets

We build several datasets made up of the combination of the features described above to run them on the pre-trained BERT model. Table 2 contains the description of the abbreviations.

The detail of the features:

- The *Hand-Crafted Features* with the features coming from the 768-dimensional vector of the initial [CLS] token as sentence embeddings ($BERT_{sent}$).
- *The Hand-Crafted Features* with the 768-dimensional vector corresponding to the target token as word embeddings ($BERT_{word}$).
- The *Hand-Crafted Features* with encodings of the [CLS] token and encodings of the target token.
- The encodings of the [CLS] token.
- The encodings of the target token.
- The encodings of the [CLS] token with the encodings of target token.

For the evaluation of the trained and fine-tuned models, those that were widely applied to LCP for the shared LCP task hosted in SemEval 2021: Mean Absolute Error (MAE), Mean Square Error (MSE), Mean Square Error (RMSE) and Pearson correlation (Shardlow et al., 2021).

4.2 BERT model *pre-trained*

We present the ten best performances corresponding to different combinations of features described in section 4.1 executed with BERT pre-trained. See Table 4.2.

As we can see, the three best results in predicting lexical complexity were achieved by the ABR and SVR algorithms. The best performance was achieved by the ABR - AdaBoost algorithm presenting the best prediction for the Spanish language with a MAE of 0.1632 and a Pearson of 0.999 in the execution with the data set made up of the combination of the features generated at the sentence level and at the word level - BERT$_{sent}\oplus$ BERT$_{word}$.

4.3 BERT model *fine-tuned*

We have applied the fine-tuned BERT model on top of the pre-trained BERT model for the purpose of the results. It can be seen in the table 4 the three best results resulting from the performance of the RFR - Random Forest Regressor algorithm and the GBR - Gradient Boosting Regressor algorithm.

The best performance was obtained with the dataset composed of the combination of the features with target word encodings together with sentence encodings from BERT fine-tuned. The same combination of features achieved the best performance in the pre-trained model but with lower results.

It should be noted that the RFR algorithm does not appear within the top ten places in the execution of the pre-trained model, but it achieves its best result when the model is refined, placing first and third within the three best executions tuned. RFR presented the best prediction for the Spanish language with a MAE of 0.1592 and a Pearson of 0.988 combining BERT$_{sent}\oplus$ BERT$_{word}$.

Features identifier	Description
HCF	Hand-crafted (linguistic) features
BERT_{sent}	Sentence encodings from pre-trained BERT models
BERT_{word}	Token encodings from pre-trained BERT models
XLMR_{sent}	Sentence encodings from pre-trained XLM-RoBERTa model
XLMR_{word}	Token encodings from pre-trained RoBERTa model
RBNE_{sent}	Sentence encodings from pre-trained RoBERTa-large-BNE model
RBNE_{word}	Token encodings from pre-trained RoBERTa-large-BNE model

Tabla 2: Feature sets

BERT *model pre-trained* with CLexIS^2					
Features	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
$\text{BERT}_{sent}\oplus$ BERT_{word}	**ABR**	**0,1632**	**0,0502**	**0,2343**	**0,9999**
BERT_{word}	SVR	0,1634	0,0432	0,2074	0,9023
BERT_{word}	ABR	0,1643	0,0512	0,2332	0,9947
BERT_{word}	GBR	0,1653	0,0494	0,0447	0,6977
$\text{BERT}_{word}\oplus$ HFC	GBR	0,1655	0,0454	0,2088	0,7040
$\text{BERT}_{sent}\oplus$ $\text{BERT}_{word}\oplus$ HFC	GBR	0,1659	0,0418	0,2074	0,7147
$\text{BERT}_{sent}\oplus$ BERT_{word}	GBR	0,1694	0,0444	0,2039	0,7167
$\text{BERT}_{sent}\oplus$ $\text{BERT}_{word}\oplus$ HFC	ABR	0,1699	0,0554	0,2334	0,9939
$\text{BERT}_{sent}\oplus$ BERT_{word}	SVR	0,1699	0,0488	0,2037	0,8799
$\text{BERT}_{word}\oplus$ HFC	SVR	0,1704	0,0503	0,2204	0,0483

Tabla 3: Results of the model BERT pre-trained with features of different nature

4.4 XLM-RoBERTa model *pre-trained*

Similar to the BERT model, the top ten sites were taken from all the runs that were done on the different data sets. The results of the three best places for the pre-trained XLM-RoBERTa model were achieved by the ABR - AdaBoots algorithm with a MAE 0.1623 and a Pearson 0.9973 result of the combination of the features with target word encodings together with sentence encodings from BERT fine -tuned and the HCF - $\text{XLMR}_{sent}\oplus$ $\text{XLMR}_{word}\oplus$ HFC, as can be seen in the table 5.

It can be clearly shown that the pre-trained XLM-RoBERTa model has a better performance compared to the pre-trained BERT model, achieving a better prediction of Lexical Complexity.

4.5 XLM-RoBERTa model *fine-tuned*

We also highlight that in the execution of the XLM-RoBERTa tuned model, it achieved a significant improvement compared to the results of the pre-trained model, reaching a MAE 0.1601 and a Pearson 0.998 result of the combination of the features with target word encodings together HCF - $\text{XLMR}_{word}\oplus$ HFC from XLMR fine-tuned. See table 6.

Comparing the results of the BERT and XLM-RoBERTa both tuned models, BERT tuned is so far the one that has an important performance achieved by a MAE 0.1592 with the execution of the RFR algorithm combining $\text{XLMR}_{sent}\oplus$ XLMR_{word}.

4.6 RoBERTa-large-BNE model *pre-trained*

The novelty of the models for Spanish is to have incorporated in this research the executions with the pre-trained model RoBERTa-large-BNE and also the adjusted model.

The ten best results are displayed in the table 8. The three best positions in predicting lexical complexity were achieved by the ABR-AdaBoost and GBR-Gradient Boost Regressor algorithms, with ABR reaching a MAE 0.1609 and a Person 0.6754 combining the sentence encodings together with the target word encodings and the HCF - $\text{RBNE}_{sent}\oplus$ $\text{RBNE}_{word}\oplus$ HFC.

It should be noted that the pre-trained model RoBERTa-large-BNE is the one that achieves a better prediction for lexical complexity in the Spanish language compared to the pre-trained models BERT and XLM-RoBERTa, reaching BERT a MAE of 0.1632, XLMR a MAE of 0.1623 and RBNE a MAE of 0.1609. See Table 7.

BERT *model fine-tuned* with CLexIS^2					
Features	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
BERT$_{sent}\oplus$ **BERT**$_{word}$	**RFR**	**0,1592**	**0,0379**	**0,1982**	**0,9883**
BERT$_{sent}\oplus$ **BERT**$_{word}\oplus$ **HFC**	GBR	0,1600	0,0367	0,1979	0,8202
BERT$_{sent}\oplus$ **BERT**$_{word}\oplus$ **HFC**	RFR	0,1610	0,0401	0,1988	0,9987
BERT$_{sent}\oplus$ **BERT**$_{word}\oplus$ **HFC**	ABR	0,1610	0,0506	0,2242	0,9998
BERT$_{sent}\oplus$ **BERT**$_{word}\oplus$ **HFC**	ABR	0,1621	0,0487	0,2300	0,9999
BERT$_{word}\oplus$ **HFC**	GBR	0,1622	0,0430	0,1984	0,8983
BERT$_{word}$	SVR	0,1622	0,0429	0,2018	0,9183
BERT$_{word}$	GBR	0,1632	0,0472	0,0429	0,7083
BERT$_{word}$	ABR	0,1632	0,0501	0,2284	0,9967
BERT$_{word}\oplus$ **HFC**	RFR	0,1640	0,0407	0,2100	0,9898

Tabla 4: Results of the model BERT tuned with features of different nature

XLM-RoBERTA *model pre-trained* with CLexIS^2					
Features	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
XLMR$_{sent}\oplus$ **XMLR**$_{word}\oplus$ **HFC**	ABR	0,1623	0,0527	0,2270	0,9973
XLMR$_{word}\oplus$ **HFC**	ABR	0,1623	0,0513	0,2273	0,9973
XLMR$_{sent}\oplus$ **HFC**	ABR	0,1630	0,0524	0,2293	0,9973
XLMR$_{word}\oplus$ **HFC**	GBR	0,1653	0,0433	0,2073	0,4848
XLMR$_{sent}\oplus$ **XMLR**$_{word}\oplus$ **HFC**	GBR	0,1658	0,0434	0,2074	0,4874
XLMR$_{sent}\oplus$ **HFC**	GBR	0,1663	0,0433	0,2082	0,4807
XLMR$_{word}\oplus$ **HFC**	SVR	0,1680	0,0483	0,2194	0,3095
XLMR$_{word}\oplus$ **HFC**	RFR	0,1690	0,0445	0,2093	0,9803
XLMR$_{sent}\oplus$ **HFC**	SVR	0,1692	0,0486	0,2204	0,2713
XLMR$_{sent}\oplus$ **XMLR**$_{word}\oplus$ **HFC**	RFR	0,1692	0,0443	0,2093	0,9803

Tabla 5: Results of the model XLMR pre-trained with features of different nature

4.7 RoBERTa-large-BNE model *fine-tuned*

Executing the RoBERTa-large-BNE tuned model, the results are encouraging, there is an improvement compared to the results of the pre-trained model. The first places were reached by the GBR-Gradient Boosting Regressor and SVR-Super Vector Regressor algorithms. It presents an acceptable improvement, achieving in its performance a MAE 0.1609 and a Pearson 0.6754 combining the sentence encodings together with the target word encodings and the HCF - RLBNE$_{sent}\oplus$ RLBNE$_{word}\oplus$ HFC, and the second and third places prove it in comparison with the pre-trained model. See Table 10.

It should be noted that the pre-trained model RoBERTa-large-BNE is the one that achieves a better prediction for lexical complexity in the Spanish language compared to the tuned models BERT and XLM-RoBERTa, reaching BERT a MAE of 0.1592, XLM-R a MAE of 0.1601 and RBNE a MAE of 0.1609. See Table 9.

It can be seen that the models based on adjusted Transformers make an important contribution to the Prediction of Lexical Complexity in the Spanish language. The Table 11 presents the best five best results of all the experiments carried out with the models both pre-trained and adjusted, where the adjusted models are those that achieved the best performance.

The BERT tuned model with the combination of the features executed on the RFR-Random Forest Regressor algorithm reached first place with a MAE 0.1592 and a Person 0.9883, the feature are RBNE$_{sent}\oplus$ RBNE$_{word}$.

It is important to mention that the set of manual features whose description we have described in section 3.3.1 make a very important contribution to the Word Complexity Prediction, since they are only based on the frequency of the word and various calculations. manually, they achieved a MAE 0.1641 and a Perarson 0.5154. See Table 12.

5 Discussion

We have presented the contribution of the BERT, XLM-RoBERTa and RoBERTa-large-BNE models for the Spanish language executed on the CLexIS^2 corpus, combining features of different nature among them, the Hand-Crafted features and those originating from combining the sentence encodings to-

XLM-RoBERTA *model fine-tuned* with CLexIS^2					
Features	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
XLMR$_{word}\oplus$ **HFC**	ABR	0,1601	0,0501	0,2251	0,9987
XLMR$_{sent}\oplus$ **XMLR**$_{word}\oplus$ **HFC**	ABR	0,1620	0,0526	0,2268	0,9987
XLMR$_{sent}\oplus$ **HFC**	ABR	0,1620	0,0519	0,2287	0,9979
XLMR$_{sent}\oplus$ **HFC**	GBR	0,1630	0,0420	0,2062	0,4790
XLMR$_{word}\oplus$ **HFC**	GBR	0,1638	0,0429	0,2034	0,4800
XLMR$_{sent}\oplus$ **XMLR**$_{word}\oplus$ **HFC**	GBR	0,1652	0,0430	0,2069	0,4930
XLMR$_{word}\oplus$ **HFC**	SVR	0,1660	0,0482	0,2172	0,3083
XLMR$_{word}\oplus$ **HFC**	RFR	0,1669	0,0427	0,2013	0,9849
XLMR$_{sent}\oplus$ **HFC**	SVR	0,1682	0,0487	0,2202	0,2729
XLMR$_{sent}\oplus$ **XMLR**$_{word}\oplus$ **HFC**	SVR	0,1688	0,0482	0,2191	0,2932

Tabla 6: Results of the model XLMR tuned with features of different nature

The *Spanish Language Models* Pre-Trained **Best Result**			
Model	**Features**	**Alg**	**MAE**
RBNE	RBNE$_{sent}\oplus$ RLBNE$_{word}\oplus$ HFC	ABR	0.1609
XLMR	XLMR$_{sent}\oplus$ XLMR$_{word}\oplus$ HFC	ABR	0.1623
BERT	BERT$_{sent}\oplus$ BERT$_{word}$	ABR	0.1632

Tabla 7: Best results models pre-trained

gether with the target word encodings and the HCF.

The BERT and ROBERTa models were widely used in the LCP competition for the shared LCP task hosted in SemEval 2021 (Shardlow et al., 2021), reaching very important results demonstrating a relevant contribution in the Prediction of Lexical Complexity, which demonstrates a valuable contribution this investigation.

We observe that according to the results of the final evaluation, especially in terms of refinement, the refined Spanish language models made an important contribution to the prediction of lexical complexity by outperforming the proposal presented after the execution of the manual features-HCF. In the case of the RoBERTa-large-BNE model, we have found a performance that exceeds the rest of the models after the execution of the pretrained model and even remains within the three best executions in the results of the tuned models such as the proposals presented. by (Gutiérrez-Fandiño et al., 2021)

6 Conclusion and Further Work

In this article, we have presented a contribution to predict the complexity of simple words in the Spanish language, combining a large number of features of different types.

We consider that, after the multiple experimentations that we carried out, it allowed us to know the maximum performance for the different combinations of the data sets by applying the regression.

In our experiments, we obtained the results after the execution of several previously trained transformer-based machine learning models on several datasets in Spanish combining features of different nature. The application of the adjusted model in the trained model achieved a better performance, which led to a MAE score of 0.1598 and a Pearson of 0.9883 achieved with the evaluation and training of the Random Forest Regressor algorithm for the transformer model. BERT fine-tuned.

As a possible alternative proposal to achieve better performance, we are very interested in continuing to test performance for LCP on datasets for Spain by testing state-of-the-art transformer models.

Acknowledgements

We appreciate Jostin Daniel Escobar Suarez, Joel Stalin Sorroza Contreras, Diego Gabriel Bernal Yucailla y Diana Geovanna Aroca Pincay graduate of the Computer Systems Engineering degree from the University of Guayaquil, for their valuable contribution to the development of our work.

RoBERTa-large-BNE *model pre-trained* with CLexIS^2					
Features	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HFC**	ABR	0,1609	0,0421	0,2047	0,6754
RBNE$_{sent}\oplus$ **RBNE**$_{word}$	ABR	0,1675	0,0556	0,2347	0,9952
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HFC**	GBR	0,1691	0,0434	0,2073	0,6607
RBNE$_{word}$	ABR	0,1693	0,0563	0,2360	0,9948
RBNE$_{word}$	GBR	0,1696	0,0447	0,2101	0,6400
RBNE$_{sent}\oplus$ **RBNE**$_{word}$	GBR	0,1698	0,0447	0,2102	0,6450
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HFC**	SVR	0,1708	0,0507	0,2224	0,2363
RBNE$_{sent}\oplus$ **HFC**	SVR	0,1708	0,0507	0,2224	0,0857
RBNE$_{word}$	SVR	0,1708	0,0507	0,2224	0,0390
RBNE$_{sent}$	SVR	0,1708	0,0507	0,2224	0,0052

Tabla 8: Results of the model RBNE pre-trained with features of different nature

The *Spanish Language Models* Fine-Tuned Best Result			
Model	**Features**	**Alg**	**MAE**
BERT	BERT$_{sent}\oplus$ BERT$_{word}$	RFR	0,1592
XLMR	XLMR$_{word}\oplus$ HFC	ABR	0,1601
RBNE	RBNE$_{sent}\oplus$ RBNE$_{word}\oplus$ HFC	GBR	0,1609

Tabla 9: Best results models fine-tuned

RoBERTa-large-BNE *model fine-tuned* with CLexIS^2					
Features	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HFC**	GBR	0,1609	0.0421	0.2047	0.6754
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HFC**	SVR	0,1630	0,0435	0,2070	0,4883
RBNE$_{word}\oplus$ **HFC**	SVR	0,1666	0,0466	0,2136	0,4220
RBNE$_{word}$	ABR	0,1677	0.0551	0,2336	0,9952
RBNE$_{sent}\oplus$ **RBNE**$_{word}$	SVR	0,1684	0.0472	0.2152	0.4425
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HFC**	GBR	0,1686	0,0432	0,2067	0,6854
RBNE$_{word}$	SVR	0,1686	0,0468	0,2146	0,5021
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HFC**	ABR	0,1689	0,0558	0,2351	0,9951
RBNE$_{word}$	GBR	0,1690	0,0442	0,2090	0,6585
RBNE$_{sent}\oplus$ **RBNE**$_{word}$	ABR	0,1699	0,0598	0,2391	0,9891

Tabla 10: Results of the model RBNE tuned with features of different nature

Summary of best results on the CLexIS^2 corpus						
Model	**Features**	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
BERT$_{fine-tuned}$	**BERT**$_{sent}\oplus$ **BERT**$_{word}$	RFR	0,1592	0,0379	0,1982	0,9883
BERT$_{fine-tuned}$	**BERT**$_{sent}\oplus$ **BERT**$_{word}\oplus$ **HFC**	GBR	0,1600	0,0367	0,1979	0,8202
XLMR$_{fine-tuned}$	**XLMR**$_{word}\oplus$ **HFC**	ABR	0,1601	0,0501	0,2251	0,9987
RBNE$_{fine-tuned}$	**RBNE**$_{sent}\oplus$ **RBNE**$_{word}$	GBR	0,1609	0,0421	0,2047	0,6754
BERT$_{fine-tuned}$	**BERT**$_{sent}\oplus$ **BERT**$_{word}\oplus$ **HFC**	RFR	0,1610	0,0401	0,1988	0,9987

Tabla 11: Summary of best results on the CLexIS^2 dataset

The *Hand-Crafted Features* HCF with CLexIS^2				
Alg	**MAE**	**MSE**	**RMSE**	**Pearson**
KNR	0.1641	0.0485	0.2190	0.5154
SVR	0.1708	0.0508	0.2225	0.1402
ABR	0.1732	0.0518	0.2266	0.9999

Tabla 12: Results of the execution of the HCF

Bibliografía

Bender, E. M., T. Gebru, A. McMillan-Major, y S. Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . En *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, página 610–623, New York, NY, USA. Association for Computing Machinery.

Bottou, L. 2010. Large-scale machine learning with stochastic gradient descent. En *Proceedings of COMPSTAT'2010*. Springer, páginas 177–186.

Breiman, L. 2001. Random forests. *Machine learning*, 45(1):5–32.

Canete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, y J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:2020.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, y V. Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Crammer, K., O. Dekel, J. Keshet, S. Shalev-Shwartz, y Y. Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.

Dale, E. y J. S. Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, páginas 37–54.

Davidson, S., A. Yamada, P. F. Mira, A. Carando, C. H. S. Gutierrez, y K. Sagae. 2020. Developing nlp tools with a new corpus of learner spanish. En *Proceedings of the 12th Language Resources and Evaluation Conference*, páginas 7238–7243.

Desai, A., K. North, M. Zampieri, y C. M. Homan. 2021. Lcp-rit at semeval-2021 task 1: Exploring linguistic features for lexical complexity prediction. *arXiv preprint arXiv:2105.08780*.

Devlin, J., M.-W. Chang, K. Lee, y K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Gooding, S. y E. Kochmar. 2018. 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*, páginas 184–194.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, y M. Villegas. 2021. Spanish language models. *arXiv preprint arXiv:2107.07253*.

Liebeskind, C., O. Elkayam, y S. Liebeskind. 2021. Jct at semeval-2021 task 1: Context-aware representation for lexical complexity prediction. En *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, páginas 138–143.

Liu, X., P. He, W. Chen, y J. Gao. 2019. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *arXiv preprint arXiv:1904.09482*.

Mc Laughlin, G. H. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Nandy, A., S. Adak, T. Halder, y S. M. Pokala. 2021. cs60075_team2 at semeval-2021 task 1: Lexical complexity prediction using transformer-based language models pre-trained on various text corpora. En *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, páginas 678–682.

Ortiz-Zambrano, J. y A. Montejo-Ráez. Sinai at semeval-2021 task 1: Complex word identification using word-level features.

Ortiz-Zambrano, J. A. y A. Montejo-Ráez. 2021. Complex words identification using word-level features for semeval-2020 task 1. En *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, páginas 126–129.

Ortiz-Zambranoa, J. A. y A. Montejo-Ráezb. 2020. Overview of alexs 2020: First workshop on lexical analysis at sepln. En *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*.

Paetzold, G. 2021. Utfpr at semeval-2021 task 1: Complexity prediction by combining bert vectors and classic features. En *Proceedings of the 15th Interna-*

tional Workshop on Semantic Evaluation (SemEval-2021), páginas 617–622.

Paetzold, G. y L. Specia. 2016. Sv000gg at semeval-2016 task 11: Heavy gauge complex word identification with system voting. En *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, páginas 969–974.

Rayner, K. y S. A. Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & cognition*, 14(3):191–201.

Rico-Sulayes, A. 2020. General lexicon-based complex word identification extended with stem n-grams and morphological engines. En *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR-WS, Malaga, Spain.*

Rojas, K. R. y F. Alva-Manchego. 2021. Iapucp at semeval-2021 task 1: Stacking fine-tuned transformers is almost all you need for lexical complexity prediction. En *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, páginas 144–149.

Ronzano, F., L. E. Anke, H. Saggion, y others. 2016. Taln at semeval-2016 task 11: Modelling complex words by contextual, lexical and semantic features. En *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, páginas 1011–1016.

Shardlow, M. 2013. A comparison of techniques to automatically identify complex words. En *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, páginas 103–109.

Shardlow, M., M. Cooper, y M. Zampieri. 2020. Complex: A new corpus for lexical complexity prediction from likert scale data. *arXiv preprint arXiv:2003.07008*.

Shardlow, M., R. Evans, G. H. Paetzold, y M. Zampieri. 2021. Semeval-2021 task 1: Lexical complexity prediction. *arXiv preprint arXiv:2106.00473*.

Shardlow, M., R. Evans, y M. Zampieri. 2021. Predicting lexical complexity in english texts. *arXiv preprint arXiv:2102.08773*.

Singh, S. y A. Mahmood. 2021. The nlp cookbook: Modern recipes for transformer based deep learning architectures. *IEEE Access*, 9:68675–68702.

Uluslu, A. Y. 2022. Automatic lexical simplification for turkish. *arXiv preprint arXiv:2201.05878*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, y I. Polosukhin. 2017. Attention is all you need. En *Advances in neural information processing systems*, páginas 5998–6008.

Vettigli, G. y A. Sorgente. 2021. Compna at semeval-2021 task 1: Prediction of lexical complexity analyzing heterogeneous features. En *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, páginas 560–564.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, y others. 2020. Transformers: State-of-the-art natural language processing. En *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, páginas 38–45.

Yaseen, T. B., Q. Ismail, S. Al-Omari, E. Al-Sobh, y M. Abdullah. 2021. Just-blue at semeval-2021 task 1: Predicting lexical complexity using bert and roberta pretrained language models. En *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, páginas 661–666.

Zaharia, G.-E., D.-C. Cercel, y M. Dascalu. 2021. Upb at semeval-2021 task 1: Combining deep learning and hand-crafted features for lexical complexity prediction. *arXiv preprint arXiv:2104.06983*.

Zambrano, J. A. O. y A. Montejo-Ráez. 2021. Clexis2: A new corpus for complex word identification research in computing studies. En *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, páginas 1075–1083.