PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR
SEDE ESMERALDAS



CARRERA:
INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

TEMA DE INVESTIGACIÓN:
SISTEMA PREDICTIVO BASADO EN TÉCNICAS DE MINERÍA DE TEXTO Y APRENDIZAJE AUTOMÁTICO PARA CLASIFICAR ESTILOS DE REDACCIÓN ACADÉMICA.

LÍNEA DE INVESTIGACIÓN:
AUTOMATISMOS Y SISTEMAS INTELIGENTES

PREVIO A LA OBTENCIÓN DEL TÍTULO DE:
INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

AUTOR:
MAILA HERRERA SAHORY NICOLE

ASESOR:
PICO VALENCIA PABLO (PhD)

ESMERALDAS, 2023

TRIBUNAL DE GRADUACION

Trabajo de tesis aprobado luego de haber dado cumplimiento a los requisitos exigidos por el Reglamento de Grado de la PUCESE previa la obtención del título de INGENIERO DE SISTEMAS Y COMPUTACIÓN.

.....

PhD. Pablo Pico Valencia

Tutor de Tesis

.....

Mgt. Jaime Sayago

Lector 1

.....

Mgt. Manuel Nevárez Toledo

Lector 2

.....

Mgt. Homero Velasteguí

Coordinador de la Escuela de Sistemas y Computación

AUTORÍA

Yo, SAHORY NICOLE MAILA HERRERA portadora de la cédula de identidad No. 0803260611 declaro que los resultados obtenidos en la investigación que presento como informe final, previo a la obtención del título de “Ingeniero de Sistemas y Computación” son absolutamente originales, auténticos y personales. En tal virtud, declaro que el contenido, las conclusiones y los efectos legales y académicos que se desprenden del trabajo propuesto de investigación y luego de la redacción de este documento son y serán de mi sola, exclusiva responsabilidad legal y académica.

.....

Sahory Nicole Maila Herrera

C.I. 0803260611

CERTIFICACIÓN

PhD. Pablo Pico Valencia, docente investigador de la PUCE-Esmeraldas, certifica que: La tesis de grado realizada por SAHORY NICOLE MAILA HERRERA, bajo el título "SISTEMA PREDICTIVO BASADO EN TÉCNICAS DE MINERÍA DE TEXTO Y APRENDIZAJE AUTOMÁTICO PARA CLASIFICAR ESTILOS DE REDACCIÓN ACADÉMICA", reúne los requisitos de calidad, originalidad y presentación exigibles a una investigación científica y que han sido incorporadas al documento final las sugerencias realizadas, en consecuencia, está en condiciones de ser sometida a la valoración del Tribunal encargado de juzgarla.

Y para que conste a los efectos oportunos, firma la presente en Esmeraldas, septiembre de 2023.

PhD. Pablo Pico Valencia

Asesor de Tesis

DEDICATORIA

Esta investigación está dedicada con amor a mis padres, quienes han sido mi inspiración para no darme por vencida cada que un obstáculo aparecía en el camino.

AGRADECIMIENTO

En este punto de mi trayectoria académica, deseo expresar mi más sincero agradecimiento a todas las personas que han sido parte fundamental en el proceso de culminación de esta tesis.

En primer lugar, quiero extender mi gratitud a mis padres, Sara Herrera y Luis Maila, cuyo amor incondicional y apoyo constante han sido la base de mi perseverancia y éxito. Su sacrificio y dedicación han sido mi mayor inspiración.

A mis queridos hermanos, Christopher y Shayenka, gracias por su aliento y confianza en mí. Siempre han estado presentes en cada paso de mi camino, compartiendo alegrías y desafíos por igual.

Un agradecimiento especial va dirigido a mis abuelos, quienes han sido ejemplo de fortaleza y sabiduría. Sus consejos y enseñanzas han sido un faro en mi travesía académica.

A mis amigos, gracias por ser pilares fundamentales en mi vida. Su amistad y aliento han sido una fuerza constante.

Mi más sincero reconocimiento va también al Profesor Pablo Pico, mi asesor de tesis, por su orientación experta, paciencia y valiosas sugerencias que han enriquecido este trabajo. Su dedicación y compromiso han sido de inmenso valor.

También, quiero expresar mi agradecimiento a todos los profesores que a lo largo de mi carrera me han brindado su conocimiento y sabiduría. Cada lección y enseñanza han contribuido a mi formación académica.

Finalmente, no puedo pasar por alto agradecer a Dios. Su guía y fortaleza me han permitido superar obstáculos y alcanzar este logro en mi vida. Su amor y gracia han sido mi mayor sustento en cada desafío.

A todas estas personas, les dedico con profundo cariño y agradecimiento esta tesis, que simboliza no solo mi esfuerzo y dedicación, sino también el apoyo incondicional que he recibido de cada uno de ustedes.

Índice general

TRIBUNAL DE GRADUACION I
AUTORIA II
CERTIFICACIÓN III
DEDICATORIA IV
AGRADECIMIENTO V
RESUMEN VIII
ABSTRACT IX
INTRODUCCIÓN 1
Presentación del tema 1
Planteamiento del problema 2
Justificación 3
Objetivos 4
Objetivo General 4
Objetivos Específicos 4
1 CAPÍTULO I: MARCO TEÓRICO 5
1.1 Bases teóricas-conceptuales 5
1.1.1. Atribución de autoría 5
1.1.2. Procesamiento de lenguaje natural 7
1.1.2.1. Minería de texto 8
1.1.3. Aprendizaje Automático 10
1.2 Antecedentes 15
1.3 Fundamentos legales 16
2 CAPÍTULO II: METODOLOGÍA 17
2.1 Delimitación de la investigación 17
2.2 Tipo de investigación 17
2.3 Métodos de investigación 17
2.4 Población y muestra 17
2.5 Variables e indicadores 18
2.6 Técnicas e instrumentos de recolección de datos 19
2.7 Técnicas de procesamiento y análisis de datos 19
2.8 Herramientas 20
2.9 Normas éticas 21
3. CAPÍTULO III: RESULTADOS 22
3 22
3.1 Perfiles de autoría 22

Investigadores seleccionados 22
Estrategia de perfil de autoría aplicada 22
Documentos 23
3.2 Diseño del clasificador 23
Set de datos 23
Descripción del sistema 24
Creación y entrenamiento del modelo 24
3.3 Métricas del clasificador 26
4. CAPÍTULO IV: DISCUSIÓN 33
5. CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES 35
5.1. Conclusiones 35
5.2. Recomendaciones 36
REFERENCIAS 40

RESUMEN

La era de la información ha impulsado la necesidad de desarrollar un sistema que detecte similitudes en el estilo de redacción en trabajos académicos utilizando procesamiento de lenguaje natural y minería de textos, con el propósito de predecir autorías y promover la integridad académica, beneficiando tanto a instituciones educativas como a docentes asesores. En consonancia con esta perspectiva, se busca crear un sistema predictivo que utilice técnicas de minería de texto y aprendizaje automático para clasificar estilos de redacción científica en informes escritos por autores de la PUCESE. A lo largo de esta investigación, se aplicaron métodos inductivos y deductivos, se emplearon diversas técnicas de procesamiento de datos y se exploraron algoritmos de aprendizaje automático para construir un modelo de clasificación de estilos de redacción académica en informes escritos. Durante el desarrollo del sistema la selección de investigadores se realizó mediante un muestreo por conveniencia, eligiendo a aquellos que tenían más de 5 artículos en plataformas específicas. La estrategia de perfil de autoría se basó en la concatenación de documentos en un solo archivo csv para la fase de entrenamiento del modelo. Se utilizó un enfoque cualitativo y experimental, preprocesando los datos mediante tokenización, eliminación de stopwords y lematización, y se emplearon técnicas como TF-IDF y Word Embeddings para convertir el texto en un formato comprensible para el modelo. Se exploraron diversos algoritmos de aprendizaje automático, como Máquinas de Vectores de Soporte, Bosques Aleatorios y Regresión Logística, para la construcción del modelo, y se evaluaron mediante métricas como precisión, recall y exactitud. A través de la matriz de confusión, se examinaron los resultados y las capacidades de los modelos en la tarea de clasificación de estilos de redacción académica. El grado de precisión del modelo con el algoritmo Máquina de Vectores de Soporte Lineal logró un alto nivel de precisión del 80.8%, destacando su capacidad para capturar con precisión instancias tanto positivas como negativas. Por otro lado, el modelo Máquina de Vectores de Soporte RBF mostró una dificultad en la clasificación precisa de ciertos estilos de escritura. En base a estos resultados, se recomienda considerar cuidadosamente el equilibrio entre precisión y recuperación al seleccionar el algoritmo adecuado para esta tarea.

Palabras clave: aprendizaje automático, minería de texto, procesamiento de lenguaje natural, estilos de redacción.

ABSTRACT

The information era has driven the need to develop a system that detects similarities in writing style in academic papers using natural language processing and text mining, with the purpose of predicting authorship and promoting academic integrity, benefiting both educational institutions and faculty advisors. In line with this perspective, we seek to create a predictive system that uses text mining and machine learning techniques to classify scientific writing styles in reports written by PUCESE authors. Throughout this research, inductive and deductive methods were applied, various data processing techniques were employed, and machine learning algorithms were explored to build a model for classifying academic writing styles in written reports. During the development of the system, the selection of researchers was done by convenience sampling, choosing those who had more than 5 articles in specific platforms. The authorship profiling strategy was based on concatenation of documents into a single csv file for the training phase of the model. A qualitative and experimental approach was used, preprocessing the data by tokenization, stopword removal and lemmatization, and techniques such as TF-IDF and Word Embeddings were employed to convert the text into a format understandable to the model. Various machine learning algorithms, such as Support Vector Machines, Random Forests and Logistic Regression, were explored for model building and evaluated using metrics such as precision, recall and accuracy. Through the confusion matrix, the results and capabilities of the models in the task of classifying academic writing styles were examined. The degree of accuracy the model with the Linear Support Vector Machine algorithm achieved a high level of accuracy of 80.8%, highlighting its ability to accurately capture both positive and negative instances. On the other hand, the RBF Support Vector Machine model showed difficulty in accurately classifying certain writing styles. Based on these results, it is recommended to carefully consider the trade-off between accuracy and recall when selecting the appropriate algorithm for this task.

Keywords: machine learning, text mining, natural language processing, writing styles.

INTRODUCCIÓN

Presentación del tema

Actualmente el mundo genera 2.5 quintillones de bytes de datos diariamente, de los cuales solo el 80% son datos que están expresados en lenguaje natural, mismo que es fácil de entender por un ser humano; pero no por los softwares tradicionales [1]. Al existir una gran cantidad de información en la web hace que el proceso de búsqueda de datos para una investigación sea complejo [2]. Por lo tanto, se requiere de técnicas y herramientas que permitan extraer, analizar y clasificar grandes cantidades de datos. En este sentido, la Inteligencia Artificial, y específicamente la minería de texto, en conjunto con las técnicas de aprendizaje automático han aportado significativamente.

La minería de texto permite que la búsqueda de información pueda realizarse con mayor eficacia y rapidez ya que mediante este proceso se extraen patrones característicos en lenguaje natural de los documentos almacenados en formato texto [2]. A diferencia de la minería de datos, la cual está caracterizada por extraer información implícita, la minería de texto obtiene información clara y precisa que se encuentra explícitamente en el texto [3]. Esto significa que se busca comprender el significado del texto tal como lo hacen los humanos.

Los campos en los que se puede aplicar la minería de texto en la actualidad son muy numerosos, esto se debe a que para realizar una investigación en cualquier ámbito se necesita trabajar con información la cual deberá ser extraída y analizada. Ejemplo de ello, se ve en el área de educación donde se está combinando el aprendizaje automático con la minería de textos para automatizar y clasificar material bibliográfico [4]. En esta misma línea, en el campo de la salud se ha usado la minería de texto, aprendizaje automático y el procesamiento del lenguaje en un entorno biosanitario para acceder fácilmente a los datos clínicos del paciente [5]. En definitiva, la minería de texto es relevante cuando se requiere comprender el significado del texto de documentos como es el caso de libros, artículos, trabajos de tesis, entre otros.

Planteamiento del problema

La pandemia ha requerido que la humanidad haya tenido que adaptarse a un modelo de trabajo y estudios a través de medios telemáticos. La educación tuvo que pasar de un modelo tradicional hacia la virtualidad. La educación se vio reemplazada por un entorno virtual de aprendizaje el cual proporciona herramientas para presentar contenido y facilitar la comunicación e interacción con los estudiantes [6]. Para que los alumnos puedan acceder a los conocimientos, los docentes proporcionan a través de medios telemáticos, una variedad de recursos didácticos en formato de diapositivas, archivos PDF, tutoriales, videos, entre otros. Todo este material sirve de apoyo para que los estudiantes puedan realizar las tareas solicitadas.

Debido a la facilidad que hoy en día tienen los alumnos para acceder a la información y poder realizar sus consultas, informes y ensayos, se teme que solo estén haciendo uso de esta al "copiar y pegar" la información, en lugar de leerla, analizarla, sintetizarla y plasmarla en sus investigaciones formativas y de titulación.

En el contexto académico, otra práctica común entre los jóvenes al momento de cumplir con sus trabajos académicos es pagar o pedir a "asesores externos" que realicen la actividad propuesta por el profesor. Lo antes expuesto deja ver claramente una problemática que está en contra de la ética académica por parte de los estudiantes al momento de elaborar sus trabajos escritos, sean estos, proyectos, informes, consultas, o trabajos de titulación. Una vez identificado el problema, surgen las siguientes interrogantes: (i) ¿en qué medida el desarrollo de una nueva herramienta ayudará a detectar similitudes de estilo de redacción en trabajos académicos? y (ii) ¿cuán efectivo puede llegar a ser un sistema de predicción de estilos de redacción académica basado en las técnicas de minería de texto y aprendizaje automático? Estas son algunas de las preguntas de investigación que la investigación ha dado respuesta.

Justificación

La redacción académica es una actividad que los estudiantes deben realizar en cada materia que cursan dentro de un programa de formación de grado o posgrado. Cuando los docentes reciben el trabajo finalizado, en algunas ocasiones se teme que no haya sido escrito por el estudiante. Por ello, esta investigación plantea crear un sistema que permita entrenar un modelo de clasificación automático el cual pueda predecir si una persona ha sido el autor de un trabajo académico. Para ello, se empleará el procesamiento del lenguaje natural y la minería de textos, de tal forma que, a partir de documentos previamente elaborados por los estudiantes en su carrera, se identifiquen los estilos de redacción de los estudiantes y posterior a ello, sirvan para realizar la predicción automática.

Es importante señalar que el sistema está orientado a evaluar en el futuro, el estilo de redacción en los trabajos de fin de carrera de grado, actualmente denominados según el Reglamento del Régimen Académico formulado por el CES como Unidad de Integración Curricular. No obstante, el piloto se realizará con los trabajos de los docentes, quienes a lo largo de su trayectoria han realizado varias publicaciones. No se tomó el caso de documentos de estudiantes debido a que no existe una base de datos con varios documentos escritos por los estudiantes, generalmente, solo se cuenta con el trabajo de grado; pero este es un único documento y no es suficiente para entrenar un clasificador. No obstante, las lecciones aprendidas en este estudio permitirán que se encamine el estudio al caso particular de los estudiantes. De esta forma, la universidad no solo podrá validar la deshonestidad académica desde el punto de vista de la similitud; sino que se hará también desde la perspectiva del estilo de escritura académica que se emplea en el diseño del plan; pero sobre todo en el informe del Trabajo de Integración Curricular. Ello implica que los beneficiarios será la Dirección Académica y de Investigación de la PUCESE, unidades encargadas de dichos procesos. Asimismo, los docentes asesores se verán beneficiados cuando tienen dudas de si un trabajo de integración curricular ha sido o no escrito por el “autor” (estudiante).

Objetivos

Objetivo General

Desarrollar un sistema predictivo mediante el uso de las técnicas de minería de texto y de aprendizaje automático para llevar a cabo la clasificación de estilos de redacción científica en informes escritos por autores de la PUCESE.

Objetivos Específicos

a) Identificar las características (o métodos) para detectar el estilo de redacción académica en informes escritos.
b) Analizar los algoritmos de aprendizaje automático que han sido ampliamente usados en procesos de minería de textos para clasificación de patrones.
c) Crear un clasificador automático capaz de predecir la autoría de un documento académico.
d) Evaluar la eficacia del modelo desarrollado en el contexto de los artículos científicos publicados por los docentes de la PUCESE.

1 CAPÍTULO I: MARCO TEÓRICO

1.1 Bases teóricas-conceptuales

1.1.1. Atribución de autoría

En los últimos años, el análisis de autoría (AA) ha ganado popularidad, ya que ha realizado grandes contribuciones a los problemas de seguridad nacional e inteligencia, la lingüística forense, el análisis de mercados y la identificación de rasgos de personalidad. Se centra en la clasificación automática de textos basándose en el estilo de redacción de los autores. Los aspectos que incluye este análisis son la atribución de autoría, la verificación del autor, la detección de plagios, la determinación del perfil del autor y la detección de inconsistencias estilísticas [7].

Abarcando la parte del análisis de autoría (AA) apoyado en otros artículos y todo lo relacionado a ello, en [8] se explica a detalle como este concepto pertenece a una subtarea de la minería de datos, teniendo como propósito minar un estilo o patrón de redacción correspondiente y característico de un autor, partiendo de los documentos redactados por el mismo. En este artículo se señala que se pueden modelar dos tipos de métodos para poder determinar la autoría en base a la representación de los documentos de los autores, esto es, basados en el perfil y en instancias.

El enfoque basado en el perfil del autor consiste en concatenar todos los documentos del autor en un solo documento el cual se convertirá en los datos de la fase de entrenamiento para crear su perfil. En la **Figura 1**, se muestra la estructura típica de este enfoque, además se observa que para realizar la predicción de un nuevo documento, se realiza un cálculo en la similitud entre el perfil ya generado y las características que presenta el nuevo documento [9].

El segundo enfoque basado en instancias que se muestra en la **Figura 2** es representado de forma vectorial en donde cada documento se ingresa de forma individual de cual se extraen un conjunto de características, las cuales son utilizadas para entrenar el modelo para luego al ingresar el texto desconocido, el modelo pueda dar como resultado el autor más probable [9].

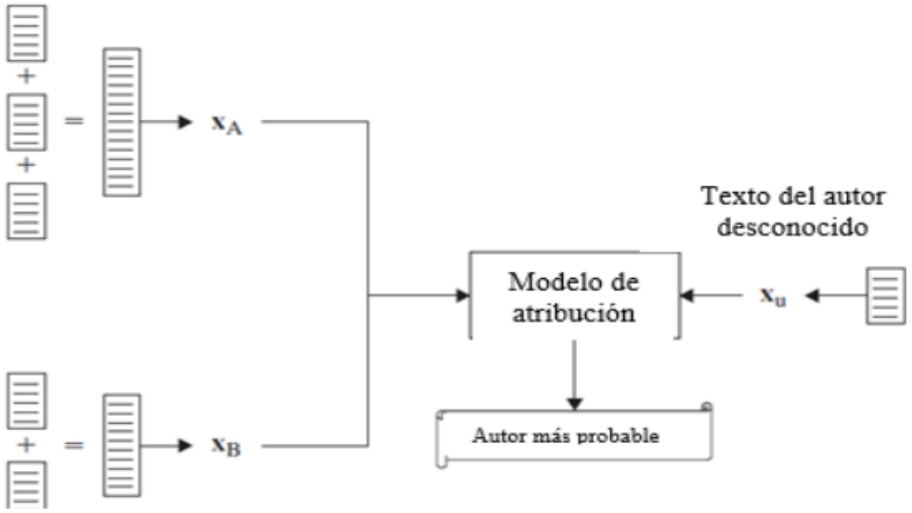


Figura 1. Enfoque basado en perfil del autor [9].

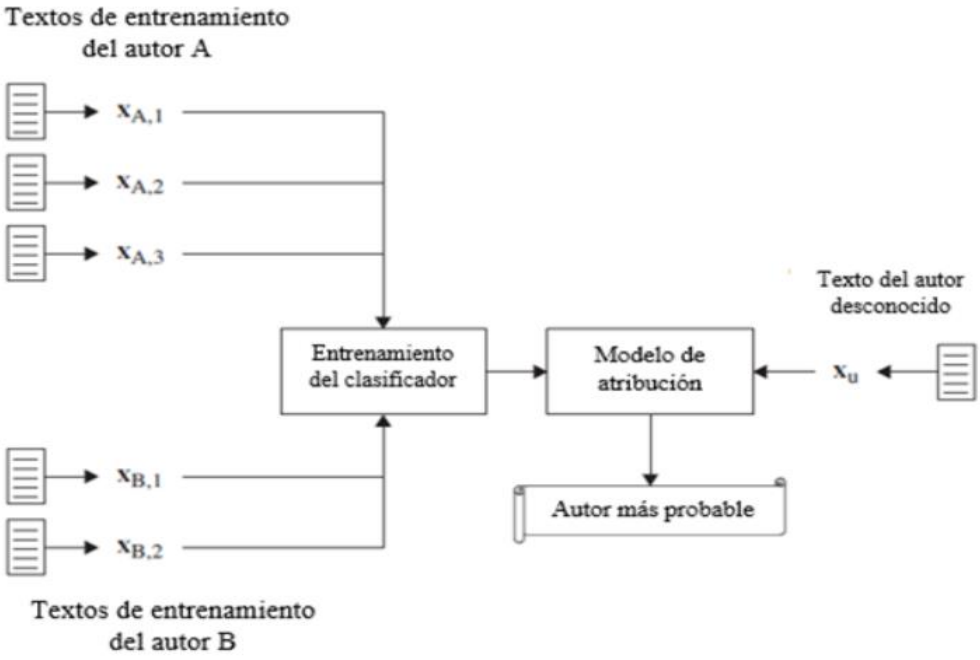


Figura 2. Enfoque basado en instancias [9].

Por otra parte, en otro estudio [7], los autores abarcan otros subcampos pertenecientes al análisis de autoría, los cuales corresponden a los siguientes: **determinación del perfil del autor** (DPA) que permite identificar patrones compartidos por un grupo de gente abordando problemas de clasificación basados en la edad, género, orientación política, etc. También, la **atribución de autoría** (ATA) que consiste en la atribución de un texto cuya de autoría es desconocida a uno de un conjunto de autores potenciales. Ambas

técnicas se combinan para proponer un clasificador de documentos periodísticos basados en sus inclinaciones políticas.

1.1.2. Procesamiento de lenguaje natural

El Procesamiento del Lenguaje Natural (PLN) es una rama de la Inteligencia Artificial (IA) cada vez más utilizada en tareas de traducciones automáticas de textos, revisión de ortografía o conteo de palabras. Su objetivo es aprender, comprender, reconocer y producir contenido de lenguaje humano [10].

En la **Figura 3**,se contempla cuatro tipos de análisis sobre la entrada de un texto en lenguaje natural, esto es, análisis morfológico, sintáctico, semántico y pragmático.



Figura 3. Proceso PLN

El análisis morfológico implica examinar las oraciones con el propósito de reconocer unidades léxicas o tokens, así como categorías de palabras. Dado que una palabra puede asumir diversos roles según el contexto, surgen desafíos de ambigüedad. En consecuencia, el análisis morfológico se encarga de establecer la forma, clase o categoría gramatical de cada palabra en una oración, asignándole su función específica en el contexto correspondiente. Este proceso es esencial para una comprensión profunda del significado textual y su estructura lingüística [11].

Por otro lado, surge el análisis sintáctico, cuyo objetivo radica en identificar las relaciones de dependencia estructural entre las palabras presentes en la oración bajo estudio. La sintaxis juega un papel central en la transmisión del significado en la mayoría de las lenguas. Tomando como ejemplo la referencia [11], se ilustra cómo las oraciones "A ella le gusta Miguel de Cervantes" y "A Miguel de Cervantes le gusta ella" difieren solo en su estructura sintáctica, pero comunican significados distintos. Debido a la ambigüedad inherente al lenguaje, este tipo de análisis representa un desafío complejo para la automatización. A pesar de su complejidad, el análisis sintáctico es fundamental para la comprensión precisa del sentido y la organización gramatical en el discurso.

El análisis que incluye la desambiguación de palabras con múltiples sentidos es el semántico. Utiliza el resto de la oración para esta desambiguación y para ello se puede implementar una amplia gama de métodos para lograrlo [11].

Finalmente, uno de los análisis más complejos es el análisis pragmático debido a que las oraciones, y las palabras que componen las oraciones, no se las puede interpretar de forma literal porque su significado dependerá en gran medida del contexto en el que se utilicen [11].

1.1.2.1. Minería de texto

La minería de textos tiene como objetivo encontrar información relevante para un propósito a partir de textos en lenguaje natural. En [12], presentan dos definiciones de autores diferentes, uno de ellos define la minería de texto como aquella que siempre está buscando patrones en un texto. Para lo cual, analiza el texto para extraer información útil para un propósito específico. Mientras que otro de los autores, lo describen como un proceso intensivo en conocimiento, en el cual el usuario interactúa con una colección de documentos a lo largo del tiempo determinado, haciendo uso de un conjunto de herramientas de análisis. Su objetivo es extraer información útil de la fuente de datos a partir de la identificación y el descubrimiento de patrones interesantes.

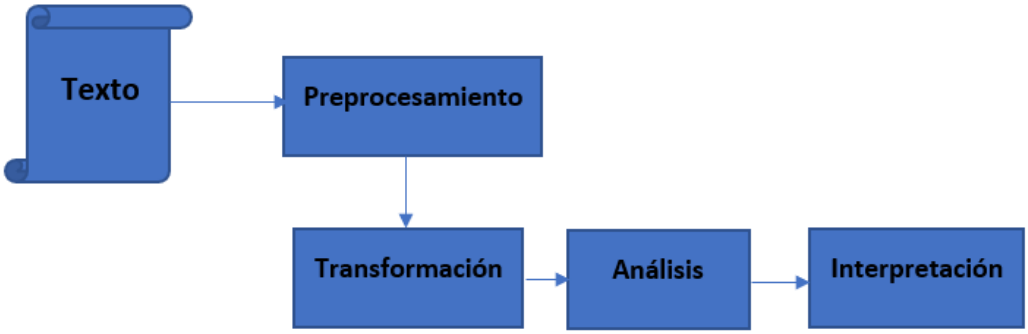


Figura 4. Esquema del flujo de la minería de texto [13]

En la **Figura 4,** muestra el flujo de la minería de texto la cual empieza por los datos de entrada, en este caso es el texto, se realiza un preprocesamiento lo cual equivale a la limpieza de este. Después se procede a realizar la transformación del texto en datos estructurados para poder procesarlos mediante las técnicas tradicionales. Luego se aplica

la técnica elegida que mejor se adapte para el caso de estudio y finalmente se interpretan los resultados obtenidos.

Las técnicas más relevantes usadas en la minería de texto son 3: Método booleano, Análisis de semántica latente (Latent Semantic Analysis; LSA), Modelo de espacio vectorial semántico [1].

- **Método booleano.** Es el método más útil y usado en el cual cada documento o texto es representado por palabras clave, y la importancia de estas palabras se miden según la presencia o ausencia de estas en el texto.
- **Análisis de semántica latente (LSA).** Es un modelo de espacio vectorial, el cual analiza que ciertas palabras, las cuales aparentemente independientes, en realidad tienen relación entre sí por temas subyacentes no observados.
- **Modelo de espacio vectorial semántico.** Esta técnica es una mejora del método booleano, siendo uno de los modelos más usados en tareas de procesamiento del lenguaje natural y recuperación de información. En este modelo vectorial, el enfoque semántico de las palabras está dividido en dos partes: la semántica distribucional que analiza los significados de cada palabra aislada y la semántica composicional que también se encarga de analizar el significado, pero de las frases o párrafos.

Otras técnicas avanzadas en minería de texto son, TF-IDF (Term Frequency-Inverse Document Frequency) y los Word Embeddings que se distinguen por su habilidad para asignar pesos contextuales a las palabras y representarlas en espacios semánticos densos y continuos.

TF-IDF (Term Frequency-Inverse Document Frequency) combina dos conceptos:

- Frecuencia del término (TF): Mide cuántas veces aparece una palabra en un documento específico. Palabras que aparecen con frecuencia pueden ser importantes para ese documento.
- Inverso de la Frecuencia en el Documento (IDF): Calcula la rareza de una palabra en toda la colección de documentos. Palabras que son raras en la colección, pero aparecen en un documento pueden ser más informativas para ese documento.

Es una técnica fundamental en el procesamiento de texto que evalúa la importancia relativa de una palabra en un documento dentro de una colección de documentos [14].

Para cada palabra en un documento, se calcula tanto su Frecuencia del Término (TF) como su Frecuencia Inversa del Documento (IDF). El puntaje TF-IDF para una palabra en el documento se obtiene multiplicando su TF por su IDF. Las palabras con puntajes TF-IDF más altos son consideradas como más relevantes para el documento y pueden ser posibles palabras clave. Esta técnica tiene ventajas como resaltar palabras clave y características importantes en un texto, lo que resulta útil para tareas como recuperación de información y clasificación de documentos. Además, no requiere grandes cantidades de datos para un aprendizaje efectivo. Sin embargo, también tiene limitaciones, ya que no captura relaciones semánticas complejas entre palabras y no considera el orden de las palabras en el documento [15].

Word Embeddings es una técnica de procesamiento de lenguaje natural (NLP) que representa las palabras en un espacio vectorial de números reales. En lugar de tratar las palabras como entidades separadas, esta técnica las convierte en vectores matemáticos que capturan las relaciones semánticas y sintácticas en un contexto específico. Estos vectores resultantes son significativos para las máquinas y permiten realizar tareas NLP complejas como análisis de sentimientos, traducción automática y recuperación de información de manera más eficiente [16].

1.1.3. Aprendizaje Automático

El aprendizaje automático es una rama de la inteligencia artificial dedicada al estudio de agentes o programas de software que aprenden o evolucionan basados en su experiencia adquirida en el entrenamiento del modelo. El objetivo principal es desarrollar técnicas que permitan a las computadoras aprender [17].

En la **Figura 5**, se muestra un proceso general que se sigue durante el aprendizaje, ya que con los datos de entrada se realiza la extracción de características obteniendo así, un conjunto de datos de entrenamiento los cuales son representados como un vector de características. Luego se procede a realizar el entrenamiento del algoritmo de clasificación para crear el modelo predictivo que cuenta también con una validación al momento de ingresar los datos nuevos o de prueba dando como resultado los datos esperados.

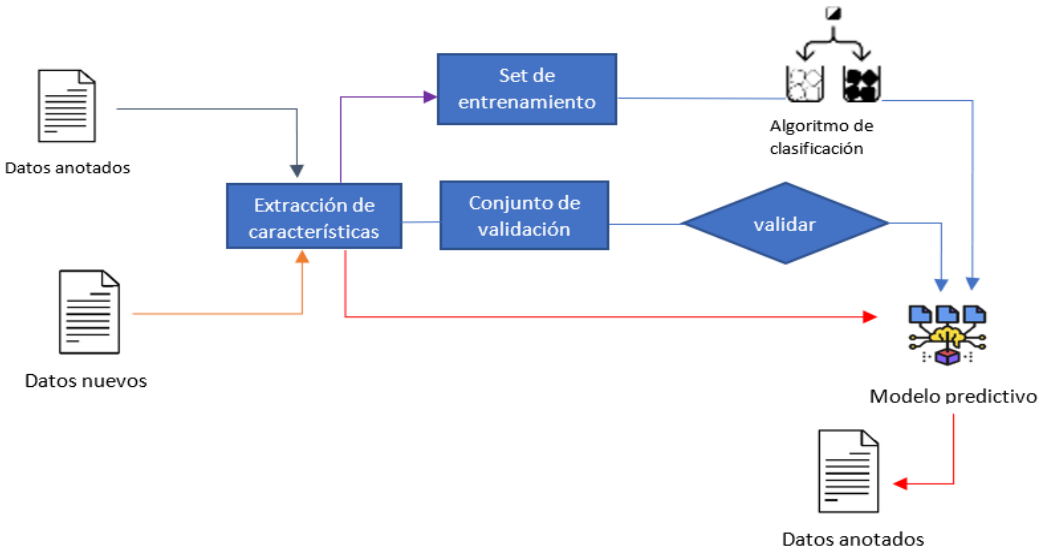


Figura 5. Proceso de aprendizaje automático [18]

Existen 3 tipos de aprendizaje automático; el aprendizaje con clase o también llamado aprendizaje supervisado tiene como objetivo aprender un patrón mediante un conjunto de datos de entrenamiento, el cual nos permite realizar predicciones al ingresar datos nuevos o que no han sido observados previamente. El aprendizaje sin clase o aprendizaje no supervisado trabaja con datos sin etiquetar o datos sin estructura de los cuales no se posee información previa que facilite la obtención de patrones, por ello este tipo de aprendizaje trata de identificar distintos subgrupos mediante una exploración de los datos indicados, que se basan en la similitud de las características. El aprendizaje reforzado tiene como objetivo conseguir un sistema que le permita mejorar su aprendizaje mediante el constante cambio de su entorno. De esa forma si la clasificación es correcta, el algoritmo va adquiriendo experiencia o confianza a través de una función de recompensa. Posteriormente se realiza una exploración de un conjunto de estados para producir la salida adecuada mediante prueba y error habiendo obtenido la mejor recompensa posible [19].

Una definición más general del aprendizaje de los algoritmos supervisados dice que, se realiza mediante unos datos de entrenamiento para deducir una función que haga lo mejor posible el mapeo entre unas entradas y una salida [18]. A continuación, se detallarán algunos de los algoritmos de aprendizaje supervisado más utilizados en la actualidad [20].

Árboles de decisión. Es un modelo de predicción que se utiliza en el ámbito de la inteligencia artificial. Dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema. La **Figura 6** es una representación gráfica de un árbol de decisión en el cual se puede identificar tres tipos de nodos; el nodo raíz que equivale a la variable con mayor importancia y que produce la primera división, el nodo hijo es un nodo intermedio que a su vez genera más divisiones en el árbol, mientras que el nodo hoja también llamado nodo terminal que el que da por terminado el esquema, dando la clasificación definitiva.

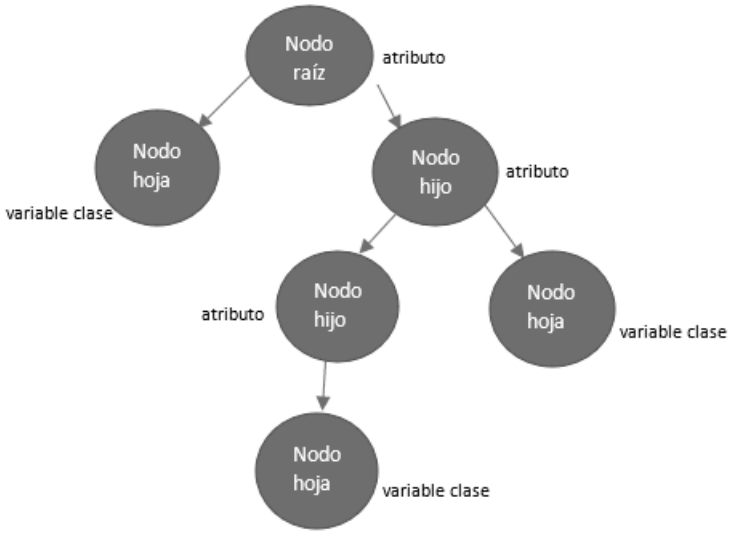


Figura 6. Representación de un árbol de decisión [21]

Bosques aleatorios. También llamado Random Forest, es un algoritmo que tiene una combinación de árboles predictores en donde cada árbol de decisión da una clasificación y finalmente la decisión con mayor "votos" es la predicción del algoritmo. La **Figura 7** muestra lo antes mencionado, este algoritmo parte de una raíz o instancia para luego dividirse en tres arboles de decisión, al finalizar cada uno propone una clasificación y al terminar, la decisión del árbol con mayor votos o datos acertados se convierte en la decisión final.

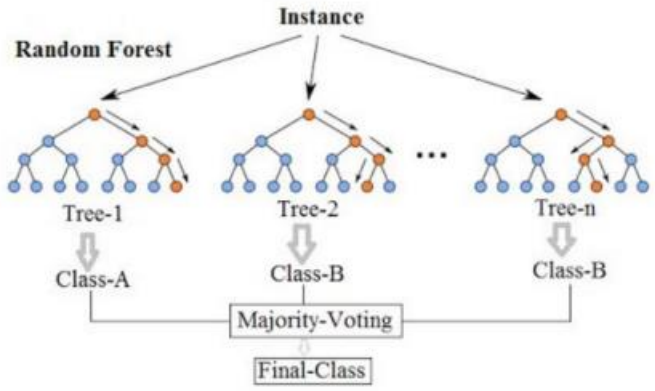


Figura 7. Representación del algoritmo Random Forest [21]

K-vecinos más cercanos. Es un algoritmo de aprendizaje supervisado que funciona como un clasificador y ayuda a encontrar de forma fácil la clasificación de los datos. El objetivo de este algoritmo es "encontrar un número predefinido de muestras de entrenamiento más cercanas en distancia al nuevo punto, y predecir la etiqueta de estos" [22].

En la **Figura 8**, muestra la forma en que trabaja este algoritmo, empezando por el aprendizaje en el cual se establece los datos en dos tipos, para pasar a la parte de la clasificación el algoritmo elige un punto de referencia para cada tipo y luego analiza los datos que tiene más cerca para así poder clasificarlos correctamente basado en su cercanía al punto referencial.

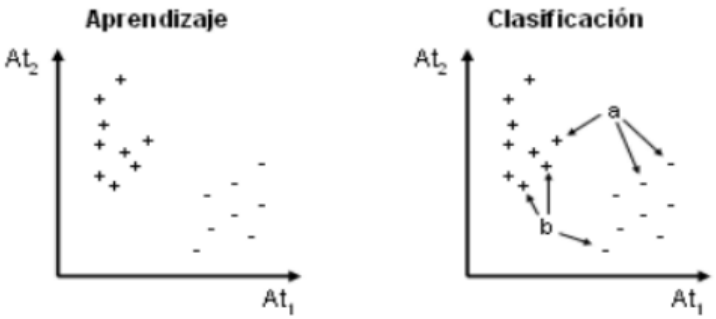


Figura 8. Ejemplo del algoritmo KNN [21]

Máquinas de vectores de soporte. Es un algoritmo de aprendizaje supervisado que funciona como un clasificador. El objetivo del algoritmo de la máquina de vectores de soporte es encontrar un hiperplano en un espacio N-dimensional para así clasificar los

puntos de datos. Entre las ventajas de utilizar este algoritmo es que son más efectivos en espacios de alta dimensión, incluso en los casos en donde el número de dimensiones es mayor que la muestra no disminuye su efectividad, por ultimo no consumen tanta memoria [22].

En la **Figura 9**, muestra una representación gráfica de este algoritmo en el cual se establece un hiperplano optimo, que se define como "la función de decisión lineal con margen máximo entre los vectores de las dos clases" [23]. Solo se necesita una cantidad mínima de los datos de entrenamiento para poder definir el hiperplano optimo el cual nos permite determinar un margen claro entre las clases.

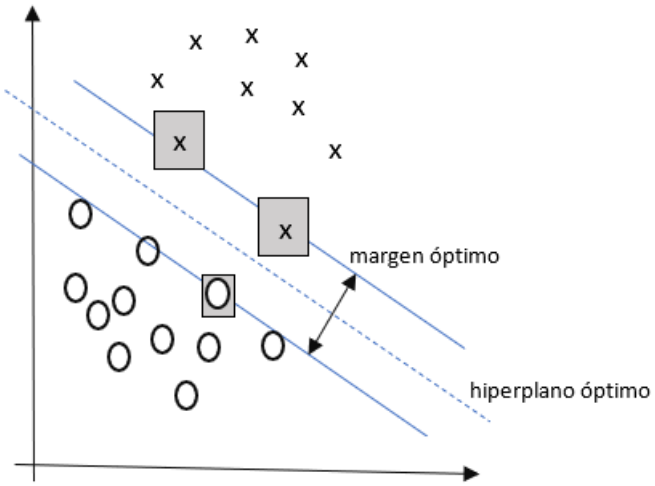


Figura 9. Representación del algoritmo SVM [21].

1.2 Antecedentes

En esta sección de la investigación se describen algunos de los artículos con mayor relevancia en el tema, teniendo como finalidad explorar que algoritmos utilizaron otros autores obteniendo buenos resultados en el entrenamiento y prueba del modelo, así como también el porcentaje de precisión.

En el año 2016, Marcial Contreras [4] desarrolló un clasificador automatizado que permite clasificar el material bibliográfico mediante la minería de texto con la finalidad de mejorar y agilizar el proceso de organización de información. El uso de la minería de texto en esta investigación permitió desarrollar el clasificador mediante el método de clasificación de documentos que es de tipo supervisado. Para el cálculo de la precisión del clasificador se realizó comparaciones entre los temas bibliográficas que fueron asignados de forma manual y automatizada, obteniendo así un grado de precisión del 75.70%. Este articulo tiene similitud con la investigación en que para obtener los datos deben analizar los documentos y así poder preparar la información mediante la minería de texto y el lenguaje natural.

De igual forma, Ángel Godoy [12] realizó un análisis de los algoritmos en su investigación que se pueden utilizar para el aprendizaje de maquina (machine learning), haciendo una comparativa en cuanto al porcentaje de utilización de cada técnica como: principal, secundaria o mencionada en artículos. El número de algoritmos comparados son 13, y como resultado de la investigación obtuvieron que la técnica *support vector machine,* tiene un porcentaje de uso como técnica principal del 22.72%, como secundaria para comparativa 20.59% y mencionada en artículos obtuvo un 16.88%.

Es necesario recalcar que en su investigación Jovany Ramirez, Maya Carrillo y Josefa Somodevilla [9], presentan dos enfoques como método para realizar la atribución de autoría, el primero es un enfoque basado en el perfil del autor, en el cual se concatenan los documentos de ese autor en uno solo para crear el perfil, mientras en el otro método se basa en instancias, es decir, se analizan todos los documentos por separado. Tuvieron como resultado que el enfoque basado en el perfil obtiene mejores resultados, siendo más eficiente y preciso.

Por el año 2020, Erika Páez y Andrés Monroy [24] realizaron una implementación de un módulo que permite analizar los sentimientos de los datos de Twitter mediante la minería de texto. Para medir el rendimiento de los algoritmos se utilizaron las medidas de

desempeño precisión, recall, F1 − Score, con los algoritmos de Random Forest y Naive Bayes. Teniendo como resultado que el mejor desempeño se dio con el algoritmo Random Forest. Las métricas utilizadas en este articulo coincide con las métricas seleccionadas en la investigación, lo cual nos permite tener mayor seguridad para continuar con el proceso de entrenamiento y pruebas del sistema una vez terminado su desarrollo.

1.3 Fundamentos legales

Dentro de las bases legales que se deben tomar en cuenta para el desarrollo del presente proyecto que protegen los derechos intelectuales y tecnológicos, se estipulan las siguientes normativas:

Esta investigación se rige al Art. 4, del capítulo uno del derecho al autor en la Ley de Propiedad Intelectual del Ecuador, donde se especifica que; “Se reconocen y garantizan los derechos de los autores y los derechos de los demás titulares sobre sus obras.” [25], por ello, todas las ideas y conceptos utilizados en este proyecto han sido citados correctamente, respetando y protegiendo las ideas de los autores.

Otra normativa que defiende los derechos de los autores a tener créditos por sus conocimientos se presenta en el capítulo sexto de la Constitución de la República del Ecuador: trabajo y producción, sección segunda: tipos de propiedad, articulo 322, especifica que “se reconoce la propiedad intelectual de acuerdo con las condiciones que señale la ley. Se prohíbe toda forma de apropiación de conocimientos colectivos, en el ámbito de las ciencias, tecnologías y saberes ancestrales” [26].

También es importante recalcar que en el Art. 142 correspondiente a las tecnologías libres del Código Orgánico de la Economía Social de los Conocimientos (COESC) del Ecuador dice que: “Se entiende por tecnologías libres al software de código abierto, los estándares abiertos, los contenidos libres y el hardware libre”[27]. Por consiguiente, el desarrollo del sistema predictivo de esta investigación se realizó con herramientas de código abierto para promover su uso y evitar infringir las leyes al usar programas comerciales sin licencia.

2 CAPÍTULO II: METODOLOGÍA

2.1 Delimitación de la investigación

Esta investigación se llevó a cabo en la Pontificia Universidad Católica del Ecuador Sede Esmeraldas siendo esta su delimitación espacial. Por otro lado, la parte práctica de esta investigación fue desarrollada en el primer semestre del año 2022, en la cual se preparó las herramientas, se desarrolló el modelo de clasificación, entrenándolo para poder obtener buenos resultados en cuanto a la precisión de este. Los artículos que se consideraron para la realización del trabajo de investigación fueron enmarcados dentro del periodo 2014-2022.

2.2 Tipo de investigación

Esta investigación es de tipo cualitativa, ya que primero se analizó la información y conocimientos sobre los algoritmos y técnicas aplicadas en otras investigaciones. En la segunda parte correspondiente a la etapa de desarrollo, pasa a ser cualitativa y experimental puesto que al aplicar los algoritmos de aprendizaje supervisado se evaluó el tiempo de entrenamiento del modelo, el rendimiento y precisión de este.

2.3 Métodos de investigación

En la etapa inicial de esta investigación se utilizó el método inductivo debido a que se realizó una búsqueda de artículos que proporciono información relacionada a la minería de texto, aprendizaje automático, procesamiento del lenguaje natural. Con ellos, se amplió los conocimientos que sustentan este proyecto, un clasificador de estilos de redacción; mientras que en la etapa de desarrollo se utilizó el método deductivo, ya que nos permitió desarrollar las reglas, e identificar patrones que nos permitirán construir el modelo de clasificación de estilos de redacción.

Además, se empleó el método experimental ya que se evaluaron los algoritmos más utilizados para identificar el de mayor eficacia y precisión en el campo de la atribución de autoría.

2.4 Población y muestra

En esta investigación se consideró como población los artículos científicos escritos por los docentes investigadores de la PUCESE. Un docente investigador se dedica a producir, desarrollar y socializar el conocimiento adquirido desde su punto de vista junto con los demás involucrados en el proceso de aprendizaje, para así poder tener una retroalimentación que constantemente se adapta a los cambios [28].

La lista de docentes investigadores de la Pucese en el Anexo 1, comprende a un total de 58 docentes, de los cuales se realizó una recopilación de los perfiles de Google Académico y ResearchGate. En la **Figura 10**, muestra que la mayoría no contaban con un perfil en Google Académico, por ello, se descartó esta plataforma para obtener los artículos. Mientras que en ResearchGate el número de docentes que tiene 5 o más artículos publicados fue 14, de los cuales se seleccionó 10 autores para realizar la base de datos.

	Scholar	ResearchGate
Nª docentes sin perfil	50	24
Nª docentes >= 5 articulos	6	14
Nº de docentes con 0 articulos	10	

Figura 10. Análisis de los datos recopilados para la muestra.

Se utilizó un muestreo por conveniencia, debido a que se definió que debían ser mínimo 5 artículos científicos por cada docente para que el modelo pueda crear el perfil del autor y realizar el entrenamiento con éxito, obteniendo así un porcentaje de mayor precisión.

2.5 Variables e indicadores

Esta investigación se llevó a cabo a partir de la variable "modelo de clasificación" con sus respectivos indicadores. Esta información en conjunto con el tipo de variable se ilustra en la **Tabla 1**.

Tabla 1. Variables e indicadores del estudio

VARIABLE	DIMENSIÓN	INDICADORES	TIPO DE VARIABLE	ENTIDADES
Modelo de clasificación	Rendimiento	Tiempo de entrenamiento Tiempo de prueba Tiempo de respuesta a usuario	Cuantitativa	Modelo predictivo entrenado
	Precisión	Métrica Accuracy Métrica Recall Métrica Precision	Cuantitativa	Modelo predictivo entrenado

Dimensión Rendimiento. Se refiere a las métricas de los porcentajes obtenidos de cada algoritmo utilizados para la clasificación de estilos de redacción académica. A continuación, se describen los indicadores que se contemplan en la investigación:

- Tiempo de entrenamiento. El porcentaje de entrenamiento que se obtiene del algoritmo.
- Tiempo de prueba. El porcentaje de prueba que se obtiene del algoritmo.
- Tiempo de respuesta al usuario. El tiempo que toma el algoritmo en dar la respuesta al usuario.

Dimensión Precision. Se refiere a las métricas de los porcentajes obtenidos de cada algoritmo utilizados para la clasificación de estilos de redacción. A continuación, se describen los indicadores que se contemplan en el estudio:

- Accuracy. Porcentaje de precisión obtenido en el algoritmo.
- Precision. La relación entre las predicciones positivas correctas y las predicciones positivas en general.
- Recall. Proporción de aciertos en las predicciones positivas correctas en relación al total de casos positivos en el conjunto de datos.

2.6 Técnicas e instrumentos de recolección de datos

Como técnica de recolección de datos se utilizó la lista de docentes detallada en el Anexo 1, para realizar la búsqueda de los perfiles en Scholar y ResearchGate de cada docente, una vez filtrada la lista descartando la plataforma Scholar y teniendo como característica que por lo menos deben ser 5 artículos, se procedió a descargar y guardar los archivos en las carpetas con el nombre de cada docente.

2.7 Técnicas de procesamiento y análisis de datos

En esta investigación se utilizó la estadística inferencial para realizar las deducciones sobre los datos de los docentes, de esta forma se llegó a la conclusión de los autores seleccionados basados en los artículos obtenidos y como herramienta para visualizar esta información se utilizó el programa Excel.

Por otro lado, para el análisis de los datos obtenidos en la fase de entrenamiento y prueba del modelo se utilizó la matriz de confusión, instrumento usado en aprendizaje automático. Se tomo en cuanta las métricas accuracy, recall y precisión. La exactitud o Accuracy es la cantidad de predicciones positivas que fueron correctas, la fórmula para calcularla es: $\frac{(VP+VN)}{(VP+FP+FN+VN)}$ en donde, VP corresponde a verdaderos positivos, mientras que VN significa verdaderos negativos, así mismo, FP es el número de falsos positivos, y FN son los falsos negativos. La sensibilidad o Recall, es la proporción de casos positivos que fueron correctamente identificados por el algoritmo, se calcula mediante: $\frac{VP}{(VP+FN)}$. Y, por último, la métrica de precisión que equivale al porcentaje de los casos positivos detectados se calcula como: $\frac{VP}{(VP+FP)}$.

2.8 Herramientas

A continuación, se describen brevemente las herramientas utilizadas para el desarrollo del modelo predictivo;

Python. Es un lenguaje de programación de alto nivel que tiene la función de poder usarse en cualquier plataforma, esto hace que el código sea simple, fácil de entender e intuitivo [29]. La sintaxis y comandos que se utilizan en son distintos que los de cualquier otro lenguaje de programación interpretado [30].

Este lenguaje de programación multiplataforma, en el cual se puede desarrollar inteligencia artificial usando diferentes tipos de algoritmos, posee librerías que facilitan el análisis de los datos las cuales permiten llevar a cabo procesos de recolección y limpieza de datos, exploración, modelado y visualización de datos de manera eficiente.

NLTK. NLTK2 (Natural Language Toolkit) se presenta como una plataforma que facilita el desarrollo de programas destinado al análisis de texto que brinda demostraciones detalladas de distintos algoritmos. Alberga una amplia colección de más de 50 conjuntos de datos y una gama de bibliotecas que simplifican significativamente tareas como segmentación, tokenización, etiquetado del habla y análisis tanto sintáctico como semántico. Su integración en Python resulta directa, ya que se encuentra disponible en el repositorio de lenguaje, permitiendo su acceso con una sola instrucción. Cabe mencionar que el etiquetador (POS) en NLTK puede ser ampliado al español mediante el entrenamiento con un corpus incorporado en dicho idioma [11].

Tensor Flow. Es un sistema de aprendizaje automático que funciona en entornos a gran escala y de forma heterogénea. Además, es una librería de código abierto para cálculo numérico que usa grafos de flujo de datos como forma de programación. Nació del trabajo de un grupo de investigadores dedicados a la Inteligencia Artificial llamados Google Brain. En 2017, Google librero el código con mejoras, es un software libre bajo licencia Apache 2, por ello tiene una comunidad muy extensa que constantemente está colaborando para implementar nuevas características [31].

2.9 Normas éticas

Esta investigación se sustenta en normativas éticas de acuerdo con los lineamientos y reglamentos de grados de la PUCESE. Durante todo el proceso de desarrollo de este proyecto se ha respetado a los autores dando reconocimiento por las ideas y conceptos que se han utilizado en la investigación.

3. CAPÍTULO III: RESULTADOS

3.1. Perfiles de autoría

Investigadores seleccionados

La selección de los investigadores se realizó por medio de un muestreo por conveniencia, debido a que favorecía a la investigación que los docentes seleccionados contaran con 5 o más artículos publicados en las plataformas determinadas.

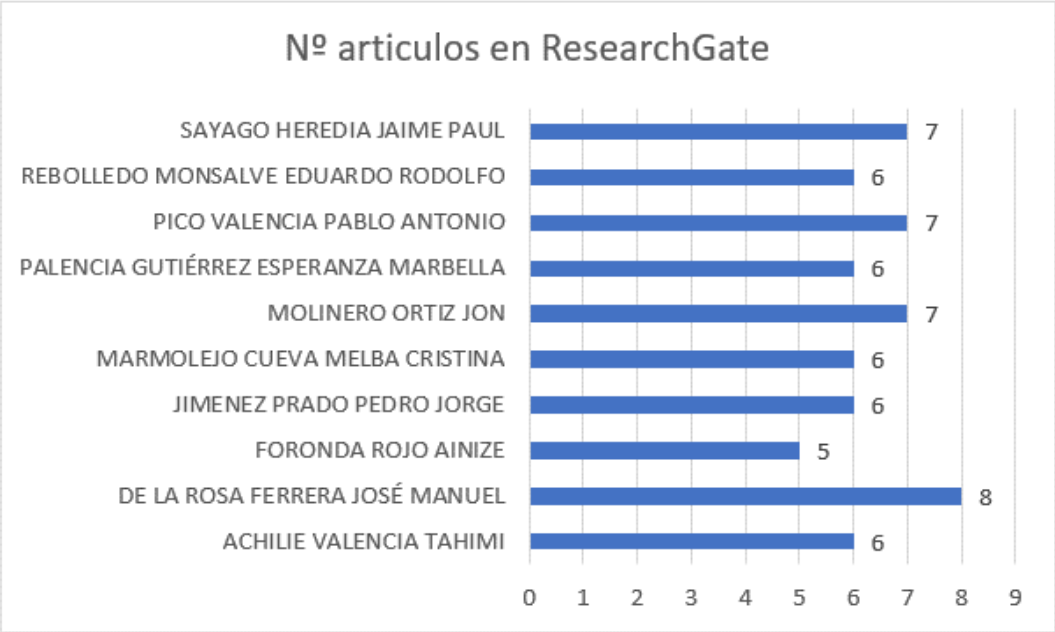


Figura 11. Docentes seleccionados por conveniencia

En la **Figura 11**¡Error! No se encuentra el origen de la referencia., muestra el número de artículos que tiene cada docente seleccionado en el muestreo por conveniencia ya que tienen más de 5 artículos publicados en ResearchGate.

Estrategia de perfil de autoría aplicada

En la parte del análisis de autoría se optó por utilizar el enfoque basado en perfil de autor el cual se realizó una concatenación todos los artículos de cada autor en un solo archivo csv el cual se convertirá en los datos de la fase de entrenamiento para crear los perfiles.

Siguiendo la distribución de los datos, en el archivos.csv está compuesto por dos columnas; “autor” y “texto”, en el cual cada texto lleva la etiqueta de un autor, por ejemplo, autor1, autor2 etc.

Documentos

Los documentos se obtuvieron del perfil de cada docente investigador en ResearchGate. Los artículos seleccionados de cada docente investigador de la universidad deben cumplir con ciertas características, las cuales son: ser artículos de revistas indexadas, artículos de memoria, capítulo de libros y no pueden ser tesis.

3.2. Diseño del clasificador

Set de datos

El dataset está compuesto por una base de datos de texto que contienen un total de 242 artículos científicos como podemos ver en la **Figura 12** . Para el conjunto de datos de entrenamiento se determinó usar el 80% de los datos lo que equivale a 189 artículos. Por consiguiente, el 20% restante se lo utilizo para realizar las pruebas necesarias y de esa forma poder medir la precisión del clasificador.

Cantidad total de texto	64
Total de texto de entrenamiento	45
Total texto de prueba	19

Figura 12. Resumen del dataset

En la **Figura 13**, se muestra la lista de los docentes seleccionados y la cantidad de artículos destinados al entrenamiento y a la prueba. Para cada docente se recopilo en un documento la introducción de cada artículo científico recopilado de los perfiles de ResearchGate de cada docente investigados.

AUTOR	Cantidad de articulos	Texto de entrenamiento (80%)	Texto de prueba(20%)
ACHILIE VALENCIA TAHIMI	6	4	2
DE LA ROSA FERRERA JOSÉ MANUEL	8	6	2
FORONDA ROJO AINIZE	5	4	1
JIMENEZ PRADO PEDRO JORGE	6	4	2
MARMOLEJO CUEVA MELBA CRISTINA	6	4	2
MOLINERO ORTIZ JON	7	5	2
PALENCIA GUTIÉRREZ ESPERANZA MARBELLA	6	4	2
PICO VALENCIA PABLO ANTONIO	7	5	2
REBOLLEDO MONSALVE EDUARDO RODOLFO	6	4	2
SAYAGO HEREDIA JAIME PAUL	7	5	2

Figura 13. Distribución de datos en datos de entrenamiento y prueba

Descripción del sistema

El sistema consta de dos partes como se muestra en la ¡Error! No se encuentra el origen de la referencia.¡Error! No se encuentra el origen de la referencia., la atribución de autoría y el modelo de predicción. La primera parte engloba todo lo que es el preprocesamiento de los datos, ya que se realiza la creación del dataset con los datos que se utilizan para el entrenamiento del modelo. En este proceso, primero se realiza el ingreso de los documentos del autor, luego de la recopilación de documentos, se selecciona el extracto del documento usado para el análisis, en este caso es la sección de la introducción de cada artículo científico.

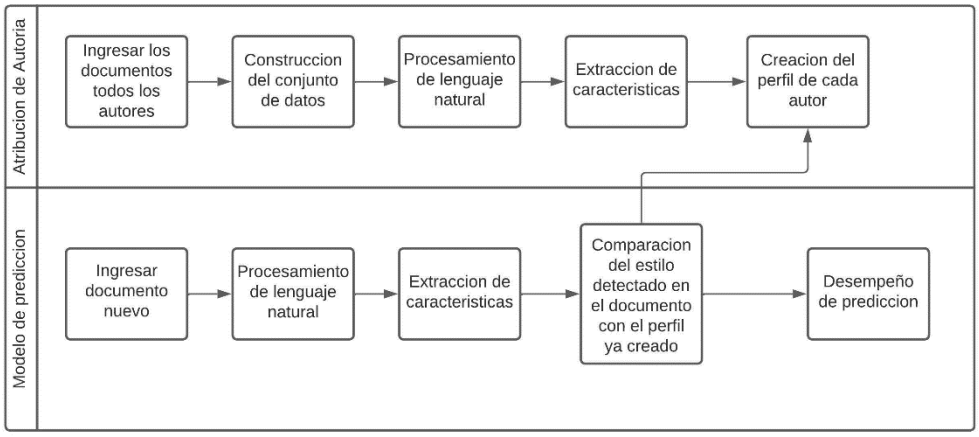


Figura 14. Diagrama del sistema

Creación y entrenamiento del modelo

La creación y entrenamiento del modelo en este proyecto es un proceso fundamental que abarca varias etapas clave. En esta sección, se explorará en detalle cómo se ha llevado a cabo la construcción y ajuste del modelo para lograr su capacidad de predicción.

Antes de que los datos puedan alimentar al modelo, es necesario realizar un preprocesamiento detallado para asegurarse de que el texto esté en un formato coherente y manejable. Esto involucra la tokenización, donde se divide el texto en palabras y signos de puntuación, la eliminación de stopwords para reducir el ruido en los datos y la lematización para llevar las palabras a su forma base.

```
# Preprocesamiento de texto en español
nltk.download('punkt')
nltk.download('wordnet')
nltk.download('stopwords')
nltk.download('words')
stop_words = set(stopwords.words('spanish'))
stemmer = SnowballStemmer('spanish')
```

```
def preprocesar_texto(texto):
    tokens = nltk.word_tokenize(texto)
    tokens = [token.lower() for token in tokens if token.isalpha() and token.lower() not in stop_words]
    tokens = [stemmer.stem(token) for token in tokens]
    return ' '.join(tokens)
```

Figura 15. Preprocesamiento del texto

Una vez que el texto ha sido preprocesado, es necesario convertirlo en una forma que el modelo pueda entender. Para lograr esto, se han utilizado dos enfoques clave: TF-IDF y Word Embeddings. La técnica TF-IDF asigna un valor numérico a cada palabra en función de su frecuencia en un documento y su importancia en el conjunto de datos. Por otro lado, los Word Embeddings, generados mediante algoritmos de entrenamiento profundo, capturan las relaciones semánticas entre palabras y permiten que el modelo comprenda el contexto en el que se utilizan.

```
# Crear el vectorizador TF-IDF
tfidf_vectorizer = TfidfVectorizer(max_features=1000)
X_tfidf = tfidf_vectorizer.fit_transform(textos_limpio)
```

```
# Modelo Word Embedding
model_w2v = Word2Vec(sentences=[word_tokenize(texto) for texto in textos_limpio], vector_size=100, window=5, min_count=1, sg=0)
```

```
# Función para obtener vector promedio de Word Embedding
def obtener_vector_promedio(texto, model):
    vectors = [model.wv[word] for word in word_tokenize(texto) if word in model.wv]
    if vectors:
        return np.mean(vectors, axis=0)
    return np.zeros(model.vector_size)  # Asegura que el vector tenga el mismo tamaño que el modelo
```

Figura 16. Técnica TF-IDF y Word Embeddings

La elección de los algoritmos de aprendizaje automático es una decisión crítica en la creación del modelo. En este proyecto, se han explorado varios algoritmos, como Máquinas de Vectores de Soporte Lineal, Máquinas de Vectores de Soporte RBF, Bosques Aleatorios y Regresión Logística. Cada algoritmo tiene sus propias fortalezas y debilidades, y la selección se basa en el tipo de problema y los datos disponibles.

```
# Entrenar y evaluar modelos
modelos = {
    'SVC Lineal': SVC(kernel='linear', probability=True),
    'SVC RBF': SVC(kernel='rbf', probability=True),
    'Random Forest': RandomForestClassifier(n_estimators=100),
    'Regresion Logistica': LogisticRegression()

}
```

Figura 17. Algoritmos utilizados en el modelo

El entrenamiento del modelo implica proporcionar los datos de entrenamiento, que consisten en las características del texto y sus etiquetas de estilo de redacción correspondientes. El modelo ajusta sus parámetros para aprender los patrones y las relaciones entre las características y las etiquetas. Durante este proceso, se realizan ajustes y se prueban diferentes configuraciones para maximizar la capacidad predictiva del modelo.

Una vez entrenado, el modelo se evalúa utilizando datos de prueba independientes que no ha visto antes. Se calculan métricas como precisión, recuperación y exactitud para medir su rendimiento. El resto del código fuente se lo puede encontrar en [1].

```
resultados = {}

for nombre, modelo in modelos.items():
    modelo.fit(X_train, y_train)
    y_pred = modelo.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred, average='weighted')
    recall = recall_score(y_test, y_pred, average='weighted')
    conf_matrix = confusion_matrix(y_test, y_pred)
    resultados[nombre] = {
        'Accuracy': accuracy,
        'Precision': precision,
        'Recall': recall,
        'Confusion Matrix': conf_matrix
    }
```

Figura 18. Métricas del modelo.

3.1 Métricas del clasificador

Se tomo en cuanta las métricas accuracy, recall y precisión. La exactitud o Accuracy es la cantidad de predicciones positivas que fueron correctas, la fórmula para calcularla es: $\frac{(VP+VN)}{(VP+FP+FN+VN)}$ en donde, VP corresponde a verdaderos positivos, mientras que VN significa verdaderos negativos, así mismo, FP es el número de falsos positivos, y FN son los falsos negativos. La sensibilidad o Recall, es la proporción de casos positivos que fueron correctamente identificados por el algoritmo, se calcula mediante: $\frac{VP}{(VP+FN)}$. Y, por

[1] https://github.com/sahory16/sistema_tesis.git

último, la métrica de precisión que equivale al porcentaje de los casos positivos detectados se calcula como: $\frac{VP}{(VP+FP)}$.

En la **Figura 19**, se presenta la matriz de confusión de los resultados de la evaluación de los modelos de clasificación. Cada fila de la matriz representa una categoría real, mientras que cada columna corresponde a la categoría predicha por el modelo. En este caso, las categorías corresponden a los estilos de redacción en los informes escritos pertenecientes a cada autor. A través de esta matriz, podemos observar cómo los modelos han asignado las instancias a cada categoría y evaluar su desempeño.

La métrica Accuracy o exactitud nos proporciona una visión general de la tasa de clasificación correcta en todos los casos. En esta evaluación, los modelos tienen tasas de exactitud que varían entre 0.55 y 0.75, lo que indica que algunos modelos tienen un mejor rendimiento en la clasificación que otros.

La precisión (Precision) es una métrica que mide la proporción de instancias clasificadas como positivas que realmente son positivas. En este contexto, los valores de precisión oscilan entre 0.54 y 0.80, lo que sugiere que los modelos están identificando correctamente instancias positivas en diferentes grados.

El Recall o recuperación, por otro lado, se centra en cuántas instancias positivas fueron identificadas correctamente. En este análisis, los valores de recuperación se encuentran en el rango de 0.55 a 0.75, lo que nos indica la capacidad de los modelos para identificar y capturar instancias positivas en diferentes proporciones.

La matriz de confusión también muestra cómo cada modelo ha clasificado incorrectamente algunas instancias. Estos errores pueden ser importantes para comprender las limitaciones y desafíos de los modelos en la tarea de clasificación. En este sentido, se puede observar la cantidad de falsos positivos y falsos negativos en cada caso, lo que proporciona información valiosa para futuras mejoras y ajustes de los modelos.

	Accuracy	Precision	Recall	Confusion Matrix
SVC Lineal	0.75	0.808333	0.75	[[2 0 0 0 0 0 0 0 0 0] [0 3 0 0 0 0 0 0 0 0] [1 0 1 0 0 0 0 0 0 0] [0 0 0 0 0 0 0 1 0 0] [0 0 0 0 2 0 0 0 0 0] [0 1 0 0 0 2 0 0 0 0] [0 2 0 0 0 0 1 0 0 0] [0 0 0 0 0 0 0 2 0 0] [0 0 0 0 0 0 0 0 2 0] [0 0 0 0 0 0 0 0 0 2]]
SVC RBF	0.6	0.75	0.6	[[2 0 0 0 0 0 0 0 0 0] [0 0 0 0 0 0 0 0 0 0] [0 0 3 0 0 0 0 0 0 0] [0 1 1 0 0 0 0 0 0 0] [0 1 0 0 0 0 0 0 0 0] [0 0 0 0 0 2 0 0 0 0] [0 0 1 0 0 0 2 0 0 0] [0 0 2 0 0 0 0 1 0 0] [0 0 1 0 0 0 0 0 1 0] [0 0 1 0 0 0 0 0 0 1]]
Random Forest	0.55	0.548333	0.55	[[1 0 0 0 0 0 0 1 0 0] [0 0 0 0 0 0 0 0 0 0] [0 0 3 0 0 0 0 0 0 0] [1 0 0 0 0 0 0 0 1 0] [0 0 0 0 0 0 0 0 1 0] [0 0 0 0 0 2 0 0 0 0] [0 1 0 0 0 0 2 0 0 0] [0 0 2 0 0 0 0 1 0 0] [0 1 0 0 0 0 0 0 1 0] [0 0 0 0 0 0 1 0 0 1]]
Regresion Logistica	0.6	0.79	0.6	[[2 0 0 0 0 0 0 0 0 0] [0 0 0 0 0 0 0 0 0 0] [0 0 3 0 0 0 0 0 0 0] [0 2 0 0 0 0 0 0 0 0] [0 1 0 0 0 0 0 0 0 0] [0 0 0 0 0 2 0 0 0 0] [0 1 0 0 0 0 2 0 0 0] [0 0 2 0 0 0 0 1 0 0] [0 1 0 0 0 0 0 0 1 0] [0 1 0 0 0 0 0 0 0 1]]

Figura 19. Evaluación del modelo

Las primeras cinco pruebas revelaron que el algoritmo Máquinas de Vectores de Soporte Lineal sobresalió al lograr una precisión del 89% al 99%, acertando en la clasificación de tres de los cinco autores. Esta capacidad para predecir abre la puerta a posibles aplicaciones valiosas. Por otro lado, la Regresión Logística también se mostró destacó alcanzando un rango de precisión del 69% al 81%, aunque ligeramente más bajo. Estos resultados iniciales señalan el potencial de estos algoritmos en la identificación de estilos de redacción.

Modelo	Autor Predicho	Probabilidad
SVC Lineal	autor1	0.465588
SVC RBF	autor2	0.195202
Random Forest	autor1	0.17
Regresion Logistica	autor1	0.771266

Prueba con nuevo documento del autor1

Modelo	Autor Predicho	Probabilidad
SVC Lineal	autor3	0.680478
SVC RBF	autor2	0.195202
Random Forest	autor3	0.17
Regresion Logistica	autor3	0.691549

Prueba con nuevo documento del autor3

Modelo	Autor Predicho	Probabilidad
SVC Lineal	autor5	0.890816
SVC RBF	autor2	0.195202
Random Forest	autor5	0.17
Regresion Logistica	autor5	0.821547

Prueba con nuevo documento del autor5

Modelo	Autor Predicho	Probabilidad
SVC Lineal	autor7	0.967337
SVC RBF	autor2	0.19235
Random Forest	autor10	0.25
Regresion Logistica	autor7	0.718298

Prueba con nuevo documento del autor7

Modelo	Autor Predicho	Probabilidad
SVC Lineal	autor9	0.991593
SVC RBF	autor2	0.19235
Random Forest	autor10	0.26
Regresion Logistica	autor9	0.719525

Prueba con nuevo documento del autor9

Figura 20. Pruebas al modelo con un artículo nuevo

Por otro lado, la evaluación de los algoritmos Random Forest y Máquinas de Vectores de Soporte RBF reveló un rendimiento por debajo del 30% en cuanto al grado de probabilidad de predicción, lo que se significa que realizó una clasificación incorrecta de los artículos en función de los estilos de redacción de los autores seleccionados. Estos resultados destacan la importancia de explorar alternativas y posibles ajustes para lograr una clasificación más precisa y confiable.

En la siguiente evaluación, cuyos resultados se muestran en la **Figura 21**, se eligieron artículos presentes en la base de datos, específicamente de los siguientes autores: autor2, autor4, autor6, autor8 y autor10. Esta selección permitió someter los algoritmos a una prueba más específica, con el propósito de comparar sus resultados con los obtenidos en la prueba previa. Al enfocarse en estos autores particulares, se buscó evaluar la capacidad

para predecir con precisión los estilos de redacción asociados al perfil creado en el entrenamiento de cada uno de ellos.

Modelo	Autor Predicho	Probabilidad
SVC Lineal	autor2	0.999905
SVC RBF	autor2	0.19235
Random Forest	autor2	0.31
Regresion Logistica	autor2	0.98708

Prueba con documento del dataset del autor2

Modelo	Autor Predicho	Probabilidad
SVC Lineal	autor4	0.999976
SVC RBF	autor2	0.195202
Random Forest	autor4	0.32
Regresion Logistica	autor4	0.999316

Prueba con documento del dataset del autor4

Modelo	Autor Predicho	Probabilidad
SVC Lineal	autor6	0.950552
SVC RBF	autor2	0.19235
Random Forest	autor10	0.27
Regresion Logistica	autor6	0.60866

Prueba con documento del dataset del autor

Modelo	Autor Predicho	Probabilidad
SVC Lineal	autor8	0.946989
SVC RBF	autor2	0.19235
Random Forest	autor10	0.25
Regresion Logistica	autor8	0.860725

Prueba con documento del dataset del autor8

Modelo	Autor Predicho	Probabilidad
SVC Lineal	autor10	0.964848
SVC RBF	autor2	0.19235
Random Forest	autor10	0.4
Regresion Logistica	autor10	0.935621

Prueba con documento del dataset del autor 10

Figura 21. Pruebas de modelo con un artículo de la base de datos.

En comparación con los resultados obtenidos en la prueba anterior, era previsible que los algoritmos Máquinas de Vectores de Soporte Lineal y Regresión Logística demostraran mejoras en su rendimiento. Estos dos algoritmos lograron rangos de precisión superiores al 85%, confirmando así la capacidad para clasificar los estilos de redacción. Sin embargo, en el caso particular de la prueba con el autor6, el algoritmo de Regresión Logística consiguió una correcta clasificación, aunque con un 60% de precisión, lo cual indica cierta dificultad en capturar las características distintivas de ese autor en particular.

Cabe mencionar que el algoritmo de Máquinas de Vectores de Soporte RBF obtuvo el menor rango de precisión en las pruebas realizadas, a pesar de tratarse de artículos que forman parte de la base de datos utilizada para entrenar el modelo. Este resultado indica que, a pesar de contar con información relevante para la clasificación, este algoritmo enfrenta desafíos al generalizar y aplicar los patrones de estilo a nuevos ejemplos. Estos hallazgos resaltan la importancia de abordar las limitaciones de cada algoritmo.

En la parte final de la evaluación cuyos resultados están en la **Figura 22**, se incluyeron dos artículos de autores cuyos estilos de redacción no habían sido previamente incorporados al modelo. Esta estrategia tenía como propósito poner a prueba el modelo en una identificación de estilos de redacción para los cuales no se tenía un perfil establecido a partir del proceso de entrenamiento. Al introducir estos artículos de autores desconocidos, se buscaba entender cómo el modelo respondía y generaba predicciones cuando se enfrentaba a estilos de escritura no familiares.

Modelo	Autor Predicho	Probabilidad
SVC Lineal	autor7	0.496155
SVC RBF	autor2	0.195202
Random Forest	autor10	0.21
Regresion Logistica	autor1	0.303728

Prueba con nuevo documento de autor sin perfil

Modelo	Autor Predicho	Probabilidad
SVC Lineal	autor7	0.4613
SVC RBF	autor2	0.19235
Random Forest	autor10	0.31
Regresion Logistica	autor7	0.246075

Prueba con nuevo documento de autor sin perfil

Figura 22. Pruebas del modelo con artículos de autores desconocidos

En el análisis de esta última prueba, los algoritmos presentaron un porcentaje de precisión inferior al 50%. Este resultado es evidente debido a que el perfil de los autores involucrados no estaba presente en la base de datos de entrenamiento. Ante esta limitación, cada algoritmo se dedicó a buscar similitudes entre los estilos de escritura de estos nuevos autores y aquellos que ya estaban registrados en el modelo.

Este resultado pone de manifiesto la dificultad que enfrentan los algoritmos cuando se encuentran con situaciones inéditas. La capacidad de los algoritmos para buscar similitudes con autores cuyo estilo no forma parte del conjunto de entrenamiento resalta su capacidad para abordar desafíos complejos. Aunque la precisión fue menor en esta instancia, esta fase de evaluación proporciona información crucial sobre las limitaciones y las áreas de mejora para el modelo. Revela la necesidad de explorar estrategias para manejar eficazmente autores no contemplados en el entrenamiento, lo que puede contribuir a mejorar la capacidad de adaptación y generalización del modelo en situaciones diversas.

4. CAPÍTULO IV: DISCUSIÓN

Después de realizar las pruebas y analizar los resultados obtenidos del modelo de clasificación de estilos de redacción, el siguiente paso es realizar una discusión comparativa con otras investigaciones para reforzar lo aprendido y de esa forma sentar las bases para futuros estudios.

Ángel Godoy [12] en su investigación "Técnicas de aprendizaje de máquina utilizadas para la minería de texto" sustenta que uno de los algoritmos más utilizados es el SVM(Maquina de vectores de soporte), demostrando que este algoritmo era utilizado en las publicaciones de estudio un 22.72% de las veces como técnica principal. En esta investigación, se utilizó el algoritmo SVM Lineal como técnica principal, acompañado de SVM RBF, Random Forest (Bosques Aleatorios) y Regresión Logística como técnicas secundarias.

En el estudio de Marcial Contreras [4], menciona que el uso de la minería de texto permitió desarrollar un clasificador automatizado, mediante el método de clasificación de documentos obteniendo un grado de precisión del 75% con el algoritmo Máquina de vectores de soporte. El modelo creado en esta investigación utilizó una combinación de las técnicas TF-IDF y Word Embeddings que le permitió capturar las relaciones semánticas entre palabras, así como el contexto en el que se utilizan. La precisión obtenida durante la etapa de entrenamiento fue de 80.8% mientras que, al realizar pruebas con un nuevo artículo de cada autor, el grado de precisión fue de 89% hasta el 99% con el algoritmo Máquina de Vectores de Soporte Lineal.

Jovany Ramirez, Maya Carrillo y Josefa Somodevilla [9] en su estudio "Atribución de autoría combinando información léxico-sintáctica mediante representaciones holográficas reducidas", presentan en su investigación dos enfoques como método para realizar la atribución de autoría, el primero es un enfoque basado en el perfil del autor, el cual consiste en concatenar los documentos de ese autor en uno solo para crear el perfil, mientras en el otro método se basa en instancias, es decir, se analizan todos los documentos por separado. Para el entrenamiento de este modelo, se optó por usar el enfoque basado en perfil, realizando una concatenación de todos los artículos de cada autor teniendo un total de 64 artículos, los cuales están etiquetados según corresponde a cada autor. De esta forma el modelo creó el perfil de cada autor y lo relaciono con las etiquetas correspondientes.

Erika Páez y Andrés Monroy [24] señala que en su investigación “Implementación de un modelo de análisis de sentimientos con respecto a la JEP basado en minería de datos en Twitter”, realizaron una implementación de un módulo que permite analizar los sentimientos de los datos de Twitter mediante la minería de texto. Las métricas que utilizaron para medir el desempeño del modelo fueron: precisión, recall y F1-Score. En contraste a el estudio antes mencionado, en esta investigación se utilizó: la métrica accuracy ya que nos proporcionó una visión general de la tasa de clasificación correcta en todos los casos, precision que permitió medir la proporción de instancias clasificadas como positivas de forma correcta y recall que indico cuántas instancias positivas fueron identificadas correctamente.

5. CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

En conclusión, este proyecto ha proporcionado una valiosa visión sobre la efectividad de diversos algoritmos de aprendizaje automático en la tarea de categorizar textos en función de sus características lingüísticas y estilísticas.

Durante esta investigación se pudieron recolectar 64 artículos en total pertenecientes a 10 diferentes autores. De este conjunto de datos se utilizó el 80% para el entrenamiento y los 20% restantes para realizar las pruebas. Por medio del análisis de métricas como precisión, recuperación y exactitud aplicadas al conjunto de datos, se ha obtenido un panorama claro de cómo aborda cada algoritmo los desafíos presentados en la clasificación de estilos de escritura.

El proyecto resalta la relevancia y la eficacia de combinar técnicas avanzadas, como TF-IDF y Word Embeddings, con algoritmos de clasificación, para abordar problemas complejos de clasificación de texto. Los resultados han demostrado que el uso de diferentes algoritmos puede influir significativamente en el rendimiento del modelo. Por ejemplo, se observó que el modelo de "Máquina de Vectores de Soporte Lineal" logró un alto nivel de precisión del 80.8% en su etapa de entrenamiento, mientras que, en las pruebas realizadas al modelo con artículos nuevos, el grado de precision se mantuvo en un rango de 89% hasta el 99%, destacando su capacidad para capturar con precisión instancias tanto positivas como negativas. Por otro lado, el modelo "Máquina de Vectores de Soporte RBF" mostró una dificultad y limitación en la clasificación de los datos nuevos.

Estos resultados permitieron dar respuesta a una de las preguntas de la investigación la cual cuestionaba; ¿cuán efectivo puede llegar a ser un sistema de predicción de estilos de redacción académica basado en las técnicas de minería de texto y aprendizaje automático? En base a estos datos, el sistema propuesto puede alcanzar un nivel de precisión relativamente confiable y efectivo. Sin embargo, se recomienda considerar cuidadosamente el equilibrio en la base de datos, y la elección del algoritmo para esta tarea.

5.2. Recomendaciones

El desarrollo de esta investigación proporciona una premisa para que otros investigadores profundicen más en el tema de la atribución de autoría, ya que es un campo que está en auge y en el cual aún falta mucho por poner a prueba.

Para la creación del dataset se recomienda obtener la mayor cantidad de datos posibles ya que de eso dependerá la efectividad y precisión del algoritmo. Sin embargo, hay que tener en cuenta que el número de artículos por cada autor sea equilibrado, para que no haya tanta dispersión en los datos que ocasione resultados poco eficientes.

En el caso específico de la utilización del corpus en español se recomienda revisar que realmente el corpus elegido sea en ese lenguaje porque en esta área en la que se combinaron el lenguaje de procesamiento del lenguaje natural, minería de texto y aprendizaje automático con la atribución de autoría son pocos los ejemplos puestos en práctica en dicho idioma. La mayoría de los artículos utilizaban datasets en inglés por lo cual las opciones para usar los corpus son variadas en dicho idioma.

En futuras investigaciones, se puede considerar la incorporación de más algoritmos de clasificación, así como la exploración de enfoques avanzados de procesamiento de lenguaje natural, como redes neuronales y modelos de transformadores.

En resumen, este proyecto ha proporcionado una base sólida para comprender y abordar la clasificación de estilos de redacción en informes académicos. Los resultados obtenidos brindan información valiosa para la toma de decisiones informadas y la mejora continua de los modelos en futuras investigaciones en este campo de la minería de texto y el procesamiento de lenguaje natural.

ANEXOS

Anexo 1, lista de docentes investigadores de la PUCESE

#	DOCENTES INVESTIGADORES	Perfil en Scholar	Nª de articulos	Perfil en ResearchGate	Nª de articulos
1	ACHILIE VALENCIA TAHIMI			https://www.researchgate.net/profile/Tahimi-Achilie-Valencia	11
2	AGUILAR QUIÑONEZ VERONICA ELIZABETH			https://www.researchgate.net/profile/Veronica-Aguilar-15	5
3	ANGULO ALEGRIA MEMI VICTORIA			https://www.researchgate.net/profile/Memi-Angulo	1
4	ARIAS CARPIO WENDY TATIANA			https://www.researchgate.net/profile/Wendy-Arias-Carpio	0
5	BAUTISTA SEGOVIA ALEXANDRA MARIBEL			https://www.researchgate.net/profile/Alexandra-Bautista-Segovia	0
6	BERNABE LILLO MARIA DE LOS LIRIOS			https://www.researchgate.net/profile/Maria-Bernabe-Lillo	2
7	BETANCOURT AGUILAR FREDDY LEONARDO				
8	CERVANTES INTRIAGO JENNY CRISTINA				
9	CHANGO SAILEMA WILSON GUSTAVO	https://schola	2		
10	CHILA GODOY MARIA ISABEL			https://www.researchgate.net/profile/Isabel-Chila	3
11	CID GARCÍA MAGDALENA AMALIA				
12	CISNEROS PALACIOS JOHANNA GABRIELA			https://www.researchgate.net/profile/Johanna-Cisneros-Palacios	4
13	CORONEL TENORIO ANA KARINA			https://www.researchgate.net/profile/Ana-Coronel-Tenorio	2
14	DE LA ROSA FERRERA JOSÉ MANUEL			https://www.researchgate.net/profile/Jose-De-La-Rosa-Ferrera	7
15	DUEÑAS MENDOZA ANDREA ESTEFANÍA				
16	ESPINOZA ÁVILA MANUEL EDUARDO			https://www.researchgate.net/profile/Manuel-Espinoza-Avila	0
17	FORONDA ROJO AINIZE	https://schola	9	https://www.researchgate.net/profile/Ainize-Foronda-2	7
18	HIDALGO TAPIA COSME ENRIQUE			https://www.researchgate.net/profile/Cosme-Hidalgo	1
19	IPIALES MIRANDA GABRIEL ALEJANDRO				
20	JIMENEZ PRADO PEDRO JORGE			https://www.researchgate.net/profile/Pedro-Jimenez-Prado	60
21	JIMENEZ YANEZ SILVIA FERNANDA				
22	MALDONADO LIRA BEATRIZ MAGALY				
23	MARMOLEJO CUEVA MELBA CRISTINA	https://schola	15	https://www.researchgate.net/profile/Melba-Cristina-Marmolejo-Cu	6
24	MÁRQUEZ ARBOLEDA VIVIANA MERCEDEZ			https://www.researchgate.net/profile/Viviana-Marquez-Arboleda	2
25	MILA CARVAJAL DIONICIO FRANCISCO				
26	MOLINERO ORTIZ JON			https://www.researchgate.net/profile/Jon-Molinero-Ortiz	49
27	MURILLO DIAZ RICARDO ANDRÉS			https://www.researchgate.net/profile/Ricardo-Murillo-Diaz	0
28	NARANJO CORRÍA REBECA				
29	OCAMPO ERAZO HUGO			https://www.researchgate.net/profile/Hugo-Ocampo-Erazo	0
30	OVIEDO VERA AIDA YISELA				

31	PALENCIA GUTIÉRREZ ESPERANZA MARBELLA			https://www.researchgate.net/profile/Esperanza-Marbella-Palencia-	5
32	PATIÑO CAGUA CELESTE ANNE				
33	PEÑA ROSAS GLORIA			https://www.researchgate.net/profile/Gloria-Pena-Rosas	4
34	PEREZ LABORDE ELENA JOHANNA			https://www.researchgate.net/profile/Elena-Perez-Laborde	0
35	PICO VALENCIA PABLO ANTONIO	https://schola	36	https://www.researchgate.net/profile/Pablo-Pico-Valencia	33
36	PINEDA CAICEDO ANDREA ESTEFANÍA			https://www.researchgate.net/profile/Andrea-Caicedo-5	0
37	PRIETO FUENMAYOR JULIA ROSA			https://www.researchgate.net/profile/Julia-Prieto-Fuenmayor	0
38	PUPO SUÑOL ANGEL EDUARDO			https://www.researchgate.net/profile/Angel-Pupo-Sunol	0
39	QUIÑÓNEZ PERLAZA GUADALUPE				
40	QUIROZ PONCE FREDDY HERNAN	https://schola	8		
41	RAMIREZ LOZADA HAYDEE			https://www.researchgate.net/profile/Haydee-Ramirez-Lozada	4
42	REBOLLEDO MONSALVE EDUARDO RODOLFO			https://www.researchgate.net/profile/Eduardo-Rodolfo-Rebolledo-M	18
43	REYES ARBOELDA PAULETT STEFANIA				
44	SANCHEZ FLORES ESTEFANIA			https://www.researchgate.net/profile/Estefania-Sanchez-18	6
45	SÁNCHEZ GUERRA IÑIGO				
46	SAYAGO HEREDIA JAIME PAUL	https://schola	37	https://www.researchgate.net/profile/Jaime-Sayago-Heredia	25
47	SOLIS MURILLO MARIA DE LOURDES			https://www.researchgate.net/profile/Maria-Solis-Murillo	2
48	SOSA BONE MIRNA PATRICIA			https://www.researchgate.net/profile/Mirna-Sosa-Bone	0
49	TORRES RODRIGUEZ MARIA TERESA	https://schola	12		
50	VELAZCO VARGAS JORGE LUIS				
51	VERA PINARGOTE SINAY CELESTE				
52	VERGARA TORO MÓNICA ANGELINA				
53	VERGARA TORRES JOSÉ LUIS				
54	VERNAZA QUIÑONEZ LUCIA MARGARITA				
55	VILA MAFFARE MARILYN MARIVEL	https://schola	2		
56	VILLARREAL PUGA JOSUE				
57	VINUEZA CHÉRREZ RUBEN ANDRES			https://www.researchgate.net/profile/Ruben-Vinueza-2	5
58	YANEZ ORTIZ VERONICA DEL CARMEN			https://www.researchgate.net/profile/Veronica-Yanez-Ortiz	5

REFERENCIAS

[1] M. V. Isabel Ana, “Tecnicas estadisticas en Mineria de Textos,” 2017.

[2] C. E. Castillo, “Implementación de técnicas para minería de texto usando modelos de tópicos.,” Escuela Politécnica Nacional, Quito, 2019.

[3] A. Consuegra, Y. M. Salazar, J. H. Garc, and D. H. Vizcaino, “Minería de texto como una herramienta para la búsqueda de artículos científicos para la investigación,” pp. 14–20, 2016.

[4] M. C. Barrera, “Minería de texto en la clasificación de material bibliográfico,” *Biblios*, vol. 0, no. 64, pp. 33–43, Nov. 2016, doi: 10.5195/biblios.2016.309.

[5] J. Alberto and G. Gutiérrez, “Minería de textos , aprendizaje automático y procesamiento del lenguaje natural en el entorno biosanitario.,” no. June, pp. 1–16, 2017, [Online]. Available: https://www.researchgate.net/publication/317668652_Mineria_de_textos_aprendizaje_automatico_y_procesamiento_del_lenguaje_natural_en_el_entorno_biosanitario.

[6] V. Ibero, a D. L. Recursos, and M. La A, “Analítica del aprendizaje: método automático para identificar sentencias que contienen información positiva y negativa utilizando técnicas de minería de texto,” *Researchgate.Net*, vol. 3, pp. 119–122, 2016, [Online]. Available: http://www.researchgate.net/publication/258644061_Libro_de_Actas_del_V_Foro_Iberomericano_de_los_Recursos_Marinos_y_la_Acuicultura/file/3deec528c8c45982e7.pdf.

[7] V. Mercado, A. Villagra, and M. Errecalde, “El Proceso de Extracción de Conocimiento en la Determinación del Perfil del Autor y la Atribución de Autoría,” pp. 261–265, 2017.

[8] D. C. Castro, “Verificación de autoría, modelos intrínsecos basados en semejanza,” Universidad de Alicante, Alicante, 2019.

[9] J. Marcos Ramírez, M. Carillo Ruíz, and M. Josefa Somodevilla, “Atribución de autoría combinando información léxico- sintáctica mediante representaciones

holográficas reducidas," *Res. Comput. Sci.*, pp. 103–113, 2014.

[10] E. Macías and A. Dennise, "Análisis del uso del procesamiento del lenguaje natural y su aplicación en sistemas conversacionales," Universidad de Guayaquil. Facultad de Ciencias Matemáticas y Físicas. Carrera de Ingeniería en Networking y Telecomunicaciones, Guayaquil, 2019.

[11] L. Talamé, A. Cardoso, and M. Amor, "Comparación de herramientas de procesamiento de textos en español extraídos de una red social para Python," *XX Simposio Argentino de Inteligencia Artificial (ASAI 2019) - JAIIO 48 (Salta)*, pp. 53–67, Sep. 2019.

[12] A. F. Godoy Viera, "Técnicas de aprendizaje de máquina utilizadas para la minería de texto," *Investig. Bibl.*, vol. 31, no. 71, pp. 103–126, Jan. 2017, doi: 10.22201/iibi.0187358xp.2017.71.57812.

[13] J. Felipe and G. Évora, "Aplicaciones de la minería de texto en la Encuesta Nacional de Transparencia 2019: una nueva alternativa de análisis para las encuestas de percepción," Universidad de Costa Rica, 2020.

[14] M. A. Ruiz Osorio, "Implementación de un sistema de diálogo automático como asistente en el proceso administrativo del examen de traductor e intérprete oficial de la Universidad de Antioquia," Universidad de Antioquia, 2020.

[15] G. Liberatore, A. Vuotto, and G. V. Fernández, "Desarrollo de una herramienta para el análisis y representación semántica de colecciones documentales a través del factor TF-IDF," *VII Jornadas Temas Actuales en Bibl.*, Jan. 2016, [Online]. Available: http://humadoc.mdp.edu.ar:8080/xmlui/handle/123456789/631.

[16] A. Cardoso, L. Talame, M. Amor, and A. Monge, "Aplicación de técnicas avanzadas de aprendizaje automático para identificar emociones en textos Este proyecto continúa la línea de investigación que el," pp. 73–77, 2021.

[17] D. Cedeno-Moreno and M. Vargas, "Aprendizaje automático aplicado al análisis de sentimientos," *I+D Tecnológico*, vol. 16, no. 2, Jul. 2020, doi: 10.33412/idt.v16.2.2833.

[18] D. Cedeno Moreno and M. Vargas, "Aprendizaje automático aplicado al análisis de sentimientos," *I+D Tecnológico*, vol. 16, no. 2, 2020, doi:

10.33412/idt.v16.2.2833.

[19] J. Z. Ruiz, “Comparativa y análisis de algoritmos de aprendizaje automático para la predicción del tipo predominante de cubierta arbórea,” Universidad Complutense de Madrid, Madrid, 2018.

[20] F. A. Ozbay and B. Alatas, “Fake news detection within online social media using supervised artificial intelligence algorithms,” *Phys. A Stat. Mech. its Appl.*, vol. 540, pp. 1–21, Feb. 2020, doi: 10.1016/J.PHYSA.2019.123174.

[21] R. Rojas, “Modelo de Aprendizaje Automático Supervisado para Identificar Patrones de Bajo Rendimiento Académico en los Ingresantes al Instituto de Educación Superior Pedagógico Público – Juliaca,” 2021.

[22] G. M. Martínez Toro, D. R. Bautista, and E. R. Riaño, “Análisis comparativo de predicción dentro de bases de datos de cáncer: una aplicación de aprendizaje automático.,” no. 17, pp. 113–122, Nov. 2018, Accessed: Aug. 24, 2022. [Online]. Available: https://www.researchgate.net/publication/331178491.

[23] R. Blanco Toledano, “Identificación de lenguaje misógino a partir de minería de textos en redes sociales,” Jun. 2021.

[24] E. P. Páez Guarnizo and A. F. Monroy, “Implementación de un modelo de análisis de sentimientos con respecto a la JEP basado en minería de datos en Twitter,” Universidad Católica de Colombia. Facultad de Ingeniería. Programa de Ingeniería de Sistemas, Bogota, 2020.

[25] C. NacionalL, *Ley de Propiedad Intelectual*. 2018, pp. 1–77.

[26] República del Ecuador, *Constitución de la República del Ecuador*, vol. 449, no. 20. Registro Oficial 449 , 2008, pp. 25–2021.

[27] COESC, “Código Orgánico de la Economía Social de los Conocimientos, Creatividad e Innovación,” 2016.

[28] E. M. Flores Nessi, A. C. Loaiza Falcón, and G. N. Rojas de Ricardo, “Rol del docente investigador desde su práctica social,” *Rev. Sci.*, vol. 5, no. 15, pp. 106–128, Feb. 2020, doi: 10.29394/SCIENTIFIC.ISSN.2542-2987.2020.5.15.5.106-128.

[29] A. Sarasa, *Gestión de la información web usando Python* -. Barcelona: Editorial UOC (Oberta UOC Publishing, SL), 2016.

[30] J. Walker, *Python: La Guía Definitiva para Principiantes para Dominar Python - Jonathan S. Walker - Google Libros*. Babelcube, Inc., 2018.

[31] J. Zamorano, "Comparación y análisis de métodos de clasificación con las bibliotecas scikit-learn y TensorFlow en Python," Escuela Técnica Superior de Ingeniería Informática, Málaga, 2019.