UNIVERSIDAD DE GUAYAQUIL

FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

DETERMINACIÓN DE GÉNERO Y PROFESIÓN DE USUARIOS DE TWITTER UTILIZANDO ESTILOMETRÍA CON PALABRAS DE USO FRECUENTE DEL ESPAÑOL Y MÉTODOS DE CLASIFICACIÓN DE MACHINE LEARNING

PROYECTO DE TITULACIÓN

Previa a la obtención del Título de:

INGENIERO EN SISTEMAS COMPUTACIONALES

AUTORES:

AUCAPIÑA CAMAS CARLOS ISMAEL

PAZMIÑO ROSALES MARIA BELÉN

TUTOR:

ING. CÉSAR ESPÍN RIOFRÍO, MSC.

GUAYAQUIL – ECUADOR

2022

REPOSITORIO NACIONAL EN CIENCIAS Y TECNOLOGÍAS	
FICHA DE REGISTRO DE TRABAJO DE TITULACIÓN	
TÍTULO: *"Determinación de género y profesión de usuarios de Twitter utilizando estilometría con palabras de uso frecuente del español y métodos de clasificación de Machine Learning"*	
AUTOR(ES): Aucapiña Camas Carlos Ismael Pazmiño Rosales María Belén	**REVISOR(A):** Ing. Jorge Avilés Monroy, M. Sc.
INSTITUCIÓN: Universidad de Guayaquil	**FACULTAD:** Ciencias Matemáticas y Físicas
CARRERA: Ingeniería en Sistemas Computacionales	
FECHA DE PUBLICACIÓN:	**N° DE PAGS:** 104
AREA TEMÁTICA: *Investigación, Machine Learning*	
PALABRAS CLAVES: *Machine Learning, Estilometría, Métodos de Clasificación, Atribución de Autoría, Validación Cruzada.*	
RESUMEN: *El objetivo de este artículo es determinar el género y la profesión de los usuarios de Twitter en Ecuador, mediante el análisis de características estilométricas y técnicas de Machine Learning (ML) para la Atribución de Autoría. El proyecto corresponde a un tipo de investigación cuantitativa-bibliográfica, con diseño experimental realizada en lenguaje de programación Python, en el ambiente de prueba Google Colab. Su desarrollo consiste inicialmente en extraer 6000 tweets de 120 usuarios, que serán divididos 5000 para entrenamiento y 1000 para pruebas. Luego, para el pre-procesamiento de la información se implementa características de tipo fraseológicas y de frecuencia de palabras utilizando el listado CREA proporcionado por la Real Academia Española. Posteriormente se entrena los cinco métodos clasificadores escogidos: Ramdom Forest (RF), Decision Tree (DT), Logistic Regression (LR), MultiLayer Perceptron (MLP) y Gradient Boosting (GB), y se evalúa su comportamiento a través de la técnica de Validación Cruzada y métricas de precisión para escoger el clasificador ideal para la predicción del género y profesión. Por último, se presentan los resultados en conductas observables y medibles. En este caso, se determinó que Random Forest obtuvo 0.63 de precisión al predecir el género y, el método MLP Classifier un 0.84 de precisión para la profesión, superando al resto de clasificadores. En conclusión, el estudio de esta investigación es de gran interés, debido a que aplica métodos tecnológicos actuales y brinda soluciones óptimas en atribución de autoría para textos cortos.*	
N° DE REGISTRO:	**N° DE CLASIFICACIÓN:**
DIRECCIÓN URL: (PROYECTO DE TITULACION EN LA WEB)	

ADJUNTO PDF	**SI** x	**NO** ☐
CONTACTO CON AUTORES:	**Teléfono:** 0998618314 0988886262	Email: maria.pazminor@ug.edu.ec carlos.aucapinac@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 Tutor del Trabajo de Titulación, "DETERMINACIÓN DE GÉNERO Y PROFESIÓN DE USUARIOS DE TWITTER UTILIZANDO ESTILOMETRÍA CON PALABRAS DE USO FRECUENTE DEL ESPAÑOL Y MÉTODOS DE CLASIFICACIÓN DE MACHINE LEARNING" elaborado por los Sr. AUCAPIÑA CAMAS CARLOS ISMAEL y PAZMIÑO ROSALES MARIA BELÉN, **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,

Ing. César Espín Riofrío, M. Sc.

TUTOR

DEDICATORIA

A mis padres, que han sido el pilar de mis éxitos, espero que estén orgullosos de lo que he alcanzado que es gracias a ellos.

Aucapiña Camas Carlos Ismael

A Dios, por las fuerzas y las ganas para culminar un proceso más en mi vida.

Pazmiño Rosales María Belén

AGRADECIMIENTO

Agradezco a Dios y a mis padres que se han esforzado por darme la educación.

Aucapiña Camas Carlos Ismael

A mi familia, por creer en mí y apoyarme incondicionalmente.

Pazmiño Rosales María Belén

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

Ing. Cesar Humberto Espín Riofrío,
M.Sc.
PROFESOR TUTOR DEL
PROYECTO
DE TITULACIÓN

Ing. Jorge Avilés Monroy, M. Sc.
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".

AUCAPIÑA CAMAS CARLOS ISMAEL

PAZMIÑO ROSALES MARIA BELÉN

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 "**Determinación de género y profesión de usuarios de Twitter utilizando estilometría con palabras de uso frecuente del español y métodos de clasificación de Machine Learning**", realizado como requisito previo para la obtención del Título de Ingeniero(a) en Sistemas Computacionales de la Universidad de Guayaquil.

Guayaquil, 04 de octubre de 2022.

Aucapiña Camas Carlos Ismael
C.I. N° 0929040202

Pazmiño Rosales María Belén
C.I. N° 0932190689

UNIVERSIDAD DE GUAYAQUIL

FACULTAD DE CIENCIAS MATEMÁTICAS Y FÍSICAS

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

DETERMINACIÓN DE GÉNERO Y PROFESIÓN DE USUARIOS DE TWITTER UTILIZANDO ESTILOMETRÍA CON PALABRAS DE USO FRECUENTE DEL ESPAÑOL Y MÉTODOS DE CLASIFICACIÓN DE MACHINE LEARNING

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

INGENIERO EN SISTEMAS COMPUTACIONALES

Autores: Aucapiña Camas Carlos Ismael

C.I. N° 0929040202

Pazmiño Rosales María Belén

C.I. N° 0932190689

Tutor: Ing. César Espín Riofrío, M.Sc.

Guayaquil, 04 de octubre de 2022.

CERTIFICADO DE ACEPTACIÓN DEL TUTOR

En mi calidad de Tutor 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 **AUCAPIÑA CAMAS CARLOS ISMAEL, PAZMIÑO ROSALES MARÍA BELÉN**, como requisito previo para optar por el Título de Ingeniero en Sistemas Computacionales cuyo proyecto es:

DETERMINACIÓN DE GÉNERO Y PROFESIÓN DE USUARIOS DE TWITTER UTILIZANDO ESTILOMETRÍA CON PALABRAS DE USO FRECUENTE DEL ESPAÑOL Y MÉTODOS DE CLASIFICACIÓN DE MACHINE LEARNING

Considero aprobado el trabajo en su totalidad.

Presentado por:

Aucapiña Camas Carlos Ismael **CI:** 0929040202

Pazmiño Rosales María Belén **CI:** 0932190689

Tutor(a): Ing. César Espín Riofrío, M.Sc.

Firma

Guayaquil, 04 de octubre de 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: Aucapiña Camas Carlos Ismael	
Dirección: Bastion Popular	
Teléfono: 0988886262	**Email:** carlos.aucapinac@ug.edu.ec

Nombre del Estudiante: Pazmiño Rosales María Belén	
Dirección: Sauces 6	
Teléfono: 0998618314	**Email:** maria.pazminor@ug.edu.ec

Facultad: Facultad de Ciencias Matemáticas y Físicas
Carrera: Carrera de Ingeniería en Sistemas Computacionales
Proyecto de Titulación al que opta: Ingeniero en Sistemas Computacionales
Profesor(a) Tutor(a): Ing. Cesar Espín Riofrío, M. Sc.

Título del Proyecto de Titulación: DETERMINACIÓN DE GÉNERO Y PROFESIÓN DE USUARIOS DE TWITTER UTILIZANDO ESTILOMETRÍA CON PALABRAS DE USO FRECUENTE DEL ESPAÑOL Y MÉTODOS DE CLASIFICACIÓN DE MACHINE LEARNING

Palabras Claves: Machine Learning, Estilometría, Métodos de Clasificación, Atribución de Autoría, Validación Cruzada.

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	**x**	Después de 1 año	

Firma Estudiante:

Aucapiña Camas Carlos Ismael — **CI:** 0929040202

Pazmiño Rosales María Belén — **CI:** 0932190689

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 [x] CDROM [ ]

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ABREVIATURAS

UG	Universidad de Guayaquil
AA	Atribución de Autoría
IA	Inteligencia Artificial
ML	Machine Learning
PLN	Procesamiento de Lenguaje Natural
RAE	Real Academia Española
CREA	Corpus de Referencia del Español Actual
RF	Ramdom Forest
DT	Decision Tree
LR	Logistic Regression
MLP	MultiLayer Perceptron
GB	Gradient Boosting

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

DETERMINACIÓN DE GÉNERO Y PROFESIÓN DE USUARIOS DE TWITTER UTILIZANDO ESTILOMETRÍA CON PALABRAS DE USO FRECUENTE DEL ESPAÑOL Y MÉTODOS DE CLASIFICACIÓN DE MACHINE LEARNING

Autores: Aucapiña Camas Carlos Ismael
C.I. N° 0929040202
Pazmiño Rosales María Belén
C.I. N° 0932190689

Tutor: Ing. César Espín Riofrío, M.Sc.

RESUMEN

El objetivo de este artículo es determinar el género y la profesión de los usuarios de Twitter en Ecuador, mediante el análisis de características estilométricas y técnicas de Machine Learning (ML) para la Atribución de Autoría. El proyecto corresponde a un tipo de investigación cuantitativa-bibliográfica, con diseño experimental realizada en lenguaje de programación Python, en el ambiente de prueba Google Colab. Su desarrollo consiste inicialmente en extraer 6000 tweets de 120 usuarios, que serán divididos 5000 para entrenamiento y 1000 para pruebas. Luego, para el pre-procesamiento de la información se implementa características de tipo fraseológicas y de frecuencia de palabras utilizando el listado CREA proporcionado por la Real Academia Española. Posteriormente se entrena los cinco métodos clasificadores escogidos: Ramdom Forest (RF), Decision Tree (DT), Logistic Regression (LR), MultiLayer Perceptron (MLP) y Gradient Boosting (GB), y se evalúa su comportamiento a través de la técnica de Validación Cruzada y métricas de precisión para escoger el clasificador ideal para la predicción del género y profesión. Por último, se presentan los resultados en conductas observables y medibles. En este caso, se determinó que Random Forest obtuvo 0.63 de precisión al predecir el género y, el método MLP Classifier un 0.84 de precisión para la profesión, superando al resto de clasificadores. En conclusión, el estudio de esta investigación es de gran interés, debido a que aplica métodos tecnológicos actuales y brinda soluciones óptimas en atribución de autoría para textos cortos.

Palabras clave: Machine Learning, Estilometría, Métodos de Clasificación, Atribución de Autoría, Validación Cruzada.

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

DETERMINACIÓN DE GÉNERO Y PROFESIÓN DE USUARIOS DE TWITTER UTILIZANDO ESTILOMETRÍA CON PALABRAS DE USO FRECUENTE DEL ESPAÑOL Y MÉTODOS DE CLASIFICACIÓN DE MACHINE LEARNING

Autores: Aucapiña Camas Carlos Ismael
C.I. N° 0929040202
Pazmiño Rosales María Belén
C.I. N° 0932190689

Tutor: Ing. César Espín Riofrío, M.Sc.

ABSTRACT

The objective of this article is to determine the gender and profession of Twitter users in Ecuador, through the analysis of stylometric characteristics and Machine Learning (ML) techniques for Authorship Attribution. The project corresponds to a quantitative-bibliographic type of research, with experimental design carried out in Python programming language, in the Google Colab test environment. Its development consists initially in extracting 6000 tweets from 120 users, which will be divided 5000 for training and 1000 for testing. Then, for the pre-processing of the information, phraseological and word frequency type features are implemented using the CREA list provided by the Real Academia Española. Subsequently, the five chosen classifier methods are trained: Ramdom Forest (RF), Decision Tree (DT), Logistic Regression (LR), MultiLayer Perceptron (MLP) and Gradient Boosting (GB), and their performance is evaluated through the Cross Validation technique and accuracy metrics to choose the ideal classifier for gender and profession prediction. Finally, the results are presented in observable and measurable behaviors. In this case, it was determined that Random Forest obtained 0.63 accuracy in predicting gender and the MLP Classifier method obtained 0.84 accuracy for profession, surpassing the rest of the classifiers. In conclusion, the study of this research is of great interest, because it applies current technological methods and provides optimal solutions in authorship attribution for short texts.

Key words: Machine Learning, Stylometry, Classification Methods, Authorship Attribution, Cross-Validation.

INTRODUCCIÓN

La atribución de autoría (AA) es la encargada de responder quién es el autor de un texto, dando algunos ejemplos previos de ese autor (Castillo Velásquez, Godoy Martínez, Zavala De Paz, et al., 2021). Desde hace mucho tiempo, los trabajos de clasificación han dado buenos resultados para textos largos, sin embargo, el estudio de textos cortos ha quedado aplazado. De modo que en este trabajo de investigación se propone un análisis de las características estilométricas de tipo fraseológicas y de frecuencia de palabras en conjunto con técnicas de Machine Learning, para determinar el género y la profesión de 120 usuarios de la red social Twitter en Ecuador.

En el capítulo I, se explica el planteamiento del problema a investigar, identificando sus causas, antecedentes y estado actual. En este caso, el análisis del texto para la clasificación de usuarios de la red social Twitter según su género y profesión, adicionalmente se detalla el alcance, objetivos, limitaciones e importancia del proyecto.

En el capítulo II, se define el marco teórico y los recursos que se serán utilizados a lo largo del proyecto, como investigaciones o teorías que se consideren válidos para el estudio y que signifique un aporte en el análisis del problema a investigar.

En el capítulo III, se describe la metodología empleada en un dataset proveniente de las publicaciones de los usuarios de Twitter, con la finalidad de extraer datos reales que puedan ser sometidos a análisis y alcanzar un alto nivel de precisión.

Finalmente, en el capítulo IV, se presenta las conclusiones que dan respuesta a los objetivos planteados inicialmente, así como las recomendaciones que deben contemplarse para futuros proyectos.

CAPÍTULO I

PLANTEAMIENTO DEL PROBLEMA

Descripción de la situación problemática

Ubicación del problema en un contexto

Establecer la atribución de autoría es considerado el principal problema de esta investigación, pues se busca determinar el género y profesión del autor de un texto mediante el entendimiento, representación y comparación de patrones que cada autor tiene como característica a la hora de escribir un texto en español.

En este proceso se construye un perfil lingüístico del autor basándose en la forma que este escribe, la omisión o la frecuencia en que utiliza ciertas palabras, por consiguiente, surge la necesidad de contar con mecanismos automáticos para facilitar el análisis de dicha información.

El analizar esta problemática representa un beneficio para las empresas u organizaciones que se interesan en conocer las necesidades de un conjunto de usuarios ecuatorianos, en este caso de la red social Twitter, Ecuador.

Situación conflicto nudos críticos

La información ubicada generalmente en redes sociales tiene un alto potencial en la investigación de mercados y publicidad dirigida, sin embargo, cuando se pretende emplear con la finalidad de elaborar perfiles de usuarios de una manera no consentida e indiscriminada esta puede verse afectada por la falta de privacidad y anonimato al revelar información sensible.

En la era de la omnipresente web social, la información personal se intercambia, cientos de millones de personas pasan interconectadas a diversas comunidades online y redes sociales: en este proceso, revelan todo tipo de detalles sobre sí mismos. De ahí el gran interés que suscita la investigación sobre aspectos de la recuperación de información y el procesamiento del lenguaje natural, con el fin de extraer la información relevante para su propósito de la fuente de conocimiento que representa esta cantidad masiva de datos en línea. (Kokkos & Tzouramanis, 2014)

Delimitación del problema

El presente proyecto se encuentra en el campo de Estilometría y Atribución de Autoría, pertenece al área de Inteligencia Artificial & Robótica, y su respectivo aspecto es la clasificación de usuarios según el texto que publican.

Tabla 1

Delimitación del problema

Delimitador	Descripción
Campo	Estilometría y Atribución de Autoría
Área	Inteligencia Artificial & Robótica
Aspecto	Clasificación de usuarios según el texto que publican
Tema	Determinación de género y profesión de usuarios de Twitter utilizando estilometría con palabras de uso frecuente del español y métodos de clasificación de Machine Learning

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

- **Delimitado:** Mediante el uso del algoritmo de Regresión Logística, se mostrará la eficacia de las técnicas usadas y el comportamiento del mismo para la detección de género y profesión de textos en español.
- **Claro:** Pretende establecer una clasificación de los usuarios de Twitter según su género y profesión.
- **Concreto:** Procura clasificar a usuarios según su género y profesión a partir de un texto publicado usando algoritmos de Machine Learning.
- **Relevante:** El impacto de los resultados del estudio, podría representar en gran beneficio a empresas u organizaciones que pretenden optimizar tiempo, costos, evitando métodos tradicionales.
- **Evidente:** Las organizaciones en algunos casos desconocen el porcentaje de distribución de usuarios por género y profesión de una base datos que han obtenido.
- **Factible:** Permite extraer información clasificada usando metodologías conocidas que no representan un elevado costo económico y que a su vez genera beneficios para las empresas u organizaciones que desean saber los intereses de los usuarios según su género y profesión en Ecuador.

Causas y consecuencias del problema

Tabla 2

Matriz de causas y consecuencias del problema

Causas	Consecuencias
C1. No conocer la audiencia de usuarios en base a sus publicaciones en internet.	E1. Inconvenientes al realizar campañas específicas a un público objetivo que se desconoce su género y profesión.
C2. Desconocimiento de métodos o algoritmos de Machine Learning para análisis de texto.	E2. No se aprovecha la potencialidad de obtener características de los usuarios.
C3. Cantidad insuficiente de mensajes de Twitter que se puedan extraer.	E3. El aprendizaje de los métodos de clasificación puede resultar poco aproximado.
C4. Texto no limpio, porque incluye símbolos , emoticones, enlaces web, etc.	E4. Dificultad en el entrenamiento y análisis en ese texto.
C5. Características estilométricas apropiadas para la clasificación.	E5. El análisis estilométrico puede verse afectado de acuerdo con el tipo de características obtenidas de los textos.
C6. Uso de métodos de clasificación acordes al análisis de textos necesarios.	E6. La elección del método utilizado afecta al aprendizaje automático para la clasificación de los textos.

Nota: Esta tabla presenta las causas y consecuencias con respecto a la elaboración del proyecto.

Formulación del problema

¿Es posible determinar el género y la profesión de usuarios de Twitter analizando características estilométricas de los mensajes que publican, con técnicas de Machine Learning para la Atribución de Autoría?

Objetivos del proyecto

Objetivo general

Determinar el género y profesión de usuarios de Twitter de Ecuador analizando características de uso de palabras frecuentes del español para estilometría y técnicas de machine learning para atribución de autoría.

Objetivos específicos

1. Establecer el estado del arte de estilometría y métodos de clasificación de machine learning para determinar el género y profesión de usuarios según los tweets publicados.
2. Determinar las técnicas de estilometría y machine learning usadas en determinación de género y profesión mediante el análisis de artículos científicos de relevancia.
3. Entrenar métodos de clasificación de machine learning para predecir el género y profesión de usuarios de Twitter.
4. Evaluar los métodos utilizados en la clasificación de machine learning para establecer el de mejores resultados en la tarea propuesta.

Alcance del proyecto

El proyecto está orientado a la detección de género y profesión de usuarios de textos en español y el uso de la atribución de autoría en la misma, como caso de estudio se recopilará los mensajes publicados en la red social Twitter, Ecuador. Los Tweets serán extraídos por medio de la API (*Tweepy*, n.d.).

Para el entendimiento de los algoritmos y métodos de Machine Learning, se establecerá características de estilométricas, fraseológicas, longitud y frecuencia de palabras. Se trata de un método de clasificación lingüística que goza de mucha popularidad en los últimos años, aunque sus orígenes se remontan a mediados de los sesenta. Se fundamenta en la idea de que

el autor de un texto imprime siempre en sus creaciones una huella estilística o autorial, un estilo propio que puede ser rastreado por medio de métodos cuantitativos (Peñarrubia Navarro, 2021).

Su ejecución tiene como objetivo proporcionar la información necesaria para su debido análisis siguiendo la guía de artículos científicos de relevancia. Se utilizará algunas de las técnicas de Machine Learning como: Regresión logística, Random Forest Classifier, Decision Tree Classifier, MLPClassifier, Gradient Boosting Classifier, estos algoritmos sirven para llevar a cabo la Atribución de Autoría, en especial, enfoques de aprendizaje supervisado utilizando un conjunto de dataset para el entrenamiento de los métodos para la predicción de género y profesión.

Con la finalidad de ver que métodos resulta más factible al momento de predecir, se evaluará cada uno de ellos con la técnica de Validación Cruzada. Cabe resaltar, que este modelo de estudio será desarrollado en Python aplicando un enfoque generativo y no parametrizado.

Justificación e importancia

Varios estudios han demostrado que la atribución de autoría puede ayudar en casos complejos como estudios forenses, acoso vía internet, atribución de libros o artículos científicos. Por ejemplo, con el uso de estas herramientas las organizaciones han podido identificar y beneficiarse conociendo cuáles son las tendencias, afinidades o gustos que tiene un cliente hacia determinado producto.

Aun así, esta investigación es de gran interés, debido a que aplica los mismos métodos tecnológicos actuales para brindar soluciones óptimas en la atribución de autoría con la diferencia que utilizará textos cortos.

Limitaciones del estudio

Inicialmente, para la recopilación de información se utilizó la API de Twitter (Tweepy) contemplando extraer datos de 120 usuarios, 50 tweets por cada uno de ellos. Teniendo en total un dataset de 6000 tweets los cuales serán seleccionados 5000 para la fase de entrenamiento y 1000 para pruebas. En el proceso de entrenamiento se etiquetan los datos dependiendo del género (hombre o mujer) y tipo de profesión, que para este caso será únicamente (periodistas o políticos) que pertenezcan a la red social Twitter en Ecuador.

CAPÍTULO II

MARCO TEÓRICO

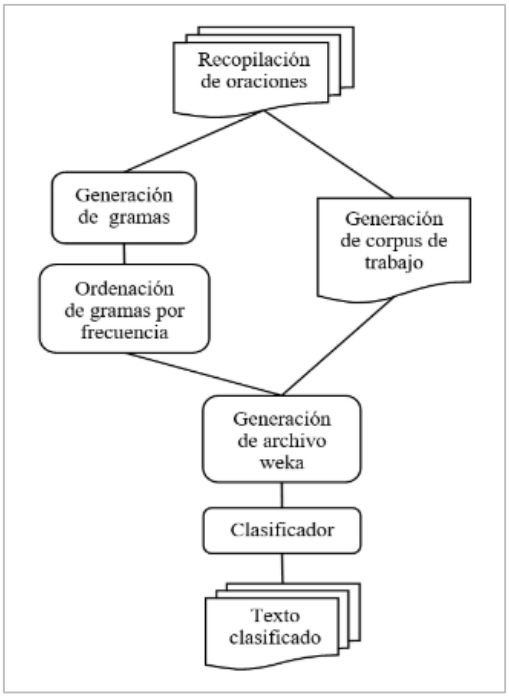
Antecedentes del estudio

En el estudio de la atribución de autoría, se encuentran varias investigaciones como la de (Castillo Velásquez et al., 2020) la cual contribuye a la verificación de la factibilidad de aplicación del análisis sintáctico automático en el proceso de AA para textos cortos, para su ejemplificación con mensajes de la red social Twitter. Sus experimentos consistieron en la generación de n-gramas (mediante una herramienta libre de extracción); la compilación de los n-gramas hace referencia al almacenamiento dinámico de los n-gramas únicos (haciendo uso de estructuras matriciales y la implementación de un algoritmo para su manipulación).

En diferentes casos la problemática de la distinción de género surge de solo encontrar referencias en inglés como dice el siguiente autor: “Los estudios hechos están dirigidos a resolver el problema del género, pero la mayoría de estas investigaciones están limitadas al inglés y a los medios tradicionales”. (Castillo Velásquez, Godoy Martínez, Zavala de Paz, et al., 2021) Este autor propone que la caracterización de la autoría puede definirse como la tarea de asignar los escritos de un autor a un conjunto de categorías de acuerdo con un perfil sociolingüístico. Algunos atributos analizados previamente en la literatura son el género, nivel de educación, idioma y antecedentes culturales.

Figura 1

Modelo propuesto para la identificación de género de textos cortos



Nota: Etapas del modelo propuesto se detallan en la investigación de (Castillo Velásquez, Godoy Martínez, Zavala de Paz, et al., 2021)

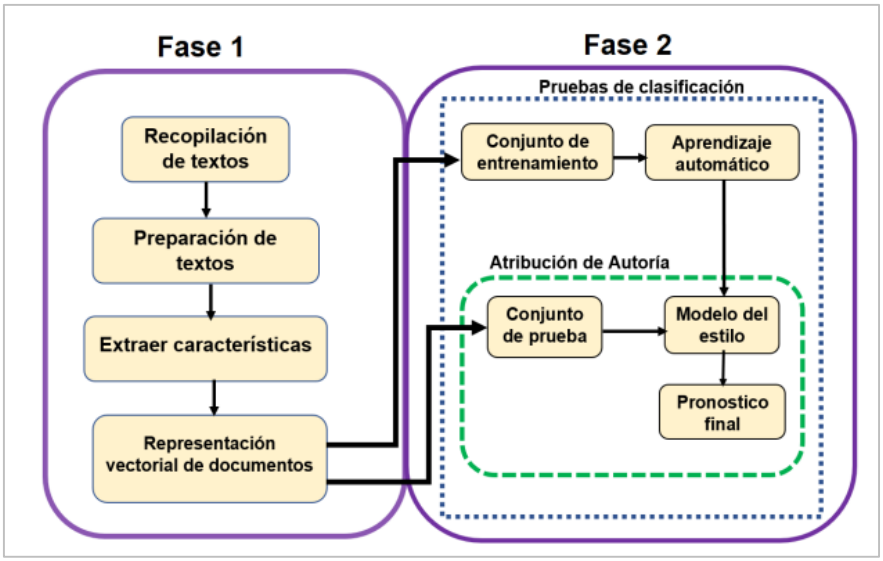
(Brito et al., 2020) aporta con un nuevo enfoque para atribución de autoría, este propone utilizar un método de clasificación de textos extrayendo las características de un solo documento del conjunto de entrenamiento sin la implementación de métodos de selección de características, analizando este nuevo enfoque con características léxicas (n-gramas y bolsa de palabras), con una representación booleana y máquina de soporte vectorial como método de aprendizaje supervisado.

En México, existen diferentes artículos de publicaciones sobre el uso de métodos tanto estilométricos y de aprendizaje supervisado donde la metodología propuesta consta de 2 fases, ver Figura 2.

La primera fase consiste en obtener las características estilométricas y la segunda en aplicar un enfoque de aprendizaje automático supervisado para clasificar documentos nuevos mediante las características seleccionadas en la etapa previa. A continuación, se describen detalladamente cada una de las actividades realizadas en cada etapa. (Favián Gutiérrez Constantino, 2020)

Figura 2

Diagrama a bloques de la propuesta.



Nota: Etapas que se realiza en la búsqueda de obras literarias para el análisis estilométrico.(Favián Gutiérrez Constantino, 2020)

El método de Regresión Logística es de gran utilidad al momento de predecir, tal es caso de la investigación hecha por estudiantes de nuestra universidad donde proponen una manera de predecir terremotos usando Regresión Logística. (Alba & Calle, 2020) afirman que: el algoritmo de Regresión Logística obtuvo una precisión del 72.25% siendo así un porcentaje mayor al 70% mínimo requerido para poder aplicarse y el valor de la curva ROC o AUC es aproximadamente 70.4% dando a entender que los datos por encima de ese valor tienden a ser

clasificadas correctamente, puesto que a medida que se acerquen más a 1, los datos tendrán una mejor clasificación (Alba & Calle, 2020).

El método de Random Forest es usando en una investigación de la universidad de Sao Paulo donde se ajustó este clasificador para usar una función de separación lineal con el factor de regularización C=1, el clasificador Random Forest utilizó 500 estimadores sin limitaciones por el número de hojas o profundidad del árbol, la versión puesta a disposición por la biblioteca Scikit-Learn utiliza el método de muestreo para crear cada árbol de decisión. (Otávio & Ferreira Frediani, 2022) utilizaron el clasificador Naive Bayes en su versión multinomial proporcionado por la biblioteca Scikit-Learn, la investigación propuesta tuvo una evaluación donde todos los clasificadores obtuvieron los mejores rendimientos en los conjuntos de datos que contenían la mayor cantidad de tweets por autor. Entre los clasificadores clásicos, dos tuvieron el mejor desempeño con todas las técnicas de extracción de características y el clasificador Random Forest logró su mejor desempeño con sólo el 4 gramo a nivel de búsqueda.

(Alejandro & Antón, 2014) de la Universidad de Valladolid, emplearon el clasificador MLP Classifier, para su realización fijaron un número de 80 características, 40 de ellas TAGS y las otras 40 FW. Se evaluaron las 15 posibles combinaciones de dos autores a partir de los 6 escogidos. Para cada una de ellas se recoge por separado la precisión de cada modelo: SVM, MLP y Naive Bayes.

Algunas investigaciones, como (Jiménez León & Martínez Vera, 2021) busca analizar diferentes patrones de Machine Learning para una predicción de fallos en áreas axiales de un recipiente toroidal de almacenamiento, en donde obtuvieron como resultado que con KNN (Vecinos Más Cercanos) se obtuvo una precisión del 62%,con el Decision Tree se obtuvo un 64%, siendo los modelos de más bajo porcentaje de predicción, mientras que SVM y Regresión Logística ofrecen un porcentaje muy aceptable que se encuentra entre el 73% y 80 %, además de esto Random Forest y Gradient Boosting ofrecen mejor precisión que los otros modelos en

el momento de predecir el fallo en las áreas axiales de un recipiente de forma toroidal de sección recta circular.

Una de las técnicas más utilizada en las investigaciones es la Validación Cruzada, ya que este método puede englobar clasificadores que ayudan al entrenamiento y predicción del objetivo. En la investigación propuesta por (Patricia Araujo Arredondo, 2009) usó la validación cruzada en 10 pliegues, con las herramientas que proporciona WEKA pudo observar que el resultado con mayor exactitud utilizando bolsa de palabras como atributos, es utilizando todas las palabras (frecuencia de 1 o más veces), logrando una exactitud de 63.27% con una validación cruzada y 60.92% en el conjunto de prueba.

Fundamentación teórica

Procesamiento de Lenguaje Natural (PLN)

"El PLN consiste en la utilización de un lenguaje natural para comunicarnos con la computadora, facilita el desarrollo de programas que realicen tareas relacionadas con el lenguaje o bien, desarrollar modelos que ayuden a comprender los mecanismos humanos relacionados con el lenguaje" (Vásquez et al., 2009).

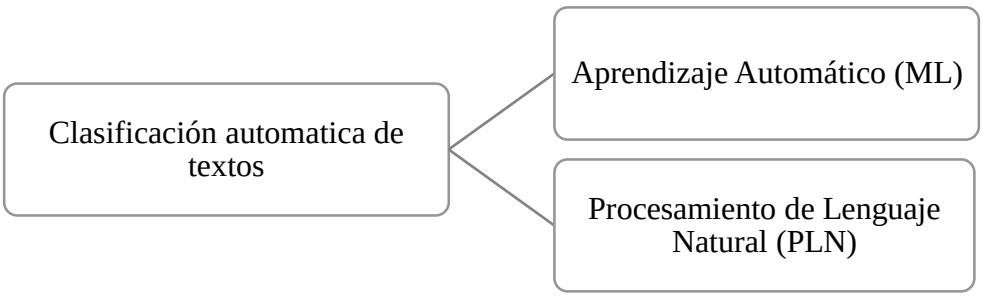
En los últimos años, la lingüística computacional se ha convertido en una apasionante área de investigación científica y práctica tecnología que se está incorporando cada vez más a los productos de consumo (por ejemplo, en aplicaciones como Siri de Apple y Traductor de Skype). Uno de los buscadores más potentes del mundo como lo es Google dispone de una herramienta llamada "Google Cloud Natural Languaje" basado en servicio de PLN que sirve para sustraer información de textos sin estructurar, donde el usuario puede proveer el contenido en diversos lenguajes multimedia, así como documentos de texto, diálogo de audio o imágenes en el que se realiza labores de análisis de entidades, extracción de conceptos clave del texto, análisis de sentimiento, detección de lenguaje o de clasificación de textos en más de 700 categorías. Además, dispone de métodos de entrenamiento basados en Machine Learning para

mejorar el servicio, los cuales tienen la posibilidad de ser creados por el usuario (Quevedo Marcos, 2020).

La clasificación automática, consiste en un conjunto de algoritmos, técnicas y sistemas capaces de asignar un documento a una o varias categorías o grupos de documentos, construidos según su afinidad temática. Para ello, véase Figura 3, se emplean técnicas de Aprendizaje Automático y de Procesamiento de Lenguaje Natural donde el PLN estudia los inconvenientes inherentes al procesamiento y manipulación de lenguajes naturales, efectuando uso de computadoras en el cual pretende conseguir conocimiento sobre el modo en que los humanos comprenden y usan el lenguaje, de tal forma que se logre realizar el desarrollo de herramientas y técnicas para lograr que las computadoras puedan comprenderlo y manipularlo, además sus fundamentos residen en un grupo bastante extenso de disciplinas como pueden ser en ciencias de la información, matemáticas, IA y robótica, entre otros. Y, por otro lado, sobre el ML y sus técnicas cubren tareas como el análisis sintáctico y morfológico de los textos, extracción de información, clasificación automática de documentos, agrupación semántica, entre otras (Pérez et al., 2017).

Figura 3

Clasificación automática de texto



Nota: Clasificación grafica automática de textos. Elaboración propia.

Atribución de Autoría (AA).

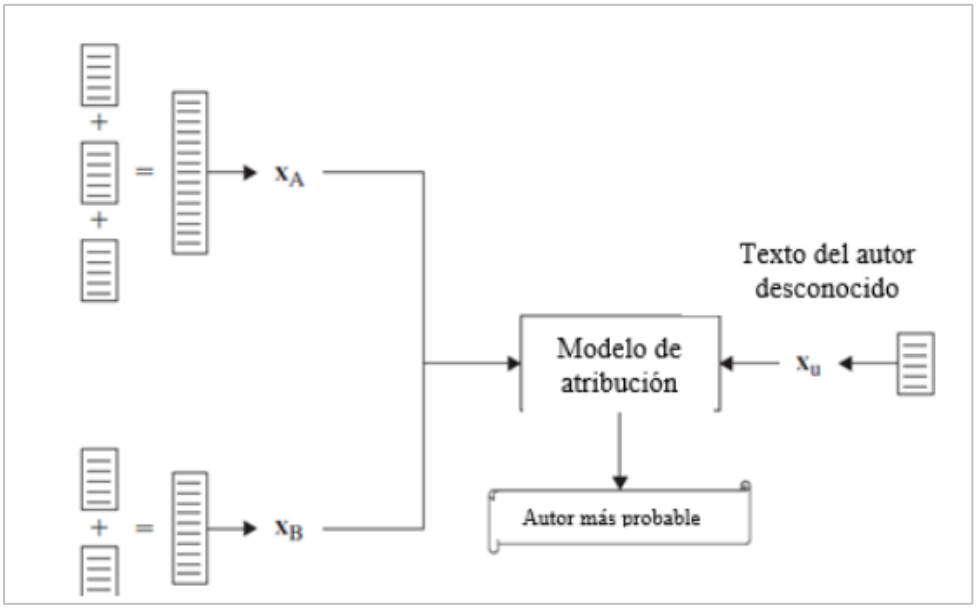
Para (Abad-García, 2019) la AA es la ciencia que estudia características textuales aplicando el conocimiento lingüístico para obtener variables lingüísticas, que analizándolas se puede determinar que autor le corresponde sea tanto de documentos o evidencias de texto escrito en cuestión. Según (Mosteller & Wallace, 1963) la AA ha sido muy estudiado en la literatura, donde el objetivo es determinar si un escrito no visto anteriormente se puede atribuir a ese autor. El problema se ha estudiado de diferentes formas con conjuntos de datos de diferentes tamaños y tipos. Las técnicas de análisis estilométrico se han utilizado para atribuir la autoría en el pasado y actualmente también.

Ahora bien, la idea principal detrás de la AA es que al extraer algunas características textuales medibles (características de estilometría), podemos discriminar al verdadero autor del texto dado que comparte características similares. Existen diferentes tipos de estilometría con una serie de características utilizadas en las obras de AA incluyen características léxicas, sintácticas, estructurales y de contenido (Stamatatos, 2009).

Existen dos formas para realizar AA, 1) el enfoque basado en el perfil del autor, en este se concatenan todos los documentos de un autor presentes en el conjunto de entrenamiento, para crear su perfil. Esto se realiza extrayendo varias características principalmente de bajo nivel, tales como n-gramas de caracteres. Para predecir la autoría de un documento nuevo, se debe calcular la similitud entre los perfiles de autor generados y las características del nuevo documento. Posteriormente el documento de autoría desconocida se asignará al autor, cuyo perfil tuvo la mayor similitud con él (Pastor López-Monroy et al., 2012), su estructura típica es mostrada en la Figura 4.

Figura 4

Atribución de Autoría combinando información léxico-sintáctica mediante representaciones.

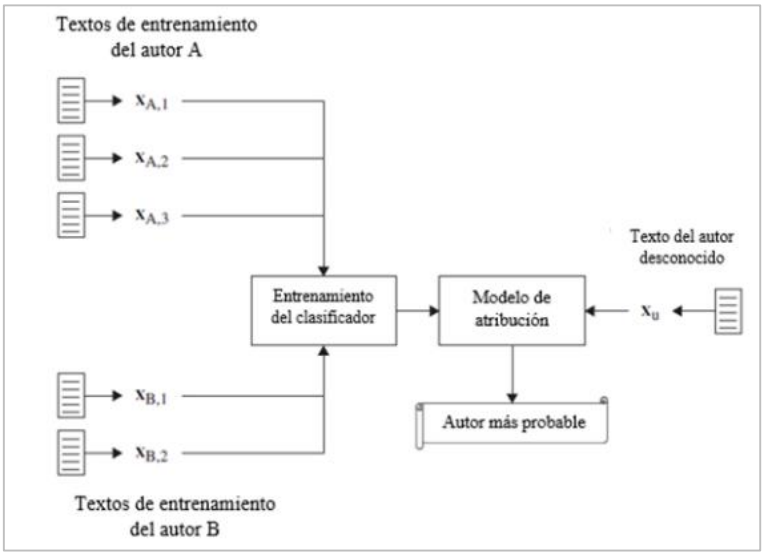


Nota: Enfoque basado en el perfil de autor. Tomado de (Marcos Ramírez et al., n.d.)

2) En contraste el enfoque basado en máquina de aprendizaje, utilizan una representación vectorial, donde cada documento se representa de forma individual por un conjunto de características. Dichos vectores serán utilizados para entrenar un algoritmo de aprendizaje automático. Estos vectores suelen contener características variadas, desde caracteres, longitud de palabras, n-gramas de caracteres, n-gramas de palabras (Pastor López-Monroy et al., 2012), su estructura típica podremos observarla en la Figura 5.

Figura 5

Enfoque basado en instancias



Nota: Enfoque basado en instancias. Tomado de (Marcos Ramírez et al., n.d.)

Estilometría

La estilometría, que en la actualidad es una de las áreas de investigación en las humanidades digitales con mayor desarrollo, es una metodología estadística para analizar textos a través de sus palabras más frecuentes (Most Frequent Words o MFW). Si bien sus bases fueron establecidas en 1890 por el filósofo polaco Wincenty Lutosławski en el libro "Principios de Estilometría", el desarrollo de la tecnología y su capacidad de procesar y analizar grandes cantidades de datos ha impulsado su uso, especialmente con la inteligencia artificial y el análisis estadístico (Owen, 2022).

Diferentes investigadores han sostenido la propuesta de que un estilo propio y único de expresión del ser humano proviene desde tiempos antiguos en los cuales famosos científicos, entre ellos matemáticos, físicos y filósofos han establecido distintos parámetros según los cuales se podría inferir la identificación de una persona. La estilometría siempre es aplicada con un propósito concreto, que de forma universal persigue identificar un patrón o curva

característica de escritura relacionada a la forma de expresión de un escritor. En otras palabras, la estilometría busca generar un perfil de escritura por medio de métodos computacionales, el mismo que debe provenir de la recolección de datos, que incluyen el recuerdo de preferencias e información recolectada a partir de interacciones previas (Bourcier, 2001).

Entre las aplicaciones de la estilometría, las principales son los estudios literarios, históricos, sociales y de género, así como análisis e investigaciones forenses, que permiten determinar la autenticidad y atribución de los mismos. Y, es que, teniendo en cuenta que cada persona tiene sus propias tendencias a la hora de escribir, esos rasgos crean una huella dactilar textual o, lo que es lo mismo, unos marcadores lexicales de estilo que actúan como un patrón (a veces más reconocible que otras) que sirve para determinar la autoría o clasificar diferentes textos (Bourcier, 2001).

Características Estilométricas

Una característica estilométrica hace referencia a un rasgo de la forma que un autor compone sus textos. La siguiente tabla muestra la clasificación de algunas características estilométricas de acuerdo con el tipo de información lingüística que representan.

Tabla 3

Clasificación de características estilométricas

Tipo de característica	Descripción
Caracteres	Alfabéticos y numéricos, letras mayúsculas y minúsculas, marcas de puntuación, *n*-gramas a nivel de carácter.
Léxicas	Longitud de palabras, longitud de sentencias, riqueza de vocabulario, hapaxes (palabras que aparecen una y dos veces), palabras frecuentes, *n*-gramas de palabras, errores de escritura.
Sintácticas	Etiquetas POS, estructura de sentencia y frase, racimos, reglas de reescritura.
Semánticas	Sinónimos, hiperónimos, dependencias semánticas.
Específicas de aplicación	Estructurales (características HTML como tipo, tamaño y color de la letra.), específicas de contenido (tipo y número de emoticones utilizados.), especificas del lenguaje (forma de saludo y despedida, tipo de firmas, indentación, modismos)

Nota: Tabla de clasificación de características estilométricas según su tipo (Germán Ríos et al., 2019).

Algunas de las características léxicas empleadas para el caso de estudio son:

Características Fraseológicas

Según (Española, 2022) define a la Fraseología como la parte de la lingüística que estudia las frases, los refranes, los modismos, los proverbios y otras unidades de sintaxis total o parcialmente fija. De manera más simple, podemos definirla como la disciplina que estudia las unidades fraseológicas.

El papel de la fraseología en la unidad textual ha sido subrayado por varios autores (J.C., 1992), expone que en la comprensión de los textos la fraseología mejora nuestra percepción de las diferencias en un texto:

1. La fraseología puede ser un potente marcador del tipo de texto, dentro de los sublenguajes de un área.
2. Las unidades fraseológicas nos ayudan a identificar las áreas mezcladas y los aspectos de los textos. Por ejemplo, una unidad fraseológica (UF) puede indicar jerga legal como específica del área específica de la medicina legal.
3. El reconocimiento de las UF puede afectar la velocidad de lectura.

Según (Ruiz Gurillo, 1997) tres principales características: la fijación, la lexicalización y la idiomaticidad. La fijación se refiere a la propiedad que muestran estas combinaciones prefabricadas de palabras donde el hablante únicamente reproduce una secuencia sin modificaciones o con cambios muy leves, como en de cabo a rabo, de pascuas a ramos, de buenas a primeras.

La lexicalización, de acuerdo con (Ruiz Gurillo, 1997), se trata de “un fenómeno presente en los lexemas complejos que, debido a su uso frecuente, tienden a convertirse en unidades léxicas simples, con pérdida, por consiguiente, de su carácter sintagmático”, esta característica

puede observarse en la locución en un pis pas, que se interpreta con el significado de ‘en un momento’ únicamente cuando sus elementos léxicos se encuentran unidos.

Finalmente, la idiomaticidad, según (Pastor, 1997), se refiere a “aquella propiedad semántica que presentan ciertas unidades fraseológicas, por la cual el significado global de dicha unidad no es deducible del significado aislado de cada uno de sus elementos constitutivos”. Un ejemplo de una UFS con un alto grado de idiomaticidad se encuentra en la expresión dar calabazas a alguien, que no guarda ninguna relación con el fruto y significa más bien según el DLE “loc. verb. coloq. Desairarlo o rechazarlo cuando requiere de amores” (Española, 2022).

Características de Frecuencia de palabras

Una lista de frecuencia de palabras es básicamente un listado de todas las palabras diferentes encontradas en una muestra de material escrito u oral con sus respectivas frecuencias de ocurrencia.

Para este caso de estudio se utilizará dicha característica estilométrica basada en el Corpus de Referencia del Español Actual (CREA) que es un conjunto de textos de diversa procedencia, almacenados en soporte informático, del que es posible extraer información para estudiar las palabras, sus significados y sus contextos.

Machine Learning

Para (Raschka & Kaufman, 2020) el aprendizaje automático -la ciencia de los algoritmos que da sentido a los datos- es el campo más significativo de todas las ciencias computacionales. En la actualidad los datos llegan en abundancia, pero gracias a los algoritmos de auto-aprendizaje es posible convertir esos datos en conocimiento, utilizando múltiples y potentes librerías de código abierto que han sido desarrolladas en la última década.

Es una especialidad en el campo de la IA (Inteligencia Artificial) que se ocupa de crear algoritmos que posean la capacidad de aprender por si solos sin tener que programarlos de una manera detallada, es decir, el tiempo de desarrollo del programador es optimizado altamente gracias al Machine Learning donde lo principal que se debe realizar es abastecer con grandes volúmenes de datos al algoritmo para que aprenda y sepa que realizar en diferentes casos o situaciones por el cual es entrenado con un alto grado de precisión (Judith Sandoval, 2018).

La definición dada por (Kaluza, 2016) detalla que el aprendizaje de maquina es un subcampo de la inteligencia artificial. Es aquel que ayuda a los ordenadores a aprender y actuar como seres humanos con la ayuda de algoritmos y datos. A partir de la definición anterior, se puede destacar que el objetivo del aprendizaje automático es el desarrollo de sistemas capaces de permitir a los ordenadores aprender y generalizar una serie de comportamientos. El aprendizaje automático no solo es cada vez más importante en las ciencias computacionales lo es también en nuestro diario vivir. Gracias al ML disfrutamos de software prácticos de reconocimiento de texto y voz, motores de búsqueda fiables, entre otros.

Algunos investigadores clasifican dos tipos de aprendizaje automático, mediante ejemplos conceptuales:

Tabla 4

Tipos de Aprendizaje Automático

1	Aprendizaje Supervisado	• Datos etiquetados • FeedBack directos • Predicción de resultados
2	Aprendizaje No Supervisado	• Sin etiquetas • Sin FeedBack • Encontrar estructuras ocultas en los datos

Nota: Tipos de aprendizaje automático .

El aprendizaje supervisado los datos etiquetados y las variables crean modelos matemáticos con el fin de predecir o clasificar observaciones que se obtendrán después, dichos métodos se estiman que son supervisados en vista que el modelo se fabrica con valores conocidos de las observaciones, dicho de otra manera, la máquina aprende de datos conocidos con el propósito de predecir futuras respuestas. Además, los algoritmos de aprendizaje supervisado se clasifican por medio de la diferenciación con respecto al tipo cuantitativa o cualitativo de la variable de salida implicada en el problema, es decir, el algoritmo de regresión se usa una vez que la respuesta es cuantitativa y el algoritmo de clasificación una vez que la respuesta es cualitativa (Huertas Mora, 2020).

Por otra parte, sobre el aprendizaje no supervisado no se entregan etiquetas al algoritmo, sino solo características, dicho de otro modo, no se encaminan por ideas anteriores de los equipos a los cuales pertenecen las muestras, por lo cual el algoritmo debe aprender como detallar la composición de los datos, dichos métodos integran primordialmente la agrupación (clustering) y los métodos de análisis de componentes principales (PCA). Además, una de las técnicas no supervisadas aplicadas en la ingeniería de mantenimiento es la segmentación, cuya intención es detectar patrones o equipos de un conjunto de datos multidimensionales

conseguidos de sensores que monitorean variables operativas de alta transcendencia. También un algoritmo del aprendizaje no supervisado puede agrupar datos según las similitudes que compartan entre ellos a través de las características que se les encuentren (Huertas Mora, 2020).

Algoritmos de Clasificación

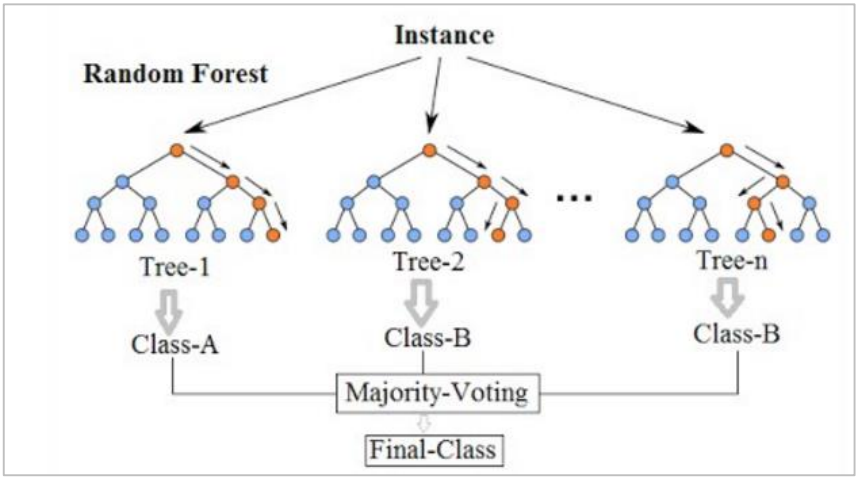
Los algoritmos de clasificación analizan la distancia o características de igual similitud entre un conjunto de datos para categorizarlos, las categorías se crean sin contar con alguna relación jerárquica entre ellas, teniendo cada categoría con rasgos únicos para las futuras calificaciones (Quispe Poccohuanca, 2018).

Algoritmo Clasificador de Bosque Aleatorio (Random Forest RF)

Un bosque aleatorio es un metaestimador que ajusta una serie de clasificadores de árboles de decisión en varias submuestras del conjunto de datos y utiliza el promedio para mejorar la precisión predictiva y controlar el sobreajuste. El tamaño de la submuestra se controla con el max_samples parámetro if bootstrap=True (predeterminado); de lo contrario, se usa todo el conjunto de datos para construir cada árbol.

Figura 6

Algoritmo Clasificador de Bosque Aleatorio



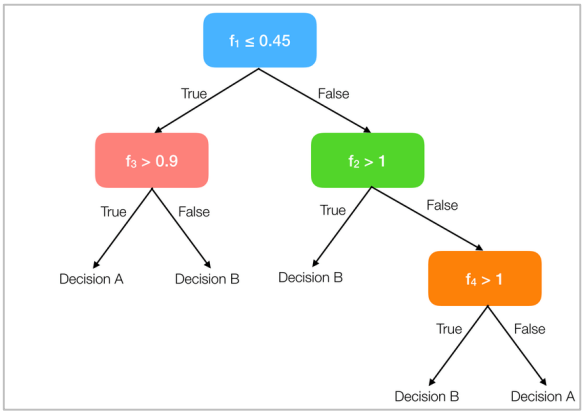
Nota: Clasificador de bosque aleatorio simplificado.

Algoritmo Clasificador de Árboles de Decisión (Decision Tree DT)

Los Árboles de decisión son modelos predictivos que recorriendo y argumentándose logran predecir la probabilidad de un resultado en cuestión, en lo cual se utilizan tanto para conflictos de clasificación como de regresión, es decir son utilizados en el aprendizaje supervisado (Mosquera et al., 2021).

Figura 7

Algoritmo Clasificador de Árboles de Decisión



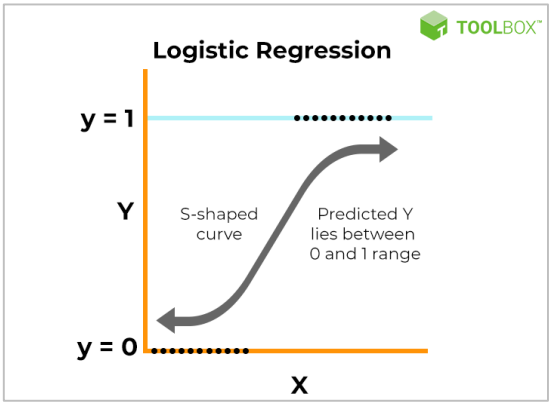
Nota: Clasificador de árbol de decisión simple con 4 características (Mollas et al., 2019).

Algoritmo Clasificador de Regresión Logística (Logistic Regression LR)

La regresión logística, a pesar de su nombre, es un modelo lineal para clasificación en lugar de regresión. La regresión logística también se conoce en la literatura como regresión logit, clasificación de máxima entropía (MaxEnt) o clasificador logarítmico lineal. En este modelo, las probabilidades que describen los posibles resultados de un solo ensayo se modelan mediante una función logística.

Figura 8

Algoritmo Clasificador de Regresión Logística



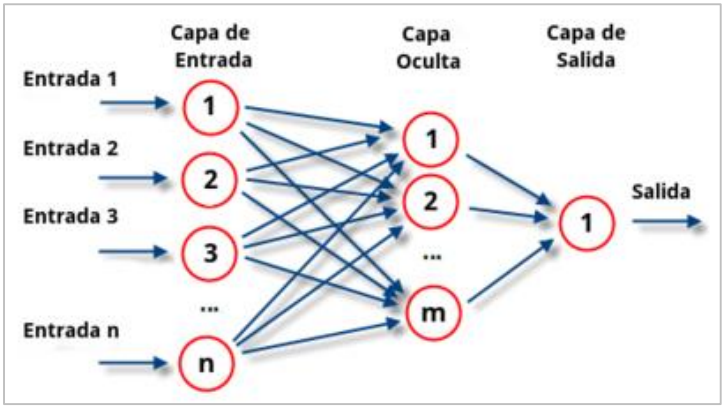
Nota: Regresión Logística – Función Sigmoidea.

Algoritmo Clasificador Perceptron Multicapa (MultiLayer Perceptron MLP)

Esta red es la más usada actualmente debido a que es capaz de actuar como aproximador universal de funciones. Además, junto con el algoritmo de backpropagation es capaz de aprender cualquier tipo de función continua entre un grupo de variables de entrada y de salida (Palmer et al., 2001). Dependiendo de la complejidad de la red, el Perceptrón puede resolver desde funciones continuas, si únicamente presenta una sola capa oculta, hasta funciones no continuas, si presenta más de una capa oculta puede resolver las funciones discontinuas.

Figura 9

Algoritmo Clasificador Perceptron Multicapa



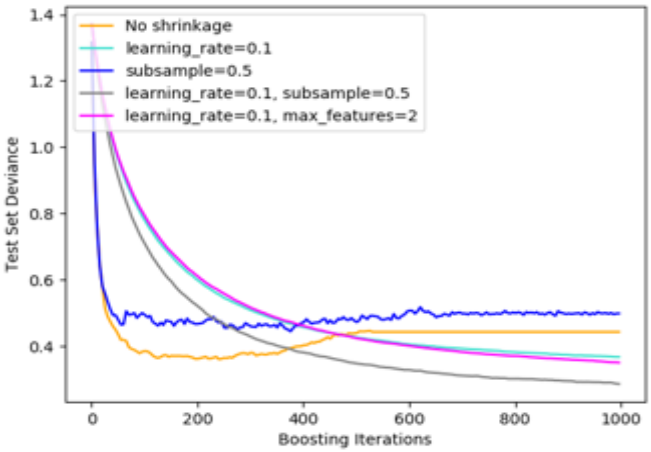
Nota: Estructura del clasificador MLP (Ordoñez Lopez, 2008).

Algoritmo Clasificador de Aumento de Gradiente (Gradient Boosting GB)

GB construye un modelo aditivo en una forma avanzada por etapas; permite la optimización de funciones de pérdida diferenciables arbitrarias. En cada etapa n_classes_, los árboles de regresión se ajustan al gradiente negativo de la función de pérdida, por ejemplo, pérdida logarítmica binaria o multiclase. La clasificación binaria es un caso especial en el que solo se induce un único árbol de regresión.

Figura 10

Algoritmo Clasificador de Aumento de Gradiente



Nota: Clasificador de Aumento de Gradiente. *(scikit-learn documentation, s. f.)*

Pre-procesamiento de Datos

El pre procesamiento de datos implica preparar y "limpiar" datos de texto para que las máquinas puedan analizarlos. El pre procesamiento pone los datos en una forma viable y resalta las características del texto con las que puede trabajar un algoritmo.

Hay varias maneras de hacer esto, incluyendo:

- Tokenización: Cuando el texto se divide en unidades más pequeñas para trabajar.
- Palabra de parada: Cuando las palabras comunes se eliminan del texto para que permanezcan las palabras únicas que ofrecen la mayor cantidad de información sobre el texto.
- Lematización y stemming: Cuando las palabras se reducen a sus formas de raíz para procesar.
- Etiquetado: Cuando las palabras se marcan según la parte del discurso en la que se encuentran, como sustantivos, verbos y adjetivos.

Extracción de Datos

La extracción de datos es un principio muy común en varios aspectos, ya que permite obtener información que posiblemente sea valiosa para diversas circunstancias, concluyendo que la extracción no es más que obtener solo datos que realmente nos interesen de un determinado tema o situación (Cabrera Arévalo & Reyes Sánchez, 2017).

En términos simples, la extracción de datos es el proceso de “extraer y recopilar datos de fuentes semiestructuradas y no estructuradas, como correos electrónicos, documentos PDF, formularios PDF, archivos de texto, redes sociales, códigos de barras e imágenes”. ¿Cómo se realiza la extracción de datos no estructurados? Una herramienta de extracción de datos de

nivel empresarial hace que los datos comerciales entrantes de fuentes no estructuradas o semiestructuradas se puedan utilizar para análisis de datos e informes.

Tweepy

Tweepy es una librería que funciona como un "wrapper" para trabajar con la API REST de Twitter, facilitando por medio de métodos y objetos en Python la interacción con estos servicios.

Python

Python es un lenguaje de programación de alto nivel que se utiliza para desarrollar aplicaciones de todo tipo. A diferencia de otros lenguajes como Java o .NET, se trata de un lenguaje interpretado, es decir, que no es necesario compilarlo para ejecutar las aplicaciones escritas en Python, sino que se ejecutan directamente por el ordenador utilizando un programa denominado interpretador, por lo que no es necesario “traducirlo” a lenguaje máquina (*¿Qué Es Python? | Blog Becas Santander*, n.d.).

Python es un lenguaje sencillo de leer y escribir debido a su alta similitud con el lenguaje humano. Además, se trata de un lenguaje multiplataforma de código abierto y, por lo tanto, gratuito, lo que permite desarrollar software sin límites. Con el paso del tiempo, Python ha ido ganando adeptos gracias a su sencillez y a sus amplias posibilidades, sobre todo en los últimos años, ya que facilita trabajar con inteligencia artificial, big data, machine learning y data science, entre muchos otros campos en auge (*¿Qué Es Python? | Blog Becas Santander*, n.d.).

Google Colab

Colaboratory, o "Colab" para abreviar, es un producto de Google Research. Permite a cualquier usuario escribir y ejecutar código arbitrario de Python en el navegador. Es especialmente adecuado para tareas de aprendizaje automático, análisis de datos y educación.

“Google Colab” es una herramienta para escribir y ejecutar código Python en la nube de Google. También es posible incluir texto enriquecido, “links” e imágenes. En caso de necesitar altas prestaciones de cómputo, el entorno permite configurar algunas propiedades del equipo sobre el que se ejecuta el código. En definitiva, el uso de “Google Colab” permite disponer de un entorno para llevar a cabo tareas que serían difíciles de realizar en un equipo personal. Por otro lado, siguiendo la idea de “Drive”, “Google Colab” brinda la opción de compartir los códigos realizados lo que es ideal para trabajos en equipo (Baume -2021, n.d.)

Scikit-Learn

Scikit-Learn es una de estas librerías gratuitas para Python. Cuenta con algoritmos de clasificación, regresión, clustering y reducción de dimensionalidad. Además, presenta la compatibilidad con otras librerías de Python como NumPy, SciPy y matplotlib.

La gran variedad de algoritmos y utilidades de Scikit-learn la convierten en la herramienta básica para empezar a programar y estructurar los sistemas de análisis datos y modelado estadístico. Los algoritmos de Scikit-Learn se combinan y depuran con otras estructuras de datos y aplicaciones externas como Pandas o PyBrain.

La ventaja de la programación en Python, y Scikit-Learn en concreto, es la variedad de módulos y algoritmos que facilitan el aprendizaje y trabajo del científico de datos en las primeras fases de su desarrollo. La formación de un Máster en Data Science hace hincapié en estas ventajas, pero también prepara a sus alumnos para trabajar en otros lenguajes. La versatilidad y formación es la clave en el campo tecnológico (Universidad de Alcalá, 2022).

Hipótesis

¿Con los métodos de clasificación de Machine Learning es posible ayudar a encontrar las diferentes características estilométricas que permitan determinar el género y profesión de los mensajes de texto de Twitter de usuarios de Ecuador?

Variables de la investigación

Variables Independientes:

- Características estilométricas del texto: Características de tipo fraseológicas y características de uso de palabras frecuentes del idioma español.

Variable Dependientes:

- Determinación de Género y Profesión de textos en español de usuarios de Twitter de Ecuador.

Definiciones conceptuales

- **Estado del Arte:** El estado del arte es una modalidad de la investigación documental que permite el estudio del conocimiento acumulado (escrito en textos) dentro de un área específica.
- **Twitter:** Un término inglés que puede traducirse como "gorjear" o "trinar", es el nombre de una red de microblogging que permite escribir y leer mensajes en Internet que no superen los 140 caracteres. Estas entradas son conocidas como tweets.
- **Api:** Es una "interfaz de programación de aplicaciones". En el contexto de las API, la palabra aplicación se refiere a cualquier software con una función distinta. La interfaz puede considerarse como un contrato de servicio entre dos aplicaciones.
- **Género:** Se refiere a las expectativas, los comportamientos y las actividades de las mujeres y los hombres socialmente construidas, que les son atribuidas sobre la base de su sexo.
- **Profesión:** Actividad habitual de una persona, generalmente para la que se ha preparado, que, al ejercerla, tiene derecho a recibir una remuneración o salario.

- **Librería:** Una librería es uno o varios archivos escritos en un lenguaje de programación determinado, que proporcionan diversas funcionalidades.
- **Wrapper:** En lenguajes de programación como JavaScript, un wrapper o envoltorio es una función que llama a una o varias funciones, unas veces únicamente por convenio y otras para adaptarlas con el objetivo de hacer una tarea ligeramente diferente. Por ejemplo, las librerías SDK de AWS son un ejemplo de wrappers
- **Token:** Un token es la representación digital en el mundo Blockchain de algo que tiene valor dentro de un contexto. Es emitido por una entidad privada y solo es válido bajo este universo concreto. Su funcionamiento es muy similar a un plan de millas dentro de una aerolínea.

CAPÍTULO III

METODOLOGÍA DE LA INVESTIGACIÓN

En este capítulo, se define la metodología y tipo de investigación que será implementada a lo largo del proyecto, determinando así aspectos sobresalientes y válidos que represente un aporte al análisis del problema a investigar.

Tipo de investigación

Partiendo de las variables dependientes o fijas mencionadas anteriormente, el proyecto se basa en un tipo de investigación experimental, que a su vez corresponde a un tipo de investigación cuantitativa pues se considera que, para determinar el comportamiento del objeto de estudio, es necesario aplicar métricas de evaluación. Asimismo, será de tipo bibliográfica para establecer el estado del arte utilizando antecedentes presentados en artículos de relevancia previamente investigados.

A continuación, se describe algunas características que soportan la selección de los siguientes tipos de investigación:

- Experimental: Se analizará la conducta de las características de tipo fraseológicas y características de uso de palabras frecuentes del idioma español aplicadas para este estudio.
- Cuantitativo: Aplicando métricas de evaluación se determinará el porcentaje de precisión y la exhaustividad entre varias soluciones.
- Bibliográfico: Mediante el uso de artículos científicos, libros, etc. es posible establecer el estado del arte y a su vez definir las bases del conocimiento para la presente investigación, como se muestra en la tabla 5.

Estado del Arte

La estilometría comenzó a tomar una forma de método en 1890, las bases fueron establecidas por el filósofo polaco Wincenty Lutoslawski en su famoso libro Principios de Estilometría de 1890, este método utilizó para componer una cronología de los Diálogos de Platón. Que de ahí surgió el famoso problema de la cronología platónica, considerado insoluble por muchos historiadores y tratado de varias formas contradictorias por otros. Comparo muestras de texto de diferente longitud, proporcionando más o menos oportunidades para las características observadas, relaciono todo lo estadístico, el estilo que buscaba marcas exclusivas, limitadas a un pequeño número de diálogos, con los diferentes métodos se ayudó para ver el estilo que tenían la manera de escribir, uso de palabras, número de diálogos, etc. para poder determinar la autoría. (Lutoslawski, 1890) el objetivo era proponer tal teoría y enunciar la hipótesis fundamental de la nueva ciencia de la estilometría, o medida de las afinidades estilísticas.

En su evolución nos encontramos en los principios de los 60, con los ordenadores y la capacidad que tienen para analizar grandes cantidades de información y datos, aunque no garantizaba la calidad del resultado, el reverendo A.Q. Morton realizo un análisis informático sobre las catorce epístolas atribuidas a San Pablo, con el que demostró que pertenecían a seis autores distintos (O'Rourke, 1967) No obstante, con el paso del tiempo y la práctica, los investigadores y estudiosos han pulido sus métodos, que hoy arrojan resultados mucho más acertados. Uno de los primeros éxitos fue la resolución de la controvertida autoría de doce de los *Federalist Papers,* escritos por Frederick Mostellar and David Wallace, esta publicación que tuvo en 1964 con el nombre *Inference and Disputed Authorship* que hizo la portada de la revista *Time,* llamo la atención de académicos y del público por igual por su uso de la metodología estadística para resolver una de las preguntas más notorias de la historia

estadounidense. Este volumen clásico aplica las matemáticas, incluido el controvertido análisis bayesiano, el estudio de palabras de uso frecuente en los textos(Mosteller & Wallace, 1963) Ya en la época actual del campo de la estilometría nos encontramos con un análisis más profundo, donde para confirmar este hallazgo se compararon similitudes léxicas de una tetralogía llamada My Brilliant Friend, estas similitudes eran cercanas entre Domenico Starnone y Elena Ferrante en 2017 el objetivo se logró cuando se aplicó el enfoque del vecino más cercano (k-NN) en todo el vocabulario llegando a la conclusión de que Domenico Starnone es el verdadero autor detrás del seudónimo de Elena Ferrante. (Savoy, 2018). Para los años 2017 y 2021, se lanzó un proyecto ESTO (Estilometría aplicada al Teatro del Siglo de Oro) liderado por Álvaro Cuéllar y Germán Vega García-Luengos (Universidad de Valladolid) donde han conseguido reunir más de 2700 obras del período aurisecular español. Tras la aplicación de análisis estilométricos se está arrojando luz a la autoría de decenas de obras del teatro del Siglo de Oro, gracias a la estilometría pueden averiguar, en una de sus más útiles funcionalidades, qué textos tienen frecuencias en léxico, cuando el autor usa las palabras en unas proporciones distintas, por lo que las obras suelen relacionarse en función de su autoría (Álvaro & Germán Vega, 2022).

En el campo de la estilometría también se aplican técnicas de Machine learning para esto hay que remontarse a sus inicios y principalmente conocer a fondo su raíz. Ya que esta herramienta es una derivación de la inteligencia artificial. En 1943 año en el que el matemático Walter Pitts y el neurofisiólogo Warren McCulloch, quienes dieron a conocer su trabajo enfocado a lo que hoy conocemos como inteligencia artificial, y así fue como en el año de 1950 el científico conocido como Alan Mathison Turing científico, matemático, filósofo y deportista, capaz de crear el conocido “Test de Turing”, cuya finalidad era la de medir que tan inteligente era una computadora. A mediados de 1979 se logró un algoritmo capaz de reconocer patrones, la herramienta principal de la inteligencia artificial que dio origen al machine

learning, ya que al poder brindarle a una máquina la capacidad de aprender patrones se podía adelantar a una respuesta o solución efectiva. (Rámirez, 2018).

La Regresión Logística fue desarrollada por David Cox en 1958, este método permite estimar la probabilidad de una variable, en el que las observaciones se clasifican en un grupo u otro dependiendo del valor que tome la variable empleada como predictor(Cox, 1958). El uso de los Árboles de Decisiones es un método usado para la predicción como técnica de Machine Learning, tuvo su origen en las ciencias sociales con los trabajos de Sonquist y Morgan el año 1964 y Morgan y Messenger el año 1979, ambos realizados en la Universidad de Michigan. De retorno a la comunidad del aprendizaje automático, la construcción de árboles de decisión es sin duda el método de aprendizaje automático más utilizado. (Guillermo Choque Aspiazu, 2009) Otro método conocido es del algoritmo para inducir un Random Forest, fue desarrollado por Leo Breiman y Adele Cutler y Random Forests es su marca de fábrica. El término aparece de la primera propuesta de Random decision forests, hecha por Tin Kam Ho de Bell Labs en 1995. El método combina la idea de bagging de Breiman y la selección aleatoria de atributos, introducida independientemente por Ho, Amit y Geman, para construir una colección de árboles de decisión con variación controlada. (Breiman, 2001) el algoritmo de Multilayer Perceptron o MLP tiene su origen en 1969 por Minsky y Papert, ellos demuestran que el perceptrón simple y adaline no puede resolver problemas no lineales se demuestra que el perceptrón multicapa es un aproximador universal. Un perceptrón multicapa puede aproximar relaciones no lineales entre los datos de entrada y salida. Podemos ver una investigación de esto en (Droua-Hamdani, 2020), uno de los métodos que usamos también es Gradient Boosting fue introducido en 1999 es una media ponderada de los modelos secuenciales calculados en cada iteración, este algoritmo es usado para las muchas predicciones basadas en el método Boosting.

La evolución de las técnicas de Machine Learning dan como resultado poder aplicarse para los diferentes campos que no solo en al área informática, sino también en áreas como la medicina, ingeniería, literatura, biología, etc. Las diferentes investigaciones y artículos que podemos encontrar detallan como estas técnicas dan resultados muy concretos, a lo largo de esta investigación vemos como la predicción de estas técnicas nos ayuda a la implementación de determinación de autoría en campos de texto.

Tabla 5

Artículos referentes a la clasificación de textos

Artículos	Método Usado
(Santosh et al., 2013)	SVM, DT
(Villena-Román & Carlos González-Cristóbal, 2014)	Naive Bayes Multinomial
(Arroju et al., 2015)	SVM, DT
(Oliveira et al., 2017)	Naive Bayes – Bernoulli
(Sezerer et al., 2018)	SVM, CNN
(Nieuwenhuis & Wilkens, 2018)	N-gramas, SVM, RL
(Kosse et al., 2018)	N-gramas, SVM, CNN
(Hacohen-Kerner et al., 2018)	MLP, RL
(Garibo Orts, 2018)	SVM, ANN
(Saman & Diana, 2018)	SVM, N-gramas
(Martín-Del-Campo-Rodríguez et al., 2019)	SVM, N-gramas
(Daelemans et al., 2019)	SVM, N-gramas
(Gągała, 2018)	SVM
(Ruseti & Rebedea, 2012)	CNN
(Moisés et al., 2021)	RF, SVM
(Buda & Bolonyai, 2020)	RL, RF, SVM
(Saeed & Shirazi, 2019)	DT, Multinomial NB
(Schaetti, 2018)	CNN
(Akhtyamova et al., 2017)	RL
(Ulea & Dichiu, 2015)	SVM, CNN

(Alroobaea et al., 2020)	DT

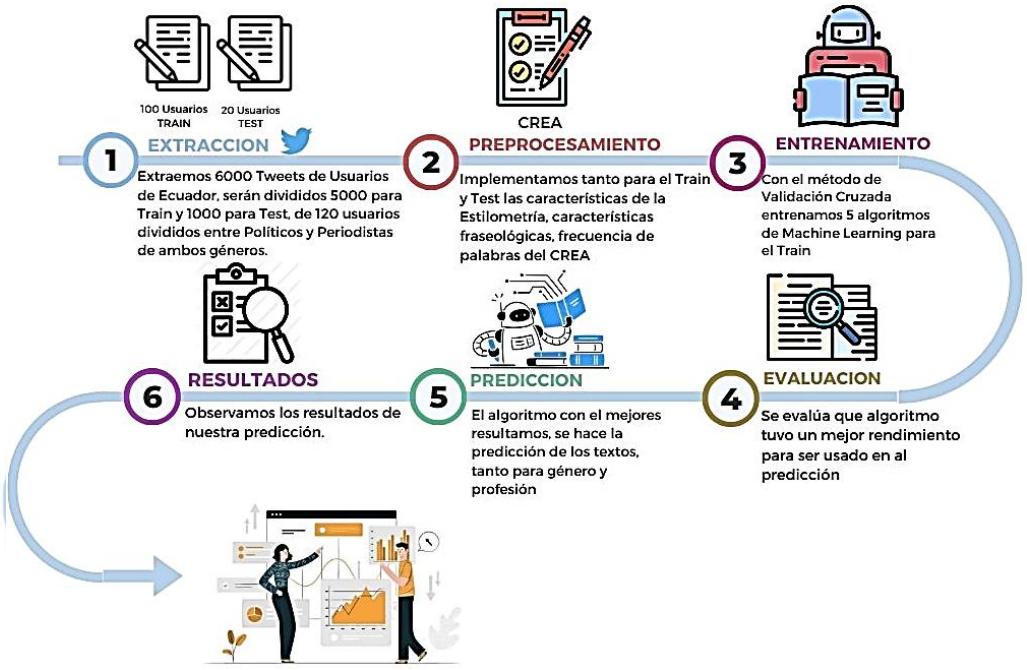
Nota: Listado de artículos fundamentado en Atribución de Autoría y métodos de clasificación de textos. Abreviaturas: Support Vector Machines (SVM), Convolutional Neural Networks (CNN), Random Forest (RF), Multi-Layer Perceptron (MLP), Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), Linear Discriminant Analysis (LDA) y Regression Logistic (RL)

Procesamiento y análisis

Para cumplir con la investigación propuesta en este proyecto, se ingresó Tweets extraídos de los usuarios de Ecuador, se presentó el género y profesión de cada uno de los usuarios, para que el algoritmo logre determinar con el análisis de texto de estos. Tenemos presente el lenguaje de programación Python en el ambiente de prueba Google Colab, se seleccionó las librerías correspondientes para el proceso de datos, posteriormente se comenzó con la extracción de datos, los cuales son procesados y estructurados con las técnicas de estilometría, palabras fraseológicas, tomando la frecuencia, longitud de palabras como nos brinda el CREA, seguido del ingreso a los cinco algoritmos de clasificación escogidos que son: Logistic Regression, Random Forest Classifier, Decision Tree Classifier, MLPClassifier y Gradient Boosting Classifier en el que se realizó el entrenamiento y aprendizaje de los tweets extraídos, con el fin de determinar el género y profesión de los usuarios.

Figura 11

Proceso de desarrollo del proyecto



Nota: Imagen de elaboración propia, correspondiente al procesamiento y análisis del proyecto definido en seis pasos.

Extracción de Datos

Para la extracción de tweets se implementó la librería Tweepy que se usa para acceder a la API de Twitter (*Tweepy*, n.d.). Se buscaron usuarios entre políticos y periodistas del Ecuador que tuvieran una participación en sus carreras respectivas y en la red social, con la herramienta de la librería Panda, estos datos se extraen en un Dataframe y son guardados en un archivo de texto de formato CSV.

Se extrajeron 50 tweets por cada 120 usarios que se dividió en 100 usuarios de entrenamiento y 20 usuarios que servirán para la predicción, como lo muestra la figura 12.

Estos datos nos servirán para el entrenamiento de los algoritmos de Machine Learning antes mencionados, de manera que tengan una mejor predicción al momento de trabajar con los textos y puedan definir con mayor exactitud en se implementó librerías y métodos que

permitan limpiar los textos, eliminando datos innecesarios, como emojis, retweets, links, tweets vacíos, no son de gran importancia al momento de que los algoritmos puedan realizar el trabajo de predicción, pues estos podrían provocar resultados poco óptimos. Como se puede observar en la siguiente Figura 12.

Figura 12
Concatenación de tweets extraídos

	Username Twitter	Nombre GÃ©nero ProfesiÃ³n			\
0	@ottosonnenh	Otto RamÃ³n Sonnenholzner Sper	male	Politico	
1	@ottosonnenh	Otto RamÃ³n Sonnenholzner Sper	male	Politico	
2	@ottosonnenh	Otto RamÃ³n Sonnenholzner Sper	male	Politico	
3	@ottosonnenh	Otto RamÃ³n Sonnenholzner Sper	male	Politico	
4	@ottosonnenh	Otto RamÃ³n Sonnenholzner Sper	male	Politico	
..	...	...	...	...	
115	@GuillermoCeli	Guillermo Alejandro Celi Santos	male	Politico	
116	@GuillermoCeli	Guillermo Alejandro Celi Santos	male	Politico	
117	@GuillermoCeli	Guillermo Alejandro Celi Santos	male	Politico	
118	@GuillermoCeli	Guillermo Alejandro Celi Santos	male	Politico	
119	@GuillermoCeli	Guillermo Alejandro Celi Santos	male	Politico	

	Tweet
0	En Manta, compartÃ con estudiantes y docentes...
1	Muy doloroso lo ocurrido en el Cristo del Cons...
2	@berecordero @ecuadortienevoz Una tragedia que...
3	En Milagro participÃ© de un encuentro con jÃ³v...
4	Desde hace dos aÃ±os he denunciado el uso frau...
..	...
115	En Cotopaxi seguimos SUMAndo con CÃ©sar Umajin...
116	@mariuxi_news Solidario contigo apreciada Mari...
117	Celebrar a #Guayaquil con hechos, llena el cor...
118	Pregunta serÃa: cÃ³mo se llama el platillo ma...
119	Los manabitas se siguen SUMAndo a la UNIDAD, e...

Nota: Resultado de Tweets extraídos y concatenados en una sola celda.

Este archivo CSV se divide en 100 usuarios de entrenamiento y 20 usuarios que nos servirán para la predicción de los datos obtenidos por el entrenamiento esperando obtener los mejores resultados y poder evaluar que algoritmo de Machine Learning, como lo planteamos en nuestro cuarto objetivo.

Extracción de características

Las características de estilométricas, las características de palabras y la frecuencia con la que son utilizadas cuando escribimos un texto, son útiles para poder datos que usan los algoritmos para el entrenamiento y predicción, las características que usamos en esta investigación son: Mean Word Lean, Lexical Diversity, Mean Sentence Len, Stdev Sentencen Len, Mean Paragraph Len, Document Len.

Cada una de estos métodos estilométricos fueron extraídos de (GitHub - Jpotts18/Stylometry: A Stylometry Library for Python, n.d.) este repositorio tiene como fin

aplicar el estudio del estilo lingüístico, por lo general lo que es el lenguaje escrito, y con otros fines de igual manera, estas metodologías nos ayudarán con la determinación de una verdadera autoría. El CREA es importante en esta investigación nos presenta los diferentes datos de las palabras más usadas que nos da la Real Academia de la lengua, esta información la encontremos con mejores detalles en (CREA | Real Academia Española, n.d.).

Entrenamiento

Para el entrenamiento, usamos nuestro dataset de entrenamiento, aplicamos los métodos clasificadores, estos métodos son utilizados de la librería Scikit-Learn (*Scikit-Learn: Machine Learning in Python — Scikit-Learn 1.1.2 Documentation*, n.d.) esta biblioteca contiene muchas herramientas para el aprendizaje automático, ofrece mucha flexibilidad, porque es fácil de comprender debido a su documentación, también usamos el método de Validación Cruzada (*3.1. Cross-Validation: Evaluating Estimator Performance — Scikit-Learn 1.1.2 Documentation*, n.d.), este evaluador que permite el rendimiento que implementamos al momento de usar nuestros algoritmos, como podemos observar en la figura 13.

Figura 13

Validación Cruzada

```
def eval_classifiers(X_train, y_train):
  clfs = [('Logistic regression', LogisticRegression(max_iter=1000, random_state=45)),
          ('Decision tree', DecisionTreeClassifier(random_state=45)),
          ('RandomForest', RandomForestClassifier(n_estimators=20, random_state=45)),
          ('MLP', MLPClassifier(max_iter=1000, random_state=45)),
          ('GBT', GradientBoostingClassifier(n_estimators=100, learning_rate=1.0, max_depth=1, random_state=0))
          ]

  # Vamos devolver los resultados como una tabla
  # Cada fila un algoritmo, cada columna un resultado
  metrics = ['accuracy', 'balanced_accuracy', 'precision', 'recall', 'f1']
  results = pd.DataFrame(columns=metrics)
  for alg, clf in clfs:
    scores = cross_validate(clf, X_train, y_train, cv=10, scoring=metrics) # por defecto, es estratificado
    results.loc[alg,:] = [scores['test_'+m].mean() for m in metrics]
```

Nota: Se observa los métodos usados en los textos del entrenamiento.

Para que los algoritmos de Machine Learning puedan realizar el entrenamiento se dan datos que puedan ser comprendidos y nos puedan dar los resultados que da la Validación Cruzada. Al aplicar este método, cada de unos de los algoritmos anteriormente seleccionados y mencionados realiza el entrenamiento de estos métodos clasificadores lo cual se busca tener una mejor predicción.

Predicción

Para la predicción, se usan los mejores puestos de nuestra evaluación, para género se usa el método de Random forest y en profesión se usa el método de MLP lo cual con el dataset, estos métodos tendran la tarea de evaluar los textos y predecir el género y profesión. Se observa en la figura 14 como estos métodos asignan la predección de las etiquetas género y profesión dándonos la predicción siguiente.

Figura 14

Predicción final del Dataset.

	user	Genero	Profesion
0	@user001	femenino	periodista
1	@user002	masculino	periodista
2	@user003	femenino	politico
3	@user004	femenino	periodista
4	@user005	femenino	politico
5	@user006	femenino	periodista
6	@user007	femenino	politico
7	@user008	masculino	politico
8	@user009	femenino	periodista
9	@user010	femenino	politico
10	@user011	femenino	politico
11	@user012	femenino	politico
12	@user013	femenino	politico
13	@user014	femenino	politico
14	@user015	masculino	politico
15	@user016	femenino	politico
16	@user017	femenino	politico
17	@user018	masculino	politico
18	@user019	masculino	politico
19	@user020	masculino	politico

Nota: Predicción al dataset.

RESULTADOS

Ya realizada la aplicación de las características estilométricas fraseológicas y de palabras de uso frecuente, y el entrenamiento de predicción, a los algoritmos propuestos, se les aplicaron métricas como el Accuracy, Balanced_Accuracy, Presicion, Recall,F1, que representan el porcentaje de casos que el algoritmo acertó en la clasificación y esta matriz determina el número de predicciones correctas o incorrectas. Como se observa en la Figura 15.

Figura 14

Métricas de evaluación

para genero

	accuracy	balanced_accuracy	precision	recall	f1
RandomForest	0.6300	0.5357	0.6625	0.8548	0.7419
MLP	0.5900	0.4917	0.6327	0.8667	0.7243
Logistic regression	0.5700	0.4518	0.6100	0.8786	0.7156
GBT	0.5500	0.4792	0.6251	0.7167	0.6600
Decision tree	0.5100	0.4643	0.6137	0.6119	0.6042

para profesion

	accuracy	balanced_accuracy	precision	recall	f1
MLP	0.8400	0.7369	0.8504	0.9571	0.8969
Logistic regression	0.8300	0.7202	0.8446	0.9571	0.8924
Decision tree	0.8400	0.8009	0.8849	0.9018	0.8876
GBT	0.8200	0.7426	0.8590	0.9018	0.8770
RandomForest	0.7800	0.6643	0.8053	0.9286	0.8595

Nota: Métricas para evaluar cada algoritmo

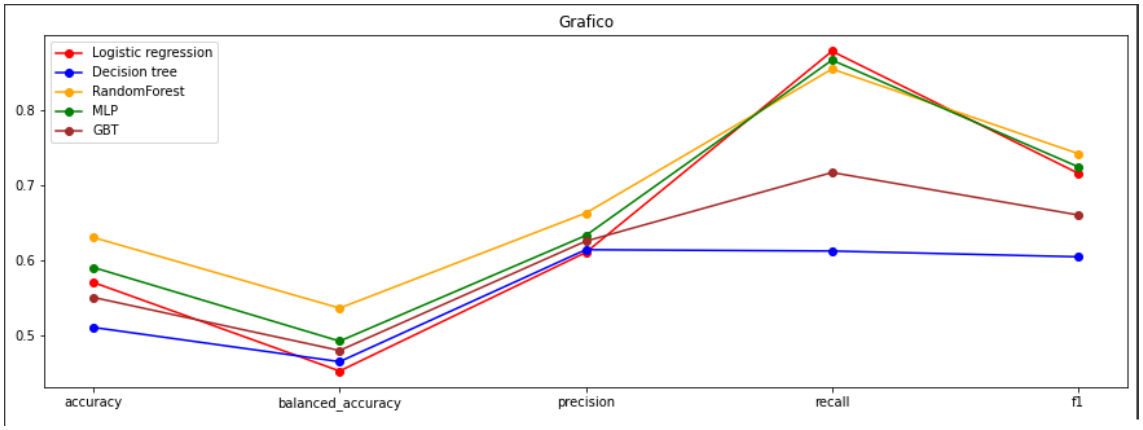
Se observa que el método de clasificación con mejores resultados se dividió tanto para género y profesión, podemos ver como para género los mejores resultados lo obtiene el método Random Forest con el 0.63 de Accuracy superando a los demás clasificadores. Para profesión resalta entre todos el método de MLP con un del 0.84 superando a los demás clasificadores.

Observamos en la figura 16, qué en la métrica Accuracy, para género, se obtiene los resultados que en el método Random Forest con un 0.63, seguido de MLP con un 059, en tercer

lugar a Logistic Regression con un 0.57, en cuarto lugar a Gradient Boosting con un 0.55 y en último lugar a Decisión Tree con un 0.51, dando como mejor resultado a Random Forest.

Figura 15

Resultados obtenidos para la determinación de género.

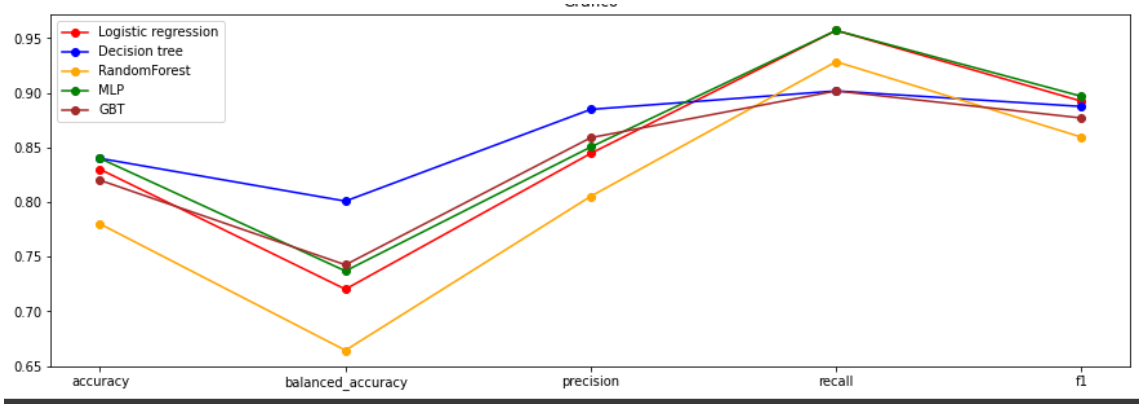


Nota: Gráfica de resultados para género

De igual manera evaluamos los métodos para profesión, en la métrica de Accuracy obteniendo el primero lugar a MLP con un 0.84, seguido de Logistic Regression con un 0.83, en tercer lugar, a Decision Tree con un 0.84, en cuarto lugar, a Gradient Boosting con un 0.82 y por último tenemos a Random Forest con un 0.78. Como lo observamos en la figura 17.

Figura 16

Resultados obtenidos para la determinación de profesión.



Nota: Gráfica de resultados para profesión.

Beneficiarios directos e indirectos del proyecto

- Beneficiarios directos: Futuros proyectos de investigación aplicados a la clasificación de textos cortos.
- Beneficiarios indirectos: Organizaciones con relación al marketing digital, agencias de encuestas, análisis de clientes, etc. en el que mediante la clasificación de tweets sea posible determinar factores que representen un beneficio.

Entregables del proyecto

Los entregables del proyecto son:

- Código fuente en lenguaje Python.
- Documentación de la investigación.
- Artículo Científico.

Propuesta

Determinar género y profesión de textos cortos en español en los usuarios de Twitter Ecuador, usando como base las características estilométricas y frecuencia de palabras (CREA) donde se entrenará con métodos de Machine Learning para su Atribución de Autoría y poder evaluar la efectividad de los métodos usados.

CAPÍTULO IV

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

Culminando este proyecto investigativo, podemos concluir que:

- Se desarrolló y presentó el estado del Arte de la Estilometría y los métodos de clasificación de machine learning, mediante la recopilación de datos históricos y el impacto que tuvieron con el paso del tiempo, a través de investigaciones y trabajos relacionados con el tema propuesto, desde sus inicios y hasta donde han evolucionado con los métodos establecidos.
- Se determinó mediante artículos científicos de relevancia, establecer las técnicas de estilometría y métodos de machine learning más usados en investigaciones similares como son: Logistic Regression, Random Forest Classifier, Decision Tree Classifier, MLP Classifier, Gradient Boosting Classifier para la determinación de género y profesión.
- Se entrenó los métodos de clasificación de machine learning, con un dataset de tweets de usuarios de Ecuador, activos en la red social Twitter, con métodos estilométricos que proporcionaron un mejor preprocesamiento de los textos, que fue de mucha importancia para la implementación que requería la metodología, que permitieron lograr predecir el género y profesión de los usuarios de Twitter.
- Se evaluó los métodos de clasificación usados en este proyecto investigativo, utilizando el método de Validación Cruzada para establecer los resultados de los clasificadores implementados, dando en la métrica Accuracy a Random Forest un total de 0.63 en la predicción de género y, para MLP Classifier un 0.84 en la predicción de profesión.

Recomendaciones

Concluido nuestro proyecto de investigación de determinar el género y profesión de usuarios de Twitter utilizando estilometría con palabras de uso frecuente del español y métodos de clasificación de Machine Learning, se recomienda lo siguiente:

- Obtener más Tweets o datos, para tener mejores resultados o resultados distintos, con el fin de que los métodos clasificatorios obtengan más amplitud de parámetros, para que así logren evaluar y así obtener mayor precisión, con el fin de que el algoritmo obtenga mucha más información al momento de predecir.
- Se recomienda el ambiente de desarrollo de Google Colab, que ayuda a entablar un proyecto con la funcionalidad de no ocupar mucho espacio en una máquina física y poder trabajar de manera multifuncional con diferentes usuarios.
- En el campo de la estilometría aparecen distintos métodos que pueden ayudar con él desarrollo de la investigación que busquen determinar la autoría de textos en español, se recomienda estar al tanto de cómo avanza el desarrollo de estos métodos con base en los algoritmos de machine learning.
- Hay diferentes métodos de machine learning que pueden ser aplicados en esta investigación, como es la Regresión Lineal, esta técnica es utilizada para estudiar la relación entre variables en el ámbito de la investigación para predecir un amplio rango de fenómenos.

Trabajos futuros

Con el fin de poder relacionar este proyecto investigativo con trabajos futuros en la misma línea de investigación, se detalla las diferentes proyecciones que puede tener este proyecto.

- Implementar diferentes modelos de clasificación de textos como son: Naive Bayes, K-Neighbors, entre otros. Diferentes métodos podrían generar mejores resultados en cuanto al entrenamiento y predicción.
- Este proyecto investigativo puede ser aplicado en distintas situaciones que puedan ayudar en varias investigaciones de casos forenses, psiquiátricos, autoría de textos literarios, donde se puedan determinar la autoría de estos textos, para beneficios de la sociedad.
- El proyecto investigativo puede ser de mucha ayuda para determinar las distintas maneras de escribir un texto entre un hombre y una mujer, ya que se podría aplicar de manera de marketing y poder determinar los distintos gustos que tiene las diferentes personas en general.
- Implementar en distintas ramas de la profesión que existe a nivel nacional, que puedan verse más beneficiadas.

BIBLIOGRAFÍA

¿Qué es Python? | Blog Becas Santander. (n.d.). Retrieved September 15, 2022, from https://www.becas-santander.com/es/blog/python-que-es.html

3.1. Cross-validation: evaluating estimator performance — scikit-learn 1.1.2 documentation. (n.d.). Retrieved August 30, 2022, from https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation

Abad-García, M. F. (2019). Plagiarism and predatory journals: A threat to scientific integrity. *Anales de Pediatria*, *90*(1), 57.e1-57.e8. https://doi.org/10.1016/J.ANPEDI.2018.11.003

Akhtyamova, L., Cardiff, J., & Ignatov, A. (2017). *Twitter Author Profiling Using Word Embeddings and Logistic Regression Notebook for PAN*.

Alba, D., & Calle, J. (2020). *Aplicación De Técnicas De Machine Learning Basado En Información Sísmica Para Profundizar La Probabilidad De Terremotos Mediante El Uso De Regresión Logística Y Redes Neuronales*. 173. http://repositorio.ug.edu.ec/handle/redug/11741

Alejandro, D., & Antón, P. (2014). Desarrollo e implementación de un algoritmo basado en text mining aplicado a la atribución de autoría. *UNIVERSIDAD DE VALLADOLID E.T.S.I. TELECOMUNICACIÓN TRABAJO*, 167.

Alroobaea, R., Almulihi, A. H., Alharithi, F. S., Mechti, S., Krichen, M., & Belguith, L. H. (2020). *A Deep learning Model to predict gender, age and occupation of the celebrities based on tweets followers Notebook for PAN at CLEF 2020*. 22–25.

Álvaro, C., & Germán Vega, G. L. (2022). *Estilometría aplicada al Teatro del Siglo de Oro | ETSO*. https://etso.es/

Arroju, M., Hassan, A., & Farnadi, G. (2015). *Age, Gender and Personality Recognition using Tweets in a Multilingual Setting Notebook for PAN*.

Baume -2021, G. L. (n.d.). *Breve introducción a Google Colab*. Retrieved September 15, 2022, from https://colab.research.google.com/

Bourcier, D. (2001). De l'intelligence artificielle à la personne virtuelle : émergence d'une entité juridique ? *Droit et Societe*, *49*(3), 847–871. https://doi.org/10.3917/DRS.049.0847

Breiman, L. (2001). *Random Forests*. *45*, 5–32.

Brito, O. G., Luis, J., Fabela, T., & Salas Hernández, S. (2020). Nuevo enfoque para la extracción de características en la clasificación de textos para la atribución de autoría New Approach for the Extraction of Characteristics in the Classification of Texts for Attribution of Authorship. *Research in Computing Science*, *149*(8), 2020–2817. https://www.proquest.com/openview/79c4d08604b8ce751f6d201abfdc6216/1?pq-origsite=gscholar&cbl=2026366&diss=y

Buda, J., & Bolonyai, F. (2020). *An Ensemble Model Using N-grams and Statistical Features to Identify Fake News Spreaders on Twitter Notebook for PAN at CLEF 2020*. 22–25. https://github.com/pan-webis-de/bolonyai20

Cabrera Arévalo, S. M., & Reyes Sánchez, Z. G. (2017). *Desarrollo de un servicio web para el análisis de tendencias de mercado, a través de la extracción de datos de la red social twitter, mediante la herramienta R, para el apoyo en la toma de decisiones en empresas comerciales*. http://repositorio.ug.edu.ec/handle/redug/19514

Castillo Velásquez, F. A., Godoy, J. L. M., Falcón, M. del C. P. T., Paz, J. P. Z. De,

Chávez, A. B., & Sierra, J. A. R. (2020). Atribución de autoría de mensajes de Twitter a través del análisis sintáctico automático Attribution of Authorship of Twitter Messages through Automatic Syntactic Analysis. *Research in Computing Science*, *149*(11), 2020–2091.

Castillo Velásquez, F. A., Godoy Martínez, J. L., Zavala De Paz, J. P., Amilcar, J., Sierra, R., Consuelo, D., & Falcón, P. T. (2021). *Identificación del género de autores de textos cortos*. https://doi.org/10.13053/CyS-25-3-3999

Castillo Velásquez, F. A., Godoy Martínez, J. L., Zavala de Paz, J. P., Rizzo Sierra, J. A., Torres Falcón, M. del C. P., Castillo Velásquez, F. A., Godoy Martínez, J. L., Zavala de Paz, J. P., Rizzo Sierra, J. A., & Torres Falcón, M. del C. P. (2021). Identificación del género de autores de textos cortos. *Computación y Sistemas*, *25*(3), 659–665. https://doi.org/10.13053/CYS-25-3-3999

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*(2), 215–232. https://doi.org/https://doi.org/10.1111/j.2517-6161.1958.tb00292.x

CREA | Real Academia Española. (n.d.). Retrieved August 30, 2022, from https://www.rae.es/banco-de-datos/crea

Daelemans, W., Kestemont, M., Manjavacas, E., Potthast, M., Rangel, F., Rosso, P., Specht, G., Stamatatos, E., Stein, B., Tschuggnall, M., Wiegmann, M., & Zangerle, E. (2019). *Bots and Gender Profiling, Celebrity Profiling, Cross-Domain Authorship Attribution and Style Change Detection*. https://doi.org/10.1007/978-3-030-28577-7_30

Droua-Hamdani, G. (2020). *ANN-MLP Classifier of Native and Nonnative Speakers Using Speech Rhythm Cues* (pp. 149–156). https://doi.org/10.1007/978-3-030-59430-5_12

Española, R. A. (2022). *fraseología | Definición | Diccionario de la lengua española | RAE - ASALE*. https://dle.rae.es/fraseología

Favián Gutiérrez Constantino. (2020, November 26). *Evaluación bolsas de palabras como característica estilométricas para atribución de autoría.* http://revistatecnologiadigital.com/pdf/10_004_evaluacion_bolsas_palabras_caracteristica_estilometrica_atribucion_autoria.pdf

Gągała, Ł. (2018). *Authorship attribution with neural networks and multiple features Notebook for PAN*.

Garibo Orts, Ò. (2018). *A Big Data approach to gender classification in Twitter. Notebook for PAN*. http://www.internetlivestats.com/twitter-statistics/

Germán Ríos, T., Castro Sánchez, N. A., Sidorov, G., & Posadas Durán, J. P. (2019). Identificación de cambios en el estilo de escritura literaria con aprendizaje automático. *Onomázein*, *46*, 102–128. http://revistadelaconstruccion.uc.cl/index.php/onom/article/view/29699/23175

GitHub - jpotts18/stylometry: A Stylometry Library for Python. (n.d.). Retrieved August 26, 2022, from https://github.com/jpotts18/stylometry

Guillermo Choque Aspiazu. (2009, February 2). *Mente Errabunda: Árboles de decisión*. Www.Eldiario.Net. http://menteerrabunda.blogspot.com/2009/05/arboles-de-decision.html

Hacohen-Kerner, Y., Yigal, Y., Shayovitz, E., Miller, D., & Breckon, T. (2018). *Author Profiling: Gender Prediction from Tweets and Images Notebook for PAN*.

Huertas Mora, A. (2020). *Algoritmos de aprendizaje supervisado utilizando datos de monitoreo de condiciones: Un estudio para el pronóstico de fallas en máquinas*. 1–77.

https://repository.usta.edu.co/bitstream/handle/11634/29886/2020alexanderhuertas.pdf?sequence=1&isAllowed=y
J.C., S. (1992). Future Developments and Research in Phraseology and Terminology related to Translation. *Terminologie et Traduction. Bruselas: Servicio de Traducción de Las Comunidades Europeas*, *2–3*, 584–585.
Jiménez León, J. I., & Martínez Vera, J. J. (2021). ANÁLISIS COMPARATIVO ENTRE MODELOS DE MACHINE LEARNING PARA LA PREDICCIÓN DE FALLO EN ÁREAS AXIALES DE UN RECIPIENTE TOROIDAL DE SECCIÓN RECTA CIRCULAR PARA EL ALMACENAMIENTO DE GNC. In *UNIVERSIDAD DE GUAYAQUIL PROYECTO DE TITULACIÓN*.
Judith Sandoval, L. (2018). ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA ANÁLISIS Y PREDICCIÓN DE DATOS. *Revista Tecnológica* , *11*, 36–40.
Kaluza, B. (2016). *Machine Learning in Java* (1 st ed.). Packt Publishing. Retrieved from. https://www.perlego.com/book/4625/machine-learning-in-java-pdf (Original work published 2016)
Kokkos, A., & Tzouramanis, T. (2014). A robust gender inference model for online social networks ad its application o LinkedIn and Twitter. *First Monday*, *19*(9). https://doi.org/10.5210/FM.V19I9.5216
Kosse, R., Schuur, Y., & Cnossen, G. (2018). *Mixing Traditional Methods with Neural Networks for Gender Prediction Notebook for PAN*. https://pan.webis.de/clef18/pan18-web/author-profiling.html
Lutoslawski, W. (1890). *Principios de estilometría aplicados a la cronología de las obras de Platón*.
Maitra, P., Ghosh, S., & Das, D. (n.d.). *Authorship Verification-An Approach based on Random Forest Notebook for PAN at CLEF 2015*. Retrieved September 8, 2022, from http://weka.wikispaces.com/Use+WEKA+in+your+Java+code
Marcos Ramírez, J., Carillo Ruíz, M., & Josefa Somodevilla, M. (n.d.). *Atribución de autoría combinando información léxico-sintáctica mediante representaciones holográficas reducidas*.
Martín-Del-Campo-Rodríguez, C., Alejandro Pérez Alvarez, D., Efraín, C., Sifuentes, M., Sidorov, G., Batyrshin, I., & Gelbukh, A. (2019). *Authorship Attribution through Punctuation n-grams and Averaged Combination of SVM Notebook for PAN*. https://docs.python.org/3/
Mendenhall, T. C. (1901). Solution of a Literary Problem THE POPULAR SCIENCE. *Wikisource*, *60*(December), 1–21.
Moisés, C., De Andrade, V., & Gonçalves, M. A. (2021). *Profiling Hate Speech Spreaders on Twitter: Exploiting Textual Analysis of Tweets and Combinations of Multiple Textual Representations*. http://ceur-ws.org
Mollas, I., Bassiliades, N., Vlahavas, I., & Tsoumakas, G. (2019). *LionForests: Local Interpretation of Random Forests*. http://arxiv.org/abs/1911.08780
Mosquera, R., Parra, L., Ledesma, A. J., Bonilla, H. F., Mosquera, R., Parra, L., Ledesma, A. J., & Bonilla, H. F. (2021). Applying data mining techniques to predict occupational accidents in the pulp and paper industry. *Información Tecnológica*, *32*(1), 133–142. https://doi.org/10.4067/S0718-07642021000100133
Mosteller, F., & Wallace, D. L. (1963). Inference in an Authorship Problem. *Journal of the American Statistical Association*, *58*(302), 275–309. http://www.jstor.org/stable/2283270
Nieuwenhuis, M., & Wilkens, J. (2018). *Twitter Text and Image Gender*

Classification with a Logistic Regression N-gram Model Notebook for PAN. https://github.com/oarriaga/face_classification

O'Rourke, J. J. (1967). Review of Paul, the Man and the Myth: A Study in the Authorship of Greek Prose, by A. Q. Morton & J. McLeman. *Journal of Biblical Literature*, *86*(1), 110–112. https://doi.org/10.2307/3263256

Oliveira, R. R., Ferreira, R., & Neto, O. (2017). *Using character n-grams and style features for gender and language variety classification Notebook for PAN*.

Ordoñez Lopez, J. P. (2008). *INTRODUCCION A LAS REDES NEURONALES ARTIFICIALES* |. https://jpordonez.wordpress.com/2008/08/06/introduccion-a-las-redes-neuronales-artificiales/

Otávio, J., & Ferreira Frediani, R. (2022). IDENTIFICAÇÃO DE AUTORIA EM TEXTOS CURTOS UTILIZANDO TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL. *UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO,"* 50.

Owen, Y. (2022). *¿Qué es la estilometría?* https://yerseyowen.com/2022/05/17/que-es-la-estilometria/

Palmer, A., Montaño, J. J., & Jiménez, R. (2001). Tutorial sobre Redes Neuronales Artificiales. *Tutorial Sobre Redes Neuronales Artificiales: El Perceptrón Multicapa*, *3*(July). http://www.psiquiatria.com/psiq_general_y_otras_areas/investigacion-86/metodologia/estadis

Pastor, G. C. (1997). Grados de equivalencia translémica de las locuciones en inglés y español. *XVIII Congreso de AEDEAN: Alcalá de Henares. 15-17 Diciembre 1994*, 329–334.

Pastor López-Monroy, A., Montes-y-Gómez, M., Villaseñor-Pineda, L., Ariel Carrasco-Ochoa, J., & Fco Martínez-Trinidad, J. (2012). *LNCS 7329 - A New Document Author Representation for Authorship Attribution*.

Patricia Araujo Arredondo, N. (2009). *Método Semisupervisado para la Clasificación Automática de Textos de Opinión*.

Peñarrubia Navarro, P. (2021). Estilometría con fines geolingüísticos aplicada al corpus COSER. *Revista de Humanidades Digitales*, *6*, 22–42. https://doi.org/10.5944/rhd.vol.6.2021.30870

Pérez, S. A., Profesor Guía, V., Alfaro, R., Profesor Co-Referente, A., Héctor, :, & Cid, A. (2017). *"Análisis y Clasificación de Textos con Técnicas Semi Supervisadas Aplicado a Área Atención al Cliente."*

Quevedo Marcos, B. (2020). *Análisis de las Herramientas de Procesamiento de Lenguaje Natural para estructurar textos médicos*.

Quispe Poccohuanca, O. E. (2018). *INTEGRACIÓN DE TÉCNICAS DE DEEP LEARNING Y ALGORITMOS DE APRENDIZAJE MULTI ETIQUETA PARA LA CLASIFICACIÓN DE TEXTOS*.

Rámirez, D. H. (2018). *EL MACHINE LEARNING A TRAVÉS DE LOS TIEMPOS, Y LOS APORTES A LA HUMANIDAD DENNIYE HINESTROZA RAMÍREZ*. 17.

Raschka, S., & Kaufman, B. (2020). *Machine learning and AI-based approaches for bioactive ligand discovery and GPCR-ligand recognition*. https://doi.org/10.1016/j.ymeth.2020.06.016

Ruiz Gurillo, L. (1997). *Aspectos de fraseología teórica española.* Valencia, Universitat. https://www.academia.edu/4078186/Ruiz_Gurillo_L_1997_Aspectos_de_fraseología_teórica_española_Valencia_Universitat

Ruseti, S., & Rebedea, T. (2012). *Authorship Identification Using a Reduced Set of*

Linguistic Features Notebook for PAN. http://www.rednoise.org/rita/documentation/ripostagger_class_ripostagger.htm

Saeed, U., & Shirazi, F. (2019). *Bots and Gender Classification on Twitter Notebook for PAN at CLEF 2019*. *2019*, 9–12. https://pan.webis.de/clef19/pan19-web/author-profiling.html

Saman, D., & Diana, I. (2018). *Gender Identification in Twitter using N-grams and LSA Notebook for PAN*. https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-

Santosh, K., Bansal, R., Shekhar, M., & Varma, V. (2013). *Author Profiling: Predicting Age and Gender from Blogs Notebook for PAN*. https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

Savoy, J. (2018). Is Starnone really the author behind Ferrante? *Digital Scholarship in the Humanities*, *33*(4), 902–918. https://doi.org/10.1093/llc/fqy016

Schaetti, N. (2018). *Character-based Convolutional Neural Network and ResNet18 for Twitter Author Profiling Notebook for PAN*. https://pan.webis.de/

scikit-learn: machine learning in Python — scikit-learn 1.1.2 documentation. (n.d.). Retrieved August 30, 2022, from https://scikit-learn.org/stable/

Sezerer, E., Polatbilek, O., Sevgili, Ö., & Tekir, S. (2018). *Gender Prediction From Tweets With Convolutional Neural Networks Notebook for PAN*. https://github.com/Darg-Iztech/Gender_Classification

Stamatatos, E. (2009). *Una encuesta sobre los métodos modernos de atribución de autoría*. 1–28.

Staykovski, T. (n.d.). *Stacked Bots and Gender Prediction from Twitter Feeds*. Retrieved September 8, 2022, from https://pan.webis.de

Tweepy. (n.d.). Retrieved August 30, 2022, from https://www.tweepy.org/

Ulea, O.-M. S., & Dichiu, D. (2015). *Automatic Profiling of Twitter Users Based on Their Tweets Notebook for PAN*.

Universidad de Alcalá. (2022). *Scikit-Learn, herramienta básica para el Data Science en Python*. https://www.master-data-scientist.com/scikit-learn-data-science/

Vásquez, A. C., huerta, H. V., Quispe, J. P., & Huayna, A. M. (2009). Procesamiento de lenguaje natural. *Revista de Investigación de Sistemas e Informática*, *6*(2), 45–54. https://revistasinvestigacion.unmsm.edu.pe/index.php/sistem/article/view/5923

Villena-Román, J., & Carlos González-Cristóbal, J. (2014). *Guessing Tweet Author's Gender and Age*. http://www.daedalus.es/

Zipf, G. K. (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press. https://doi.org/doi:10.4159/harvard.9780674434929

ANEXOS

Anexo 1. Planificación de las actividades del proyecto

Nombre de la tarea	Fecha de inicio	Fecha de finalización	13.06.2022	21.06.2022	29.06.2022	06.07.2022	14.07.2022	22.07.2022	30.07.2022	07.08.2022	15.08.2022	23.08.2022	31.08.2022	08.09.2022	16.09.2022
Determinación de género y profesión de usuarios de Twitter utilizando características estilométricas y métodos de clasificación de Machine Learning	13.06.2022	21.06.2022													
Planteamiento del problema	21.06.2022	29.06.2022													
Elaboración marco teórico	29.06.2022	06.07.2022													
Desarrollo estado del arte	06.07.2022	14.07.2022													
Análisis de técnicas de machine learning para la clasificación de textos cortos	14.07.2022	22.07.2022													
Recolección de datos	22.07.2022	30.07.2022													
Evaluación métodos de machine learning	30.07.2022	07.08.2022													
Implementación de métodos	07.08.2022	15.08.2022													
Análisis de resultados para establecer el mejor método ML	15.08.2022	23.08.2022													
Elaboración de articulo científico	23.08.2022	31.08.2022													
Juicios de expertos	31.08.2022	08.09.2022													
Anexos	08.09.2022	16.09.2022													
Entrega del proyecto	16.09.2022	16.09.2022													

Anexo 2. Fundamentación Legal

Las Normas Legales en un Proyecto de Titulación

Apoyo en leyes, estatutos, acuerdos, reglamentos, especialmente para proyectos especiales y factibles, debe escribir únicamente los artículos citados en la CONSTITUCIÓN DE LA REPUBLICA DEL ECUADOR; LEY ORGÁNICA DE EDUCACIÓN SUPERIOR (art. 21), REGLAMENTO DEL CONSEJO DE EDUCACIÓN SUPERIOR; LEY ÓRGANICA DE TRANSPARENCIA Y ACCESO A LA INFORMACIÓN PUBLICA; LEY DEL SISTEMA NACIONAL DE REGISTRO DE DATOS PÚBLICOS; CÓDIGO ORGÁNICO DE LA ECONOMIA SOCIAL DE LOS CONOCIMIENTOS, CREATIVIDAD E INNOVACIÓN, BUEN VIVIR, etc.

- Considerar solo artículos relacionados al tema.
- Debe iniciar la redacción con un breve antecedente de la base legal para realizar el proyecto de titulación.

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

ARTÍCULO DE LA LOES	**CONTEXTO**
¿Qué regula la LOES? 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
¿Cuál es el Objeto de esta Ley? 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, al acceso universal, permanencia, movilidad y egreso sin discriminación alguna.
Entre las funciones ART. 4 DERECHO A LA EDUCACION SUPERIOR	a) Garantizar el derecho a la educación superior mediante la docencia, la investigación y su vinculación con la sociedad, y asegurar crecientes niveles de calidad, excelencia académica y pertinencia; n) Garantizar la producción de pensamiento y conocimiento articulado con el pensamiento universal; y, ñ) Brindar niveles óptimos de calidad en la formación
Principio de Igualdad y Principio de Calidad	El **principio de igualdad** de oportunidades consiste en garantizar a todos los actores del Sistema de Educación Superior las mismas posibilidades en el acceso, permanencia, movilidad y egreso del sistema, sin discriminación de género, credo, orientación sexual, etnia, cultura, preferencia política, condición socioeconómica o discapacidad. El **principio de calidad** consiste en la búsqueda constante y sistemática de la excelencia, la pertinencia, producción óptima, transmisión del conocimiento y desarrollo del pensamiento mediante la autocrítica, la crítica externa y el mejoramiento permanente
ART. 87	Como requisito previo a la obtención del título, los y las estudiantes deberán acreditar servicios a la comunidad mediante prácticas o pasantías pre profesionales. debidamente monitoreadas. en los campos de su especialidad,

	de conformidad con los lineamientos generales definidos por el Consejo de Educación Superior.
ARTÍCULO 19.- DEL REGLAMENTO.- NÓMINA DE GRADUADOS Y NOTIFICACIÓN A LA SENESCYT	Las instituciones de educación superior notificarán obligatoriamente a la SENESCYT la nómina de los graduados y las especificaciones de los títulos que expida, en un plazo no mayor de treinta días contados a partir de la fecha de graduación. (…) este será el único medio oficial a través del cual se verificará el reconocimiento y validez del título en el Ecuador.
ARTÍCULO 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.

Elaboración: Investigadores.
Fuente: Ley Orgánica de Educación Superior.

ARTÍCULO DE LA CONSTITUCIÓN	CONTEXTO
ARTÍCULO 22	Establece: las personas tienen derecho a desarrollar su capacidad creativa, al ejercicio digno y sostenido de las actividades culturales y artísticas, y a beneficiarse de la protección de los derechos morales y patrimoniales que les correspondan por las producciones científicas, literarias o artísticas de su autoría.
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 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
ARTÍCULO 355 primer y segundo inciso	El Estado reconocerá a las universidades y escuelas politécnicas autonomía académica, administrativa, financiera y orgánica, acorde con los objetivos del régimen de desarrollo y los principios establecidos en la Constitución

Elaboración: Investigadores.
Fuente: Ley Orgánica de Educación Superior.

FACTIBILIDAD LEGAL. - Comprende la viabilidad legal del proyecto, es decir, conocer los alcances y limitaciones relacionadas con el desarrollo del mismo.

- La viabilidad legal busca principalmente determinar la existencia de alguna restricción legal en la realización de un proyecto.
- Se busca determinar la existencia de normas o regulaciones legales que impidan la ejecución u operación del proyecto.
- Promover el desarrollo de proyectos sin problemas y dentro de las disposiciones legales.
- Pueden ser registrados y patentados.
- Este proyecto no transgrede ninguna norma, leyes o reglamentos establecidos en la Constitución del Ecuador ni en estamentos legales, por tanto, es factible su desarrollo y aplicación.

CODIGO ORGÁNICO DE LA ECONOMÍA SOCIAL DE LOS CONOCIMIENTOS, CREATIVIDAD E INVENCIÓN

Artículo 104.- Obras susceptibles de protección. - La protección reconocida por el presente Título recae sobre todas las obras literarias, artísticas y científicas, que sean originales y que puedan reproducirse o divulgarse por cualquier forma o medio conocido o por conocerse. 12.- SOFTWARE

Artículo 131.- Protección de software. - El software se protege como obra literaria. Dicha protección se otorga independientemente de que hayan sido incorporados en un ordenador y cualquiera sea la forma en que estén expresados, ya sea como código fuente; es decir, en forma legible por el ser humano; o como código objeto; es decir, en forma legible por máquina, ya sea sistemas operativos o sistemas aplicativos, incluyendo diagramas de flujo, planos, manuales de uso, y en general, aquellos elementos que conformen la estructura, secuencia y organización del programa. Se excluye de esta protección las formas estándar de desarrollo de software. En este sentido, los documentos y textos producidos en las Instituciones de Educación Superior desarrollados con el objeto de obtener sus grados académicos y/o trabajos de facultad, son autores intelectuales con el patrocinio de cada institución, por lo tanto, son acreedores a los derechos de protección intelectual dispuestos en la normativa vigente.

Adicionalmente, considere revisar las siguientes fuentes:

Artículos de la Constitución Política vigente (Año 2018).

Se recomienda lo siguiente:

- ✓ Artículo 22
- ✓ Artículo 26
- ✓ Artículo 28
- ✓ Artículo 350
- ✓ Artículo 355 primer y segundo inciso
- ✓ Artículo 424 primer inciso

Se entiende por inciso a un párrafo.

Artículos de la Ley Orgánica de Educación Superior.

Se recomienda lo siguiente:

- ✓ Artículo 1
- ✓ Artículo 2
- ✓ Artículo 4
- ✓ Artículo 19
- ✓ Artículo 21
- ✓ Principio de Igualdad y Principio de Calidad
- ✓ Artículo 87
- ✓ Artículo 144
- ✓ Artículo 204

En la siguiente lámina se expresan textualmente los principios de Igualdad y de Calidad.

Artículos de la Ley Orgánica de Educación Superior

El ***principio de igualdad*** de oportunidades consiste en garantizar a todos los actores del Sistema de Educación Superior las mismas posibilidades en el acceso, permanencia, movilidad y egreso del sistema, sin discriminación de género, credo, orientación sexual, etnia, cultura, preferencia política, condición socio económica o discapacidad.

El ***principio de calidad*** consiste en la búsqueda constante y sistemática de la excelencia, la pertinencia, producción óptima, transmisión del conocimiento y desarrollo del pensamiento mediante la autocrítica, la crítica externa y el mejoramiento permanente.

ANEXO 3. VALIDACIÓN DE EXPERTOS

DATOS GENERALES

APELLIDOS Y NOMBRES DEL EXPERTO	**TITULO PROFESIONAL DEL EXPERTO**	**AUTORES**	
Santos Díaz Lilia Beatriz	Ingeniera en Sistemas Computacionales	Aucapiña Camas Carlos Ismael	Pazmiño Rosales María Belén
TÍTULO DEL PROYECTO	"Determinación de género y profesión de usuarios de Twitter utilizando estilometría con palabras de uso frecuente del español y métodos de clasificación de Machine Learning"		

INDICADOR	**CRITERIO**	**DEFICIENTE 0-20**				**REGULAR 21-40**				**BUENA 41- 60**				**MUY BUENA 61- 80**				**EXCELENTE 81 - 100**			
		5	**10**	**15**	**20**	**25**	**30**	**35**	**40**	**45**	**50**	**55**	**60**	**65**	**70**	**75**	**80**	**85**	**90**	**95**	**100**
CLARIDAD	¿El proyecto está formulado en un lenguaje de programación apropiado y comprensible?																				X
OBJETIVIDAD	¿El proyecto está expresado en conductas observables y medibles?																				X
ACTUALIDAD	¿Considera que el proyecto está acorde con los métodos tecnológicos actuales y aportan soluciones a temáticas de la actualidad?																				X
SUFICIENCIA	¿Considera que la técnica de validación cruzada y los métodos clasificadores implementados son acordes a los resultados esperados?																				X
INTENCIONALIDAD	¿Los resultados obtenidos reflejan el cumplimiento de los métodos usados en el proyecto?																				X
CONSISTENCIA	¿El desarrollo de la investigación está basado en aspectos teóricos y científicos?																				X
METODOLOGÍA	¿La metodología implementada responde al propósito del proyecto?																				X

APLICABILIDAD	¿Considera que la determinación de género y profesión de los usuarios de Twitter usando algoritmos de Machine Learning demuestra aplicabilidad en este y otros campos a investigar?																				X

CONSTANCIA DE JUICIO DE EXPERTO

Estimado Ingeniero

Ing. César Espín Riofrío, M.sc.

DOCENTE TUTOR DEL TRABAJO DE TITULACIÓN

Guayaquil. -

El presente instrumento certifica que se realizó la revisión del proyecto de titulación "**DETERMINACIÓN DE GÉNERO Y PROFESIÓN DE USUARIOS DE TWITTER UTILIZANDO ESTILOMETRÍA CON PALABRAS DE USO FRECUENTE DEL ESPAÑOL Y MÉTODOS DE CLASIFICACIÓN DE MACHINE LEARNING**" cuyos criterios e indicadores empleados permitieron articular el trabajo según se muestra en el Anexo 3, por tanto, **AUCAPIÑA CAMAS CARLOS ISMAEL** y **PAZMIÑO ROSALES MARÍA BELÉN** 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 3, se procede a validar el trabajo de titulación.

Sin otro particular.

LILIA BEATRIZ SANTOS DIAZ

Firmado digitalmente por:LILIA BEATRIZ SANTOS DIAZ
DN: cn=LILIA BEATRIZ SANTOS DIAZ c=EC o=SECURITY DATA S.A. 2 ou=ENTIDAD DE CERTIFICACION DE INFORMACION
Motivo:Soy el autor de este documento
Ubicación:
Fecha:2022-09-11 18:17-05:00

Ing. Lilia Beatriz Santos Díaz
C.I. N° 0702064452

ANEXO 3. VALIDACIÓN DE EXPERTOS

DATOS GENERALES

APELLIDOS Y NOMBRES DEL EXPERTO	**TITULO PROFESIONAL DEL EXPERTO**	**AUTORES**	
Ing. Jácome Choez William Jessie	Ingeniería en sistemas computacionales	Aucapiña Camas Carlos Ismael	Pazmiño Rosales María Belén
TÍTULO DEL PROYECTO	"Determinación de género y profesión de usuarios de Twitter utilizando estilometría con palabras de uso frecuente del español y métodos de clasificación de Machine Learning"		

INDICADOR	**CRITERIO**	**DEFICIENTE 0-20**				**REGULAR 21-40**				**BUENA 41- 60**				**MUY BUENA 61- 80**				**EXCELENTE 81 - 100**			
		5	**10**	**15**	**20**	**25**	**30**	**35**	**40**	**45**	**50**	**55**	**60**	**65**	**70**	**75**	**80**	**85**	**90**	**95**	**100**
CLARIDAD	¿El proyecto está formulado en un lenguaje de programación apropiado y comprensible?																			X	
OBJETIVIDAD	¿El proyecto está expresado en conductas observables y medibles?																			X	
ACTUALIDAD	¿Considera que el proyecto está acorde con los métodos tecnológicos actuales y aportan soluciones a temáticas de la actualidad?																			X	
SUFICIENCIA	¿Considera que la técnica de validación cruzada y los métodos clasificadores implementados son acordes a los resultados esperados?																X				
INTENCIONALIDAD	¿Los resultados obtenidos reflejan el cumplimiento de los métodos usados en el proyecto?																X				
CONSISTENCIA	¿El desarrollo de la investigación está basado en aspectos teóricos y científicos?																		X		
METODOLOGÍA	¿La metodología implementada responde al propósito del proyecto?																		X		
APLICABILIDAD	¿Considera que la determinación de género y profesión de los usuarios de Twitter usando algoritmos de Machine Learning demuestra aplicabilidad en este y otros campos a investigar?																		X		

CONSTANCIA DE JUICIO DE EXPERTO

Estimado Ingeniero

Ing. Cesar Espín Riofrio, M.sc.

DOCENTE TUTOR(A) DEL TRABAJO DE TITULACIÓN

Guayaquil. -

El presente instrumento certifica que se realizó la revisión del proyecto de titulación "**DETERMINACIÓN DE GÉNERO Y PROFESIÓN DE USUARIOS DE TWITTER UTILIZANDO ESTILOMETRÍA CON PALABRAS DE USO FRECUENTE DEL ESPAÑOL Y MÉTODOS DE CLASIFICACIÓN DE MACHINE LEARNING**"" cuyos criterios e indicadores empleados permitieron articular el trabajo según se muestra en el Anexo 3, por tanto, **AUCAPIÑA CAMAS CARLOS ISMAEL** y **PAZMIÑO ROSALES MARÍA BELÉN** 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 3, se procede a validar el trabajo de titulación.

Sin otro particular.

Ing. William Jessie Jácome Choez
C.I. Nº 0930670112

ANEXO 3. VALIDACIÓN DE EXPERTOS

DATOS GENERALES

APELLIDOS Y NOMBRES DEL EXPERTO	**TITULO PROFESIONAL DEL EXPERTO**	**AUTORES**	
Ing. Alfonso Guijarro Rodríguez, Mgs.	Máster universitario en modelado computacional en ingeniería	Aucapiña Camas Carlos Ismael	Pazmiño Rosales María Belén
TÍTULO DEL PROYECTO	"Determinación de género y profesión de usuarios de Twitter utilizando estilometría con palabras de uso frecuente del español y métodos de clasificación de Machine Learning"		

INDICADOR	**CRITERIO**	**DEFICIENTE 0-20**				**REGULAR 21-40**				**BUENA 41- 60**				**MUY BUENA 61- 80**				**EXCELENTE 81 - 100**			
		5	**10**	**15**	**20**	**25**	**30**	**35**	**40**	**45**	**50**	**55**	**60**	**65**	**70**	**75**	**80**	**85**	**90**	**95**	**100**
CLARIDAD	¿El proyecto está formulado en un lenguaje de programación apropiado y comprensible?																				X
OBJETIVIDAD	¿El proyecto está expresado en conductas observables y medibles?																			X	
ACTUALIDAD	¿Considera que el proyecto está acorde con los métodos tecnológicos actuales y aportan soluciones a temáticas de la actualidad?																				X
SUFICIENCIA	¿Considera que la técnica de validación cruzada y los métodos clasificadores implementados son acordes a los resultados esperados?																			X	
INTENCIONALIDAD	¿Los resultados obtenidos reflejan el cumplimiento de los métodos usados en el proyecto?																				X
CONSISTENCIA	¿El desarrollo de la investigación está basado en aspectos teóricos y científicos?																			X	
METODOLOGÍA	¿La metodología implementada responde al propósito del proyecto?																			X	

APLICABILIDAD	¿Considera que la determinación de género y profesión de los usuarios de Twitter usando algoritmos de Machine Learning demuestra aplicabilidad en este y otros campos a investigar?																			X	

CONSTANCIA DE JUICIO DE EXPERTO

Estimado Ingeniero

Ing. César Espín Riofrío, M.sc.

DOCENTE TUTOR DEL TRABAJO DE TITULACIÓN

Ciudad. -

El presente instrumento certifica que se realizó la revisión del proyecto de titulación "**DETERMINACIÓN DE GÉNERO Y PROFESIÓN DE USUARIOS DE TWITTER UTILIZANDO ESTILOMETRÍA CON PALABRAS DE USO FRECUENTE DEL ESPAÑOL Y MÉTODOS DE CLASIFICACIÓN DE MACHINE LEARNING**" cuyos criterios e indicadores empleados permitieron articular el trabajo según se muestra en el Anexo 3, por tanto, **AUCAPIÑA CAMAS CARLOS ISMAEL** y **PAZMIÑO ROSALES MARÍA BELÉN** 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 3, se procede a validar el trabajo de titulación.

Sin otro particular.

Firmado electrónicamente por:
ALFONSO ANIBAL GUIJARRO RODRIGUEZ

Ing. Alfonso Guijarro Rodríguez, Mgs.
C.I. N° 0914312509

ANEXO 3. VALIDACIÓN DE EXPERTOS

DATOS GENERALES

APELLIDOS Y NOMBRES DEL EXPERTO	**TITULO PROFESIONAL DEL EXPERTO**	**AUTORES**	
		Aucapiña Camas Carlos Ismael	Pazmiño Rosales María Belén
TÍTULO DEL PROYECTO	"Determinación de género y profesión de usuarios de Twitter utilizando estilometría con palabras de uso frecuente del español y métodos de clasificación de Machine Learning"		

INDICADOR	**CRITERIO**	**DEFICIENTE 0-20**				**REGULAR 21-40**				**BUENA 41- 60**				**MUY BUENA 61- 80**				**EXCELENTE 81 - 100**			
		5	**10**	**15**	**20**	**25**	**30**	**35**	**40**	**45**	**50**	**55**	**60**	**65**	**70**	**75**	**80**	**85**	**90**	**95**	**100**
CLARIDAD	¿El proyecto está formulado en un lenguaje de programación apropiado y comprensible?																				X
OBJETIVIDAD	¿El proyecto está expresado en conductas observables y medibles?																			X	
ACTUALIDAD	¿Considera que el proyecto está acorde con los métodos tecnológicos actuales y aportan soluciones a temáticas de la actualidad?																			X	
SUFICIENCIA	¿Considera que la técnica de validación cruzada y los métodos clasificadores implementados son acordes a los resultados esperados?																				X
INTENCIONALIDAD	¿Los resultados obtenidos reflejan el cumplimiento de los métodos usados en el proyecto?																			X	
CONSISTENCIA	¿El desarrollo de la investigación está basado en aspectos teóricos y científicos?																			X	
METODOLOGÍA	¿La metodología implementada responde al propósito del proyecto?																				X

APLICABILIDAD	¿Considera que la determinación de género y profesión de los usuarios de Twitter usando algoritmos de Machine Learning demuestra aplicabilidad en este y otros campos a investigar?																				X

CONSTANCIA DE JUICIO DE EXPERTO

Estimado Ingeniero

Ing. César Espín Riofrío, M.sc.

DOCENTE TUTOR DEL TRABAJO DE TITULACIÓN

Guayaquil. -

El presente instrumento certifica que se realizó la revisión del proyecto de titulación "**DETERMINACIÓN DE GÉNERO Y PROFESIÓN DE USUARIOS DE TWITTER UTILIZANDO ESTILOMETRÍA CON PALABRAS DE USO FRECUENTE DEL ESPAÑOL Y MÉTODOS DE CLASIFICACIÓN DE MACHINE LEARNING**" cuyos criterios e indicadores empleados permitieron articular el trabajo según se muestra en el Anexo 3, por tanto, **AUCAPIÑA CAMAS CARLOS ISMAEL** y **PAZMIÑO ROSALES MARÍA BELÉN** 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 3, se procede a validar el trabajo de titulación.

Sin otro particular.

Firmado electrónicamente por:
VERONICA DEL ROCIO MENDOZA MORAN

Ing. Verónica Mendoza Morán. M.Sc.
C.I. 0703243832

CONSTANCIA DE JUICIO DE EXPERTO

Estimado Ingeniero

Ing. César Espín Riofrío, M.sc.

DOCENTE TUTOR DEL TRABAJO DE TITULACIÓN

Ciudad. -

El presente instrumento certifica que se realizó la revisión del proyecto de titulación **"DETERMINACIÓN DE GÉNERO Y PROFESIÓN DE USUARIOS DE TWITTER UTILIZANDO ESTILOMETRÍA CON PALABRAS DE USO FRECUENTE DEL ESPAÑOL Y MÉTODOS DE CLASIFICACIÓN DE MACHINE LEARNING"** cuyos criterios e indicadores empleados permitieron articular el trabajo según se muestra en el Anexo 3, por tanto, **AUCAPIÑA CAMAS CARLOS ISMAEL** y **PAZMIÑO ROSALES MARÍA BELÉN** 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 3, se procede a validar el trabajo de titulación.

Sin otro particular.

Ing. Diana Cruz Rosales.
C.I. N° 0924558117

ANEXO 3. VALIDACIÓN DE EXPERTOS

DATOS GENERALES

APELLIDOS Y NOMBRES DEL EXPERTO	TITULO PROFESIONAL DEL EXPERTO	AUTORES	
Ing. Diana Cruz Rosales.	Ingeniera en Sistemas Computacionales	Aucapiña Camas Carlos Ismael	Pazmiño Rosales María Belén
TÍTULO DEL PROYECTO	"Determinación de género y profesión de usuarios de Twitter utilizando estilometría con palabras de uso frecuente del español y métodos de clasificación de Machine Learning"		

INDICADOR	CRITERIO	DEFICIENTE 0-20				REGULAR 21-40				BUENA 41- 60				MUY BUENA 61- 80				EXCELENTE 81 - 100			
		5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
CLARIDAD	¿El proyecto está formulado en un lenguaje de programación apropiado y comprensible?															X					
OBJETIVIDAD	¿El proyecto está expresado en conductas observables y medibles?																X				
ACTUALIDAD	¿Considera que el proyecto está acorde con los métodos tecnológicos actuales y aportan soluciones a temáticas de la actualidad?																			X	
SUFICIENCIA	¿Considera que la técnica de validación cruzada y los métodos clasificadores implementados son acordes a los resultados esperados?																	X			
INTENCIONALIDAD	¿Los resultados obtenidos reflejan el cumplimiento de los métodos usados en el proyecto?																			X	
CONSISTENCIA	¿El desarrollo de la investigación está basado en aspectos teóricos y científicos?																			X	
METODOLOGÍA	¿La metodología implementada responde al propósito del proyecto?																			X	
APLICABILIDAD	¿Considera que la determinación de género y profesión de los usuarios de Twitter usando algoritmos de Machine Learning demuestra aplicabilidad en este y otros campos a investigar?																			X	

Anexo 4. Artículo Científico

Determinando género y profesión de usuarios de Twitter utilizando características estilométricas y métodos de clasificación de Machine Learning.

Determining gender and profession of Twitter users using stylometric features and Machine Learning classification methods

César Espin-Riofrio[1[0000-0001-8864-756X]], María Pazmiño-Rosales[1[0000-0002-2227-9570]], Carlos Aucapiña-Camas[1[0000-0002-8163-2055]], Verónica Mendoza-Morán[1[0000-0001-7520-3505]] and Arturo Montejo-Ráez[2[0000-0002-8643-2714]]

[1] Universidad de Guayaquil, Guayaquil 090510, Ecuador
{cesar.espinr, maria.pazminor, carlos.aucapinac, veronica.mendozam}@ug.edu.ec
[2] Universidad de Jaén, Jaén 23071, España
amontejo@ujaen.es

Resumen: El objetivo de este artículo es determinar el género y la profesión de los usuarios de Twitter en Ecuador, mediante el análisis de características estilométricas y técnicas de Machine Learning (ML) para la Atribución de Autoría. El proyecto corresponde a un tipo de investigación cuantitativa-bibliográfica, con diseño experimental realizada para evaluar con diversas características estilométricas y algoritmos de clasificación. Su desarrollo consiste en extraer tweets de usuarios de Ecuador, que serán divididos para entrenamiento y para prueba. Para el pre-procesamiento de la información se implementa características de tipo fraseológicas y de frecuencia de palabras. Posteriormente se entrena los métodos clasificadores Ramdom Forest (RF), Decision Tree (DT), Logistic Regression (LR), MultiLayer Perceptron (MLP) y Gradient Boosting (GB), y se evalúa su comportamiento a través de la técnica de Validación Cruzada y métricas para escoger el clasificador ideal para la predicción del género y profesión. Por último, se presentan los resultados en conductas observables y medibles. Se determinó Random Forest como mejor clasificador para predecir el género y el método MLP para profesión, superando al resto de clasificadores. Esta investigación es de gran interés, debido a que aplica métodos tecnológicos actuales y brinda soluciones óptimas en atribución de autoría utilizando textos cortos.

Palabras clave: Machine Learning, Estilometría, Procesamiento de Lenguaje Natural, Atribución de Autoría.

Abstract. The objective of this article is to determine the gender and profession of Twitter users in Ecuador, through the analysis of stylometric characteristics and Machine Learning (ML) techniques for Authorship Attribution. The project corresponds to a type of quantitative-bibliographic research, with experimental design conducted to evaluate with various stylometric features and classification algorithms. Its development consists of extracting tweets from users in Ecuador, which will be divided for training and for testing. For the pre-processing of the information, phraseological and word frequency features are implemented. Subsequently, the classifier methods Ramdom Forest (RF), Decision Tree (DT), Logistic Regression (LR), MultiLayer Perceptron (MLP) and Gradient Boosting (GB) are trained, and their behavior is evaluated through the Cross Validation technique and metrics to choose the ideal classifier for gender and profession prediction. Finally, the results are presented in observable and measurable behaviors. Random Forest was determined as the best classifier for predicting gender and the MLP method for profession, outperforming the other classifiers. This research is of great interest because it applies current technological methods and provides optimal solutions in authorship attribution using short texts.

Keywords: Machine Learning, Stylometrics, Natural Language Processing, Authorship Attribution..

Introducción

La Atribución de Autoría (AA) se encarga de responder a la cuestión de quién es el autor de un texto, dando algunos ejemplos previos de ese autor. Desde hace mucho tiempo, los trabajos de clasificación han dado buenos resultados para textos largos, sin embargo, el estudio de textos cortos ha quedado aplazado. De modo que en este trabajo de investigación se propone un análisis de las características estilométicas de tipo fraselógicas y de frecuencia de palabras en conjunto con técnicas de ML, para determinar el género y la profesión de 120 usuarios de la red social Twitter en Ecuador. Se dice que el estilo es algo que nace del subconsciente, por tal razón cada quien tiene su propio estilo. Esta es la base de la estilometría, conocido por ser el análisis estadístico de los textos literarios a diferentes niveles.

Se conoce, que las bases de la estilometría fueron establecidas en el siglo XIX por (Lutoslawski, 1890) en su famoso libro "Principios de Estilometría". (Mendenhall, 1901) fue uno de los primeros en identificar a un autor basándose en técnicas estilométricas. Años más tarde, (Zipf, 1932) decide emplear métodos estadísticos basándose en la cantidad de palabras en un texto, lo que da inicio a contemplar diferentes enfoques basados en el aprendizaje automático o como lo define (Raschka & Kaufman, 2020) la ciencia de los algoritmos que dan sentidos a los datos.

Para el entendimiento de los algoritmos y métodos de Machine Learning, se establecerá características de estilométricas, fraseológicas, longitud y frecuencia de palabras. Se trata de un método de clasificación lingüística que goza de mucha popularidad en los últimos años, aunque sus orígenes se remontan a mediados de los sesenta. Se fundamenta en la idea de que el autor de un texto imprime siempre en sus creaciones una huella estilística o autorial, un estilo propio que puede ser rastreado por medio de métodos cuantitativos (Peñarrubia Navarro, 2021).

Su ejecución tiene como objetivo proporcionar la información necesaria para su debido análisis siguiendo la guía de artículos científicos de relevancia. Se utiliza técnicas de Machine Learning que es Validacion Cruzada, estos métodos comprenden Regresión logística, Random Forest Classifier, Decision Tree Classifier, MLPClassifier, Gradient Boosting Classifier, estos diferentes algoritmos y estrategias son para llevar a cabo la Atribución de Autoría, en especial, enfoques de aprendizaje supervisado utilizando un conjunto de dataset para el entrenamiento de los métodos para la predicción de género y profesión.

En los principios de los 60, con los ordenadores y la capacidad que tienen para analizar grandes cantidades de información y datos, aunque no garantizaba la calidad del resultado, el reverendo A.Q. Morton realizo un análisis informático sobre las catorce epístolas atribuidas a San Pablo, con el que demostró que pertenecían a seis autores distintos (O'Rourke, 1967) No obstante, con el paso del tiempo y la práctica, los investigadores y estudiosos han pulido sus métodos, que hoy arrojan resultados mucho más acertados. Uno de los primeros éxitos fue la resolución de la controvertida autoría de doce de los Federalist Papers, escritos por Frederick Mostellar and David Wallace, esta publicación que tuvo en 1964 con el nombre Inference and Disputed Authorship que hizo la portada de la revista Time, llamo la atención de académicos y del público por igual por su uso de la metodología estadística para resolver una de las preguntas más notorias de la historia estadounidense. Este volumen clásico aplica las matemáticas, incluido el controvertido análisis bayesiano, el estudio de palabras de uso frecuente en los textos (Mosteller & Wallace, 1963). Ya en la época actual del campo de la estilometría nos encontramos con un análisis más profundo, donde para confirmar este hallazgo se compararon similitudes léxicas de una tetralogía llamada My Brilliant Friend, estas similitudes eran cercanas entre Domenico Starnone y Elena Ferrante en 2017 el objetivo se logró cuando se aplicó el enfoque del vecino más cercano (k-NN) en todo el vocabulario llegando a la conclusión de que Domenico Starnone es el verdadero autor detrás del seudónimo de Elena Ferrante (Savoy, 2018). Para los años 2017 y 2021, se lanzó un proyecto ESTO (Estilometría aplicada al Teatro del Siglo de Oro) liderado por Álvaro Cuéllar y Germán Vega García-Luengos (Universidad de Valladolid) donde han conseguido reunir más de 2700 obras del período aurisecular español. Tras la aplicación de análisis estilométricos se está arrojando luz a la autoría de decenas de obras del teatro del Siglo de Oro, gracias a la estilometría pueden averiguar, en una de sus más útiles funcionalidades, qué textos tienen frecuencias en léxico, cuando el autor usa las palabras en unas proporciones distintas, por lo que las obras suelen relacionarse en función de su autoría (Álvaro & Germán Vega, 2022).

En el campo de la estilometría también se aplican técnicas de Machine learning para esto hay que remontarse a sus inicios y principalmente conocer a fondo su raíz. Ya que esta herramienta es una derivación de la inteligencia artificial. En 1943 año en el que el matemático Walter Pitts y el neurofisiólogo Warren McCulloch, quienes dieron a conocer su trabajo enfocado a lo que hoy conocemos como inteligencia artificial, y así fue como en el año de 1950 el científico conocido como Alan Mathison Turing científico, matemático, filósofo y deportista, capaz de crear el conocido "Test de Turing", cuya finalidad

era la de medir que tan inteligente era una computadora. A mediados de 1979 se logró un algoritmo capaz de reconocer patrones, la herramienta principal de la inteligencia artificial que dio origen al machine learning, ya que al poder brindarle a una máquina la capacidad de aprender patrones se podía adelantar a una respuesta o solución efectiva. (Rámirez, 2018)

Algunos autores detallan cómo los problemas de aprendizaje automático se pueden dividir en: aprendizaje supervisado y no supervisado. Particularmente el supervisado, utiliza métodos de clasificación cuyo fin es identificar a qué categoría pertenece un objeto. Entre los seleccionados para este estudio consta Logistic Regression (LR), un método desarrollado por (Cox, 1958) que permite estimar la probabilidad de una variable cualitativa binaria en función de una variable cuantitativa. Decision Tree (DT), tuvo origen en los estudios realizados de Sonquist y Morgan en el año 1964, fue uno de los primeros métodos en demostrar la relación que existe entre cada condición y el grupo de acciones permisibles asociado con ella. Multilayer Perceptron (MLP) se origina en 1969 por Minsky y Papert, quienes demostraron que el método es un aproximador universal. Gradient Boosting (GB) fue introducido en 1999, este algoritmo es usado para las muchas predicciones basadas en el método Boosting. Y a inicios del siglo XXI (Breiman, 2001) desarrolló Random Forest (RF), una técnica de aprendizaje supervisado que genera múltiples árboles de decisión sobre un conjunto de datos de entrenamiento donde los resultados obtenidos se combinan a fin de obtener un modelo único más robusto en comparación con los resultados de cada árbol por separado.

Existen pocas referencias para trabajos de AA en Twitter utilizando los clasificadores antes menciados. (Maitra et al., n.d.) trabajaron con Random Forest para clasificar documentos desconocidos basándose en las características extraídas, su enfoque obtuvo una puntuación de 75% de precisión. (Akhtyamova et al., 2017) realizaron pruebas con un clasificador Logistic Regression para identificar el género del autor y la variedad lingüística de los tweets. Los autores experimentaron con un corpus obtenido de Twitter de más de 11 mil autores. El modelo propuesto logró una precisión del 64% para la tarea de identificación del género. (Staykovski, n.d.) logró perfilar el género de un autor con la finalidad de reconocer si el texto hubiera sido escrito por un bot o un humano. Para ello aplicó el método Multilayer Perceptron para la clasificación de tweets.

Hoy, la estilometría nos proporciona instrumentos de análisis que permiten revisar una atribución de autoría con la objetividad que proporciona la cuantificación obtenida por los macro análisis.

La evolución de las técnicas de Machine Learning dan como resultado poder aplicarse para los diferentes campos que no solo en al área informática, si no también en áreas como la medicina, ingeniería, literatura, biología, etc. Las diferentes investigaciones y artículos que podemos encontrar detallan como estas técnicas dan resultados muy concretos, a lo largo de esta investigación vemos como la predicción de estas técnicas nos ayuda a la implementación de determinación de autoría en campos de texto.

Método

El proyecto se basa en un tipo de investigación experimental, que a su vez corresponde a un tipo de investigación cuantitativa pues se considera que, para determinar el comportamiento del objeto de estudio, es necesario aplicar métricas de evaluación.

Inicialmente, se propuso utilizar un lenguaje de programación apropiado y comprensible, en este caso Python en el ambiente de prueba Google Colab, se seleccionó las librerías correspondientes para el proceso de datos, posteriormente se comenzó con la extracción de datos, los cuales son procesados y estructurados con las técnicas de estilometría de tipo fraseológicas, frecuencia de uso de palabras, seguido del ingreso a los algoritmos de clasificación escogidos con los que se realizó el entrenamiento y aprendizaje, con el fin de determinar el género y profesión de los usuarios.

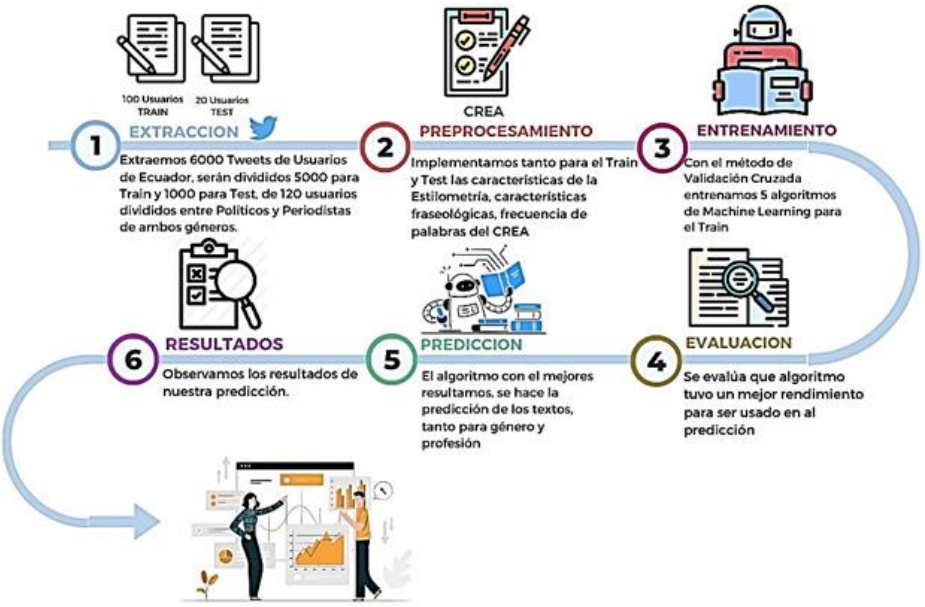


Fig. 1. Proceso de desarrollo del proyecto

Extracción del dataset

Para la extracción de tweets se implementó la librería Tweepy que se usa para acceder a la API de Twitter (*Tweepy*, n.d.). Se buscaron usuarios entre políticos y periodistas del Ecuador que tuvieran una participación en sus carreras respectivas y en la red social, con la herramienta de la librería Panda.

Se extrajeron 50 tweets por cada 120 usuarios que se dividió en 100 usuarios de entrenamiento y 20 usuarios que servirán para la predicción, como lo muestra la figura 2.

```
@fabovillamar https://t.co/PlwQpcLsGN
61) Por lo que se ha visto el último mes, alguien debería sugerirle que puede ser que el responsable no sea el CNE; que investigue en s
62) El COE vive su propia realidad.

Intelectuales https://t.co/gSg9JyeDiR
63) El IESS debe dejar de ser tratado como la caja chica de los gobiernos de turno. Es fundamental que el Ejecutivo deje de tener el cc

Hace un año presenté la reforma, hoy votaremos sobre el texto hasta ahora consensuado. https://t.co/ERMc49HWYf https://t.co/gzgWaLXM87
64) Impresentable. https://t.co/zi1NHlVopr
65) @elizaldehot Que barbaridad.
66) Solicité a @SENAE_Aduana un informe acerca de los reclamos que ha recibido ante el requerimiento del "Registro del Importador" para

@IvanOntanedaB @Vice_Ec https://t.co/ddNU2jfGxV
67) @Stevenneira Que así sea!
Igual para ti Steven.
68) Gracias Dios.

Venga el 2021. https://t.co/SNL1kdytNO
69)

 @LeninArtieda https://t.co/adw7sdsIxi
70) Un solo ídolo!!! https://t.co/rJVxeFeQkr
```

Fig. 2. Ejemplos de Tweets sin procesar

Estos datos servirán para el entrenamiento de los algoritmos de ML, durante la extracción se encuentran caracteres de poca relevancia, por lo que se implementó técnicas que permitieron limpiar los textos, eliminando datos innecesarios como emoticones, retweets, links, tweets vacíos, y así predecir con mayor exactitud, pues estos podrían alterar los resultados. En la figura 3, se observa el nuevo dataset depurado.

	Username Twitter	Nombre GÃ©nero ProfesiÃ³n		
0	@ottosonnenh	Otto RamÃ³n Sonnenholzner Sper	male	Politico
1	@ottosonnenh	Otto RamÃ³n Sonnenholzner Sper	male	Politico
2	@ottosonnenh	Otto RamÃ³n Sonnenholzner Sper	male	Politico
3	@ottosonnenh	Otto RamÃ³n Sonnenholzner Sper	male	Politico
4	@ottosonnenh	Otto RamÃ³n Sonnenholzner Sper	male	Politico
..	...	...	...	...
115	@GuillermoCeli	Guillermo Alejandro Celi Santos	male	Politico
116	@GuillermoCeli	Guillermo Alejandro Celi Santos	male	Politico
117	@GuillermoCeli	Guillermo Alejandro Celi Santos	male	Politico
118	@GuillermoCeli	Guillermo Alejandro Celi Santos	male	Politico
119	@GuillermoCeli	Guillermo Alejandro Celi Santos	male	Politico

	Tweet
0	En Manta, compartÃ con estudiantes y docentes...
1	Muy doloroso lo ocurrido en el Cristo del Cons...
2	@berecordero @ecuadortienevoz Una tragedia que...
3	En Milagro participÃ© de un encuentro con jÃ³v...
4	Desde hace dos aÃ±os he denunciado el uso frau...
..	...
115	En Cotopaxi seguimos SUMAndo con CÃ©sar Umajin...
116	@mariuxi_news Solidario contigo apreciada Mari...
117	Celebrar a #Guayaquil con hechos, llena el cor...
118	Pregunta serÃa: cÃ³mo se llama el platillo ma...
119	Los manabitas se siguen SUMAndo a la UNIDAD, e...

Fig. 3. Concantenación de tweets extraidos

Extracción de características

Las características de tipo fraseológicas forman parte de las características estilométricas, comúnmente son empleadas para recopilar datos que usarán los algoritmos de ML para el entrenamiento y predicción. Algunas de las características seleccionadas en esta investigación son: Mean Word Lean, Lexical Diversity, Mean Sentence Len, Stdev Sentencen Len, Mean Paragraph Len, Document Len. Para ello, se utilizaron las librerías creadas por Jeff Potter (*GitHub - Jpotts18/Stylometry: A Stylometry Library for Python*, n.d.) este repositorio tiene como fin aplicar el estudio del estilo lingüístico, por lo general en el lenguaje escrito, y con otros fines de igual manera.

Por otra parte, para el uso de frecuencia de palabras se utiliza las 1000 primeras del listado CREA (Corpus de Referencia del Español Actual) tomado de la Real Academia Española (RAE) , el cual presenta las palabras más usadas en el idioma español.

Orden	Frec.absoluta		Frec.normalizada
1.	de	9,999,518	65545.55
2.	la	6,277,560	41148.59
3.	que	4,681,839	30688.85
4.	el	4,569,652	29953.48
5.	en	4,234,281	27755.16
6.	y	4,180,279	27401.19
7.	a	3,260,939	21375.03
8.	los	2,618,657	17164.95
9.	se	2,022,514	13257.31
10.	del	1,857,225	12173.87
11.	las	1,686,741	11056.37
12.	un	1,659,827	10879.95
13.	por	1,561,904	10238.07
14.	con	1,481,607	9711.74
15.	no	1,465,503	9606.18
16.	una	1,347,603	8833.36
17.	su	1,103,617	7234.06
18.	para	1,062,152	6962.26
19.	es	1,019,669	6683.79
20.	al	951,054	6234.03
21.	lo	866,955	5682.77
22.	como	773,465	5069.96
23.	más	661,696	4337.33
24.	o	542,284	3554.60
25.	pero	450,512	2953.04

Fig. 4. Archivo CREA

Entrenamiento y Evaluación

Como parte del entrenamiento, se utiliza el dataset de 100 usuarios y se aplica los métodos clasificadores, tomados de la librería Scikit-Learn (*Scikit-Learn: Machine Learning in Python — Scikit-Learn 1.1.2 Documentation*, n.d.) esta biblioteca contiene muchas herramientas para el estudio de aprendizaje automático, es fácil de comprender gracias a la documentación proporcionada.

Luego, se procede a emplear la técnica de Validación Cruzada (*3.1. Cross-Validation: Evaluating Estimator Performance — Scikit-Learn 1.1.2 Documentation*, n.d.), este evaluador permitirá conocer el rendimiento y la confiabilidad de los clasificadores, como se observa en la figura 5.

```
def eval_classifiers(X_train, y_train):
  clfs = [('Logistic regression', LogisticRegression(max_iter=1000, random_state=45)),
          ('Decision tree', DecisionTreeClassifier(random_state=45)),
          ('RandomForest', RandomForestClassifier(n_estimators=20, random_state=45)),
          ('MLP', MLPClassifier(max_iter=1000, random_state=45)),
          ('GBT', GradientBoostingClassifier(n_estimators=100, learning_rate=1.0, max_depth=1, random_state=0))
          ]

  # Vamos devolver los resultados como una tabla
  # Cada fila un algoritmo, cada columna un resultado
  metrics = ['accuracy', 'balanced_accuracy', 'precision', 'recall', 'f1']
  results = pd.DataFrame(columns=metrics)
  for alg, clf in clfs:
    scores = cross_validate(clf, X_train, y_train, cv=10, scoring=metrics) # por defecto, es estratificado
    results.loc[alg,:] = [scores['test_'+m].mean() for m in metrics]
```

Fig. 5. Evaluación de clasificadores con Validación Cruzada

Predicción

Los clasificadores con mejores resultados se utilizan para la predicción. En el caso de género se usa el método de Random Forest y en profesión el método de MLP Classifier, disponiendo para ello de los 20 usuarios restantes destinados para la predicción. Se observa en la figura 6 como estos métodos asignan la predicción de las etiquetas género y profesión.

	user	Genero	Profesion
0	@user001	femenino	periodista
1	@user002	masculino	periodista
2	@user003	femenino	politico
3	@user004	femenino	periodista
4	@user005	femenino	politico
5	@user006	femenino	periodista
6	@user007	femenino	politico
7	@user008	masculino	politico
8	@user009	femenino	periodista
9	@user010	femenino	politico
10	@user011	femenino	politico
11	@user012	femenino	politico
12	@user013	femenino	politico
13	@user014	femenino	politico
14	@user015	masculino	politico
15	@user016	femenino	politico
16	@user017	femenino	politico
17	@user018	masculino	politico
18	@user019	masculino	politico
19	@user020	masculino	politico

Fig. 6. Predicción del Dataset

Resultados

Una vez realizada la aplicación de las características estilométricas, el uso de palabras frecuentes, el entrenamiento y predicción, se evalúa por medio de las métricas Accuracy, Balanced_Accuracy, Presicion, Recall, F1. En la figura 7 se observa la matriz con los valores en el que cada algoritmo acertó en la clasificación.

```
para genero
                    accuracy balanced_accuracy precision recall     f1
RandomForest          0.6300            0.5357    0.6625 0.8548 0.7419
MLP                   0.5900            0.4917    0.6327 0.8667 0.7243
Logistic regression   0.5700            0.4518    0.6100 0.8786 0.7156
GBT                   0.5500            0.4792    0.6251 0.7167 0.6600
Decision tree         0.5100            0.4643    0.6137 0.6119 0.6042
para profesion
                    accuracy balanced_accuracy precision recall     f1
MLP                   0.8400            0.7369    0.8504 0.9571 0.8969
Logistic regression   0.8300            0.7202    0.8446 0.9571 0.8924
Decision tree         0.8400            0.8009    0.8849 0.9018 0.8876
GBT                   0.8200            0.7426    0.8590 0.9018 0.8770
RandomForest          0.7800            0.6643    0.8053 0.9286 0.8595
```

Fig. 7. Resultados obtenidos aplicando métricas de evaluación

Para género, la mejor puntuación lo obtiene el método Random Forest con f1 de 0.7419 superando al resto de clasificadores, para profesión resalta el método MLP Classifier con 0.8969.

Con la técnica de Validación Cruzada, implementada en los cinco métodos de clasificación de ML, se obtiene los siguientes resultados que se visualizan en la figura 8, se escoge la métrica f1 para observar qué, para género, el método Random Forest obtiene un 0.7419, seguido del MLP Classifier con un 0.7243, en tercer lugar a Logistic Regression con un 0.7156, en cuarto lugar a Gradient Boosting con un 0.6600 y en último lugar Decisión Tree con un 0.6042, dando como mejor resultado el entrenamiento con Random Forest.

De igual manera se evaluaron los métodos para profesión, obteniendo el primer lugar a MLP Classifier con un f1 de 0.8969, seguido de Logistic Regression con un 0.8924, en tercer lugar, a Decision Tree con un 0.8876, en cuarto lugar, a Gradient Boosting con un 0.8770 y por último Random Forest con un 0.8595, como se muestra en la figura 9.

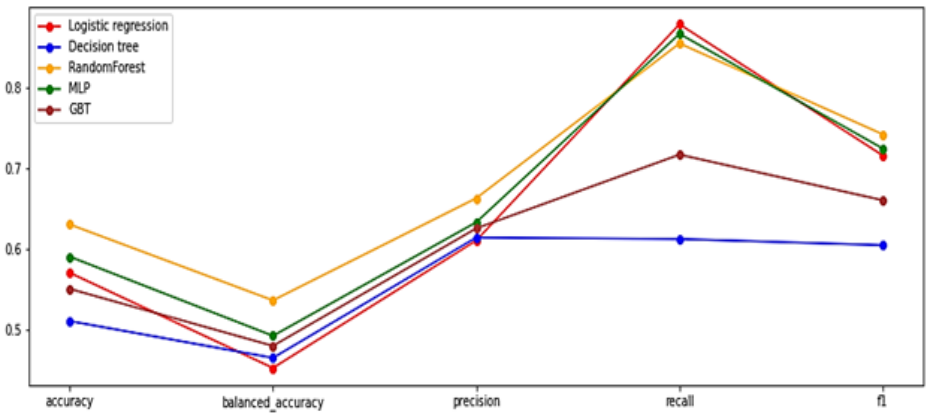


Fig. 8. Resultados obtenidos para la determinación de género

.

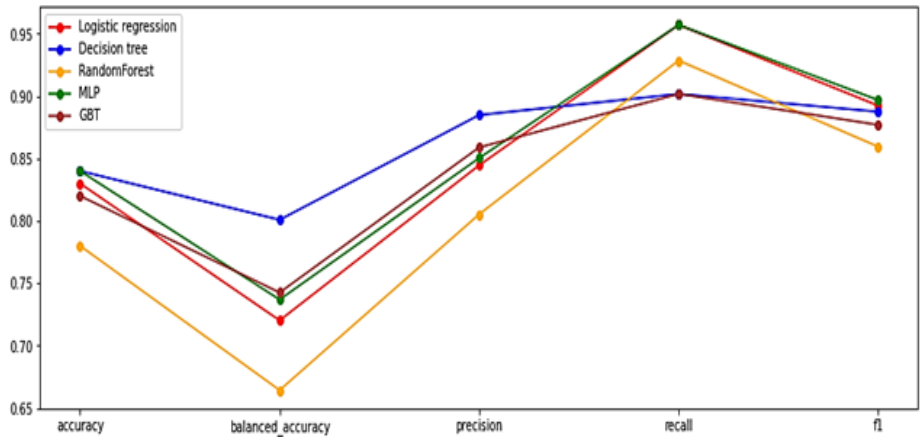


Fig. 9. Resultados obtenidos para la determinación de profesión

Discusión

Los clasificadores seleccionados cumplieron las espectativas del proyecto, aun así existen otros métodos de aprendizaje supervisado que podrían aplicarse al mismo caso de estudio y dar mejores resultados de los obtenidos, por ejemplo los métodos Naive Bayes, Suport Vector Machine (SVM), K-Neighbors, entre otros.

Por otro lado, existen otras técnicas para medir el rendimiento de los modelos de ML, por ejemplo validación simple, bootstrap, matrices de confusión sin embargo la técnica de validación cruzada demostró ser efectiva al medir el rendimiento de todo el conjunto de datos para el entrenamiento y la evaluación, en lugar de una parte.

Las características estilométricas utlizadas de tipo fraseológico y uso frecuente de palabras del idioma español, hicieron se obtengan buenos resultados, sería interesante seguir experimentando con distinto tipo de características estilométricas

Conclusiones

Los resultados de la predicción demuestran la factibilidad de usar modelos de clasificación para la tarea de AA de textos cortos. La contribución más importante es la generalidad de los modelos de ML que pueden ser aplicados a cualquier lenguaje de programación.

Se determinó a Random Forest con un f1 de 0.7419 como mejor clasificador para la etiqueta género y, MLP con f1 de 0.8969 como el mejor para la etiqueta profesión. En ambos casos hubo valores de rendimiento muy cercanos con el uso de los siguientes tres clasificadores.

Las caracterísiterísticas estilométricas de palabras de uso frecuente del idioma español haciendo referencia al listado CREA de la RAE, fueron importantes para un entrenamiento adecuado de los clasificadores, y así obtener una predicción aceptable. Sería bueno experimentar qué tanta influencia tiene la cantidad de palabras de uso frecuente utilizada, podría haber un umbral superior al cual ya no tengan mayor influencia para el entrenamiento y predicción.

Es recomendable implementar más datos en para tener mejores resultados o resultados distintos, para que la precisión tenga un mejor rendimiento.

Se espera trabajar a futuro con la definición de nuevas características estilométicas que mejoren sustancialmente los resultados de clasificación; usar datasets estándares, como los que proporciona PAN-CLEF en sus campañas, para poder comparar nuestro trabajo con otros que ya lo han implementado; y, por último, probar otros métodos de clasificación.

Referencias

¿Qué es Python? | Blog Becas Santander. (n.d.). Retrieved September 15, 2022, from https://www.becas-santander.com/es/blog/python-que-es.html

3.1. Cross-validation: evaluating estimator performance — scikit-learn 1.1.2 documentation. (n.d.). Retrieved August 30, 2022, from https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation

Abad-García, M. F. (2019). Plagiarism and predatory journals: A threat to scientific integrity. *Anales de Pediatria*, *90*(1), 57.e1-57.e8. https://doi.org/10.1016/J.ANPEDI.2018.11.003

Akhtyamova, L., Cardiff, J., & Ignatov, A. (2017). *Twitter Author Profiling Using Word Embeddings and Logistic Regression Notebook for PAN*.

Alba, D., & Calle, J. (2020). *Aplicación De Técnicas De Machine Learning Basado En Información Sísmica Para Profundizar La Probabilidad De Terremotos Mediante El Uso De Regresión Logística Y Redes Neuronales*. 173. http://repositorio.ug.edu.ec/handle/redug/11741

Alejandro, D., & Antón, P. (2014). Desarrollo e implementación de un algoritmo basado en text mining aplicado a la atribución de autoría. *UNIVERSIDAD DE VALLADOLID E.T.S.I. TELECOMUNICACIÓN TRABAJO*, 167.

Alroobaea, R., Almulihi, A. H., Alharithi, F. S., Mechti, S., Krichen, M., & Belguith, L. H. (2020). *A Deep learning Model to predict gender, age and occupation of the celebrities based on tweets followers Notebook for PAN at CLEF 2020*. 22–25.

Álvaro, C., & Germán Vega, G. L. (2022). *Estilometría aplicada al Teatro del Siglo de Oro | ETSO*. https://etso.es/

Arroju, M., Hassan, A., & Farnadi, G. (2015). *Age, Gender and Personality Recognition using Tweets in a Multilingual Setting Notebook for PAN*.

Baume -2021, G. L. (n.d.). *Breve introducción a Google Colab*. Retrieved September 15, 2022, from https://colab.research.google.com/

Bourcier, D. (2001). De l'intelligence artificielle à la personne virtuelle : émergence d'une entité juridique ? *Droit et Societe*, *49*(3), 847–871. https://doi.org/10.3917/DRS.049.0847

Breiman, L. (2001). *Random Forests*. *45*, 5–32.

Brito, O. G., Luis, J., Fabela, T., & Salas Hernández, S. (2020). Nuevo enfoque para la extracción de características en la clasificación de textos para la atribución de autoría New Approach for the Extraction of Characteristics in the Classification of Texts for Attribution of Authorship. *Research in Computing Science*, *149*(8), 2020–2817. https://www.proquest.com/openview/79c4d08604b8ce751f6d201abfdc6216/1?pq-origsite=gscholar&cbl=2026366&diss=y

Buda, J., & Bolonyai, F. (2020). *An Ensemble Model Using N-grams and Statistical Features to Identify Fake News Spreaders on Twitter Notebook for PAN at CLEF 2020*. 22–25. https://github.com/pan-webis-de/bolonyai20

Cabrera Arévalo, S. M., & Reyes Sánchez, Z. G. (2017). *Desarrollo de un servicio web para el análisis de tendencias de mercado, a través de la extracción de datos de la red social twitter, mediante la herramienta R, para el apoyo en la toma de decisiones en empresas comerciales*. http://repositorio.ug.edu.ec/handle/redug/19514

Castillo Velásquez, F. A., Godoy, J. L. M., Falcón, M. del C. P. T., Paz, J. P. Z. De, Chávez, A. B., & Sierra, J. A. R. (2020). Atribución de autoría de mensajes de Twitter a través del análisis sintáctico automático Attribution of Authorship of Twitter Messages through Automatic Syntactic Analysis. *Research in Computing Science*, *149*(11), 2020–2091.

Castillo Velásquez, F. A., Godoy Martínez, J. L., Zavala De Paz, J. P., Amilcar, J., Sierra, R., Consuelo, D., & Falcón, P. T. (2021). *Identificación del género de autores de textos cortos*. https://doi.org/10.13053/CyS-25-3-3999

Castillo Velásquez, F. A., Godoy Martínez, J. L., Zavala de Paz, J. P., Rizzo Sierra, J. A., Torres Falcón, M. del C. P., Castillo Velásquez, F. A., Godoy Martínez, J. L., Zavala de Paz, J. P., Rizzo Sierra, J. A., & Torres Falcón, M. del C. P. (2021). Identificación del género de autores de textos cortos. *Computación y Sistemas*, *25*(3), 659–665. https://doi.org/10.13053/CYS-25-3-3999

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*(2), 215–232. https://doi.org/https://doi.org/10.1111/j.2517-6161.1958.tb00292.x

CREA | Real Academia Española. (n.d.). Retrieved August 30, 2022, from https://www.rae.es/banco-de-datos/crea

Daelemans, W., Kestemont, M., Manjavacas, E., Potthast, M., Rangel, F., Rosso, P., Specht, G., Stamatatos, E., Stein, B., Tschuggnall, M., Wiegmann, M., & Zangerle, E. (2019). *Bots and Gender Profiling, Celebrity Profiling, Cross-Domain Authorship Attribution and Style Change Detection*. https://doi.org/10.1007/978-3-030-28577-7_30

Droua-Hamdani, G. (2020). *ANN-MLP Classifier of Native and Nonnative Speakers Using Speech Rhythm Cues* (pp. 149–156). https://doi.org/10.1007/978-3-030-59430-5_12

Española, R. A. (2022). *fraseología | Definición | Diccionario de la lengua española | RAE - ASALE*.

https://dle.rae.es/fraseología
Favián Gutiérrez Constantino. (2020, November 26). *Evaluación bolsas de palabras como característica estilométricas para atribución de autoría.* http://revistatecnologiadigital.com/pdf/10_004_evaluacion_bolsas_palabras_caracteristica_estilometrica_atribucion_autoria.pdf
Gągała, Ł. (2018). *Authorship attribution with neural networks and multiple features Notebook for PAN*.
Garibo Orts, Ò. (2018). *A Big Data approach to gender classification in Twitter. Notebook for PAN*. http://www.internetlivestats.com/twitter-statistics/
Germán Ríos, T., Castro Sánchez, N. A., Sidorov, G., & Posadas Durán, J. P. (2019). Identificación de cambios en el estilo de escritura literaria con aprendizaje automático. *Onomázein*, *46*, 102–128. http://revistadelaconstruccion.uc.cl/index.php/onom/article/view/29699/23175
GitHub - jpotts18/stylometry: A Stylometry Library for Python. (n.d.). Retrieved August 26, 2022, from https://github.com/jpotts18/stylometry
Guillermo Choque Aspiazu. (2009, February 2). *Mente Errabunda: Árboles de decisión*. Www.Eldiario.Net. http://menteerrabunda.blogspot.com/2009/05/arboles-de-decision.html
Hacohen-Kerner, Y., Yigal, Y., Shayovitz, E., Miller, D., & Breckon, T. (2018). *Author Profiling: Gender Prediction from Tweets and Images Notebook for PAN*.
Huertas Mora, A. (2020). *Algoritmos de aprendizaje supervisado utilizando datos de monitoreo de condiciones: Un estudio para el pronóstico de fallas en máquinas*. 1–77. https://repository.usta.edu.co/bitstream/handle/11634/29886/2020alexanderhuertas.pdf?sequence=1&isAllowed=y
J.C., S. (1992). Future Developments and Research in Phraseology and Terminology related to Translation. *Terminologie et Traduction. Bruselas: Servicio de Traducción de Las Comunidades Europeas*, *2–3*, 584–585.
Jiménez León, J. I., & Martínez Vera, J. J. (2021). ANÁLISIS COMPARATIVO ENTRE MODELOS DE MACHINE LEARNING PARA LA PREDICCIÓN DE FALLO EN ÁREAS AXIALES DE UN RECIPIENTE TOROIDAL DE SECCIÓN RECTA CIRCULAR PARA EL ALMACENAMIENTO DE GNC. In *UNIVERSIDAD DE GUAYAQUIL PROYECTO DE TITULACIÓN*.
Judith Sandoval, L. (2018). ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA ANÁLISIS Y PREDICCIÓN DE DATOS. *Revista Tecnológica* , *11*, 36–40.
Kaluza, B. (2016). *Machine Learning in Java* (1 st ed.). Packt Publishing. Retrieved from. https://www.perlego.com/book/4625/machine-learning-in-java-pdf (Original work published 2016)
Kokkos, A., & Tzouramanis, T. (2014). A robust gender inference model for online social networks ad its application o LinkedIn and Twitter. *First Monday*, *19*(9). https://doi.org/10.5210/FM.V19I9.5216
Kosse, R., Schuur, Y., & Cnossen, G. (2018). *Mixing Traditional Methods with Neural Networks for Gender Prediction Notebook for PAN*. https://pan.webis.de/clef18/pan18-web/author-profiling.html
Lutoslawski, W. (1890). *Principios de estilometría aplicados a la cronología de las obras de Platón*.
Maitra, P., Ghosh, S., & Das, D. (n.d.). *Authorship Verification-An Approach based on Random Forest Notebook for PAN at CLEF 2015*. Retrieved September 8, 2022, from http://weka.wikispaces.com/Use+WEKA+in+your+Java+code
Marcos Ramírez, J., Carillo Ruíz, M., & Josefa Somodevilla, M. (n.d.). *Atribución de autoría combinando información léxico-sintáctica mediante representaciones holográficas reducidas*.
Martín-Del-Campo-Rodríguez, C., Alejandro Pérez Alvarez, D., Efraín, C., Sifuentes, M., Sidorov, G., Batyrshin, I., & Gelbukh, A. (2019). *Authorship Attribution through Punctuation n-grams and Averaged Combination of SVM Notebook for PAN*. https://docs.python.org/3/
Mendenhall, T. C. (1901). Solution of a Literary Problem THE POPULAR SCIENCE. *Wikisource*, *60*(December), 1–21.
Moisés, C., De Andrade, V., & Gonçalves, M. A. (2021). *Profiling Hate Speech Spreaders on Twitter: Exploiting Textual Analysis of Tweets and Combinations of Multiple Textual Representations*. http://ceur-ws.org
Mollas, I., Bassiliades, N., Vlahavas, I., & Tsoumakas, G. (2019). *LionForests: Local Interpretation of Random Forests*. http://arxiv.org/abs/1911.08780
Mosquera, R., Parra, L., Ledesma, A. J., Bonilla, H. F., Mosquera, R., Parra, L., Ledesma, A. J., & Bonilla, H. F. (2021). Applying data mining techniques to predict occupational accidents in the pulp and paper industry. *Información Tecnológica*, *32*(1), 133–142. https://doi.org/10.4067/S0718-07642021000100133
Mosteller, F., & Wallace, D. L. (1963). Inference in an Authorship Problem. *Journal of the American Statistical Association*, *58*(302), 275–309. http://www.jstor.org/stable/2283270
Nieuwenhuis, M., & Wilkens, J. (2018). *Twitter Text and Image Gender Classification with a Logistic Regression N-gram Model Notebook for PAN*. https://github.com/oarriaga/face_classification

O'Rourke, J. J. (1967). Review of Paul, the Man and the Myth: A Study in the Authorship of Greek Prose, by A. Q. Morton & J. McLeman. *Journal of Biblical Literature*, *86*(1), 110–112. https://doi.org/10.2307/3263256

Oliveira, R. R., Ferreira, R., & Neto, O. (2017). *Using character n-grams and style features for gender and language variety classification Notebook for PAN*.

Ordoñez Lopez, J. P. (2008). *INTRODUCCION A LAS REDES NEURONALES ARTIFICIALES |*. https://jpordonez.wordpress.com/2008/08/06/introduccion-a-las-redes-neuronales-artificiales/

Otávio, J., & Ferreira Frediani, R. (2022). IDENTIFICAÇÃO DE AUTORIA EM TEXTOS CURTOS UTILIZANDO TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL. *UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO,"* 50.

Owen, Y. (2022). *¿Qué es la estilometría?* https://yerseyowen.com/2022/05/17/que-es-la-estilometria/

Palmer, A., Montaño, J. J., & Jiménez, R. (2001). Tutorial sobre Redes Neuronales Artificiales. *Tutorial Sobre Redes Neuronales Artificiales: El Perceptrón Multicapa*, *3*(July). http://www.psiquiatria.com/psiq_general_y_otras_areas/investigacion-86/metodologia/estadis

Pastor, G. C. (1997). Grados de equivalencia translémica de las locuciones en inglés y español. *XVIII Congreso de AEDEAN: Alcalá de Henares. 15-17 Diciembre 1994*, 329–334.

Pastor López-Monroy, A., Montes-y-Gómez, M., Villaseñor-Pineda, L., Ariel Carrasco-Ochoa, J., & Fco Martínez-Trinidad, J. (2012). *LNCS 7329 - A New Document Author Representation for Authorship Attribution*.

Patricia Araujo Arredondo, N. (2009). *Método Semisupervisado para la Clasificación Automática de Textos de Opinión*.

Peñarrubia Navarro, P. (2021). Estilometría con fines geolingüísticos aplicada al corpus COSER. *Revista de Humanidades Digitales*, *6*, 22–42. https://doi.org/10.5944/rhd.vol.6.2021.30870

Pérez, S. A., Profesor Guía, V., Alfaro, R., Profesor Co-Referente, A., Héctor, :, & Cid, A. (2017). *"Análisis y Clasificación de Textos con Técnicas Semi Supervisadas Aplicado a Área Atención al Cliente."*

Quevedo Marcos, B. (2020). *Análisis de las Herramientas de Procesamiento de Lenguaje Natural para estructurar textos médicos*.

Quispe Poccohuanca, O. E. (2018). *INTEGRACIÓN DE TÉCNICAS DE DEEP LEARNING Y ALGORITMOS DE APRENDIZAJE MULTI ETIQUETA PARA LA CLASIFICACIÓN DE TEXTOS*.

Rámirez, D. H. (2018). *EL MACHINE LEARNING A TRAVÉS DE LOS TIEMPOS, Y LOS APORTES A LA HUMANIDAD DENNIYE HINESTROZA RAMÍREZ*. 17.

Raschka, S., & Kaufman, B. (2020). *Machine learning and AI-based approaches for bioactive ligand discovery and GPCR-ligand recognition*. https://doi.org/10.1016/j.ymeth.2020.06.016

Ruiz Gurillo, L. (1997). *Aspectos de fraseología teórica española.* Valencia, Universitat. https://www.academia.edu/4078186/Ruiz_Gurillo_L_1997_Aspectos_de_fraseología_teórica_española_Valencia_Universitat

Ruseti, S., & Rebedea, T. (2012). *Authorship Identification Using a Reduced Set of Linguistic Features Notebook for PAN*. http://www.rednoise.org/rita/documentation/ripostagger_class_ripostagger.htm

Saeed, U., & Shirazi, F. (2019). *Bots and Gender Classification on Twitter Notebook for PAN at CLEF 2019*. *2019*, 9–12. https://pan.webis.de/clef19/pan19-web/author-profiling.html

Saman, D., & Diana, I. (2018). *Gender Identification in Twitter using N-grams and LSA Notebook for PAN*. https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-

Santosh, K., Bansal, R., Shekhar, M., & Varma, V. (2013). *Author Profiling: Predicting Age and Gender from Blogs Notebook for PAN*. https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

Savoy, J. (2018). Is Starnone really the author behind Ferrante? *Digital Scholarship in the Humanities*, *33*(4), 902–918. https://doi.org/10.1093/llc/fqy016

Schaetti, N. (2018). *Character-based Convolutional Neural Network and ResNet18 for Twitter Author Profiling Notebook for PAN*. https://pan.webis.de/

scikit-learn: machine learning in Python — scikit-learn 1.1.2 documentation. (n.d.). Retrieved August 30, 2022, from https://scikit-learn.org/stable/

Sezerer, E., Polatbilek, O., Sevgili, Ö., & Tekir, S. (2018). *Gender Prediction From Tweets With Convolutional Neural Networks Notebook for PAN*. https://github.com/Darg-Iztech/Gender_Classification

Stamatatos, E. (2009). *Una encuesta sobre los métodos modernos de atribución de autoría*. 1–28.

Staykovski, T. (n.d.). *Stacked Bots and Gender Prediction from Twitter Feeds*. Retrieved September 8, 2022, from https://pan.webis.de

Tweepy. (n.d.). Retrieved August 30, 2022, from https://www.tweepy.org/

Ulea, O.-M. S., & Dichiu, D. (2015). *Automatic Profiling of Twitter Users Based on Their Tweets Notebook for PAN*.

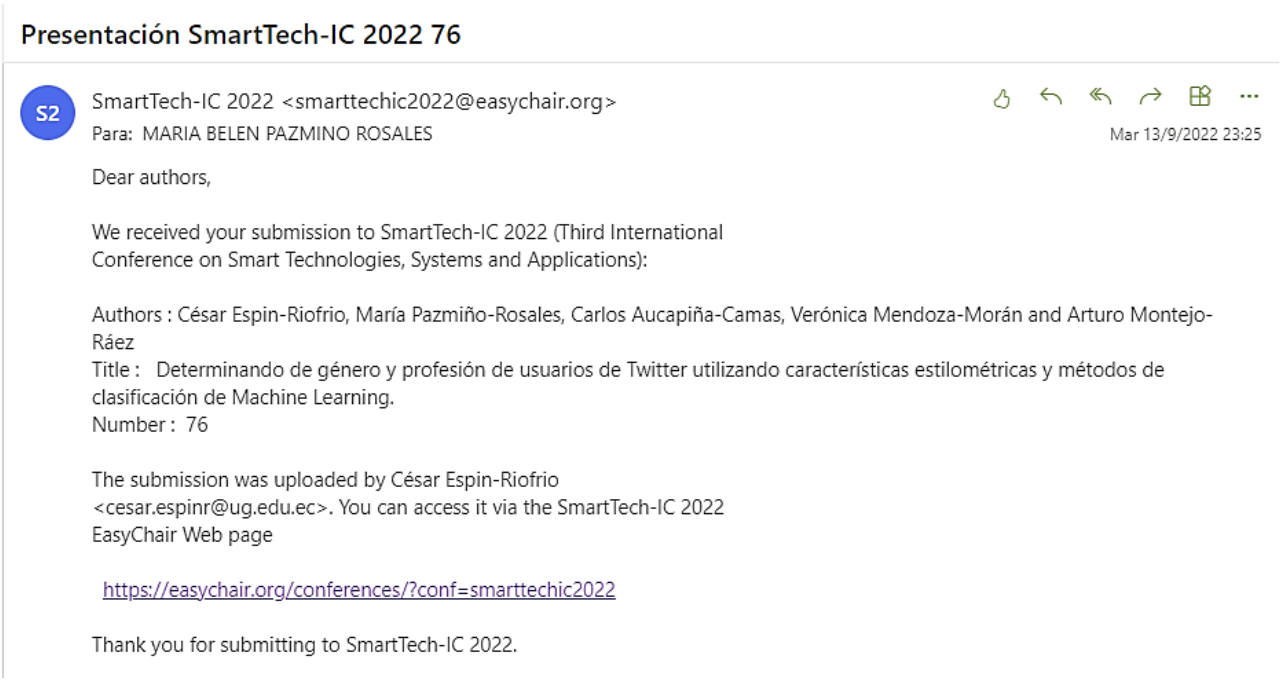
Universidad de Alcalá. (2022). *Scikit-Learn, herramienta básica para el Data Science en Python*. https://www.master-data-scientist.com/scikit-learn-data-science/

Vásquez, A. C., huerta, H. V., Quispe, J. P., & Huayna, A. M. (2009). Procesamiento de lenguaje natural. *Revista de Investigación de Sistemas e Informática*, *6*(2), 45–54. https://revistasinvestigacion.unmsm.edu.pe/index.php/sistem/article/view/5923

Villena-Román, J., & Carlos González-Cristóbal, J. (2014). *Guessing Tweet Author's Gender and Age*. http://www.daedalus.es/

Zipf, G. K. (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press. https://doi.org/doi:10.4159/harvard.9780674434929

Anexo 5. Recepción del Artículo Científico



Presentación SmartTech-IC 2022 76

S2 SmartTech-IC 2022 <smarttechic2022@easychair.org>

Para: MARIA BELEN PAZMINO ROSALES

Mar 13/9/2022 23:25

Dear authors,

We received your submission to SmartTech-IC 2022 (Third International Conference on Smart Technologies, Systems and Applications):

Authors : César Espin-Riofrio, María Pazmiño-Rosales, Carlos Aucapiña-Camas, Verónica Mendoza-Morán and Arturo Montejo-Ráez
Title : Determinando de género y profesión de usuarios de Twitter utilizando características estilométricas y métodos de clasificación de Machine Learning.
Number : 76

The submission was uploaded by César Espin-Riofrio <cesar.espinr@ug.edu.ec>. You can access it via the SmartTech-IC 2022 EasyChair Web page

https://easychair.org/conferences/?conf=smarttechic2022

Thank you for submitting to SmartTech-IC 2022.