INSTITUTO POLITÉCNICO NACIONAL
Escuela Superior de Ingeniería Mecánica y Eléctrica
Unidad Zacatenco

Aplicación para el perfilado de autores de habla hispana

PROYECTO TERMINAL

PRESENTAN
Hernández Santana Diana Estephani
Suárez Flores María Belén

Directores
Dr. Juan Pablo Francisco Posadas Durán
Dra. Karla Sandra Arellano García

Ciudad de México, Junio 2024

Agradecimientos

Quiero expresar mi sincero agradecimiento al Instituto Politécnico Nacional (IPN) y a la Escuela Superior de Ingeniería Mecánica y Eléctrica (ESIME), Unidad Zacatenco, por haberme proporcionado el soporte académico y los recursos necesarios para llevar a cabo este trabajo terminal. Agradezco a todos los profesores por su valiosa orientación, apoyo y enseñanzas que han enriquecido mi formación profesional. Estoy realmente agradecida por haber tenido la oportunidad de estudiar en esta reconocida institución.

Hernández Santana Diana Estephani

Doy gracias a Dios por la salud, porque aunque yo me descuidé, él siempre me sostuvo y me permitió superar todas las enfermedades y desafíos que enfrenté a lo largo de mi carrera. También quiero expresar mi gratitud a mis padres por su esfuerzo, sacrificio, apoyo y confianza. Gracias, papá y mamá, por dejarme salir del castillo donde desde pequeña siempre me hicieron sentir una princesa. Hoy me enorgullece decirles que soy una guerrera que ha salido adelante por sus propios medios. Por último pero no menos importante, agradezco el apoyo de mis amigos y demás personas que conocí en el transcurso de mi vida académica, en especial el de mi amiga Yare, porque llegué a esta ciudad siendo foránea y en ella encontré un hogar. En las buenas, en las malas y en las peores siempre tuve su cariño. Sin duda su amistad incondicional es el tesoro más valioso que la universidad me pudo brindar.

Suárez Flores María Belén

Resumen

En la actualidad, las redes sociales son nuestro principal medio de comunicación. Plataformas sociales como X (anteriormente Twitter) y Facebook, nos permiten expresar opiniones, gustos y pensamientos, por otro lado también permite mantenernos informados sobre noticias y tendencias, sin embargo, la poca censura de dichas plataformas y la gran cantidad de información generada día con día nos mantiene expuestos a temas sensibles y noticias falsas, por lo tanto, el contenido presentado a cada usuario no siempre es el adecuado.

En el presente proyecto se propone el diseño de un sistema de perfilado de autor como una herramienta que permita clasificar usuarios en base a su género y variedad linguística, de esta forma políticos, empresas, creadores de contenido, organizaciones, entre otros, serán capaces de ajustar su contenido a cada usuario para tener mayor alcance y respuesta positiva de parte de su público.

Se entrenó al sistema con una base de datos previamente etiquetados, técnicas de aprendizaje automático y procesamiento de lenguaje natural. Posteriormente se realizaron experimentos con datos de prueba y con tweets extraídos directamente de perfiles de famosos hispanohablantes.

El sistema propuesto obtuvo una precisión del 77.1 % para el género femenino y del 76.7 % para el género masculino implementando el algoritmo de Máquinas de Soporte Vectorial. Por otro lado la mayor precisión para la predicción de la varaiable linguística, fue del 91 %, implementando el algoritmo Multinomial Naive Bayes.

Contenido

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En este capítulo se abordan las generalidades de un proyecto enfocado a la creación de un algoritmo para el perfilado de autor. Se presenta el planteamiento del problema y su justificación, al igual que el objetivo general y los objetivos particulares que se desean alcanzar.

1.1. Planteamiento del problema

La aparición de Internet proporcionó los medios necesarios para el desarrollo de plataformas sociales como Facebook, X (anteriormente Twitter), Instagram, WhatsApp, entre otras, que permiten la interacción y el intercambio de mensajes de texto y contenido multimedia entre millones de usuarios de todo el mundo. Estas redes sociales se han convertido en sitios donde los usuarios expresan sus inquietudes y opiniones sobre temas políticos, económicos y sociales, además de ser utilizados para mantenerse informados sobre noticias y tendencias.

El crecimiento exponencial de usuarios de la plataforma X y la cantidad de información generada mundialmente día con día (456,000 tweets por minuto, 656 millones de tweets al día), han provocado que múltiples empresas y organizaciones utilicen esta red social como su principal canal de difusión para anunciar nuevos productos y servicios. También para mantenerse al tanto de las opiniones y necesidades de sus clientes, además de buscar nuevas posibilidades de mercado. Sin embargo, la propagación instantánea de información a través de

Internet ha creado un entorno digital donde surgen diversas problemáticas: 1) identificación de información relevante y confiable, que puede llevar a la propagación de desinformación y noticias falsas; 2) la recopilación masiva de datos personales expone a las personas a robos de identidad, fraudes en línea y otros delitos cibernéticos; 3) la sobrecarga de información y la exposición constante a noticias negativas pueden tener un impacto negativo en la salud mental de las personas, aumentando los niveles de estrés, ansiedad y fatiga informativa.

Esta plataforma mantiene una relación estrecha con el usuario, ya que no notifica quién ve tu perfil o tus publicaciones y gracias a ello, las personas se sienten libres de expresarse. En comparación con otras plataformas, no presenta tantas restricciones respecto al tipo de contenido que puede publicarse, por lo tanto, es común encontrar discursos de odio, amarillismo, noticias falsas, contenido sexual explícito y palabras altisonantes. Debido a la poca censura de la plataforma, el tipo de contenido mostrado no suele ser el adecuado para cada usuario y debido a ello, temas sensibles son visualizados diariamente.

Mediante la implementación de un sistema de perfilado de autor, es posible deducir características demográficas de los autores de textos, tales como edad, género, ocupación, nacionalidad e incluso rasgos de personalidad. Este enfoque permite clasificar a los usuarios con el propósito de mostrar temas de interés y preferencias acorde a su perfil.

Existen diversos desafíos en la creación de perfiles de autor en textos extraídos de redes sociales en idioma español. Algunos de ellos son, las variantes lingüísticas del español hablado en diferentes países, la interpretación del significado de las palabras en diversos contextos, el uso de emojis para representar estados de ánimo, objetos e ideas, las faltas de ortografía, la gramática y el lenguaje informal de los usuarios.

El presente proyecto de titulación propone un sistema para inferir de manera automática la variedad ligüística y género de los autores de textos digitales en un corpus extraído de la plataforma X en idioma español, aplicando técnicas de Procesamiento de Lenguaje Natural y aprendizaje automático.

1.2. Objetivos

1.2.1. Objetivo general

Desarrollar un sistema para el perfilado de autor, utilizando aprendizaje supervisado y procesamiento de lenguaje natural.

1.2.2. Objetivos particulares

- ✓ Identificar las características de un sistema para el perfilado de autor.
- ✓ Identificar las herramientas necesarias para el desarrollo del prototipo.
- ✓ Diseño de un sistema para perfilado de autor orientado a la predicción de género y variedad linguística
- ✓ Programar el sistema utilizando las herramientas seleccionadas.
- ✓ Evaluar el sistema propuesto.

1.3. Justificación

La inteligencia artificial (IA) se ha convertido en una herramienta versátil y poderosa con una amplia gama de aplicaciones. Su función principal radica en la capacidad de las máquinas para aprender, razonar y tomar decisiones de manera similar a los humanos. Esta capacidad permite a la IA analizar grandes cantidades de datos de manera eficiente y encontrar patrones significativos que pueden impulsar la toma de decisiones informadas en diversos campos, desde la atención médica hasta la logística empresarial. La importancia de la IA se refleja en su capacidad para mejorar la eficiencia, la productividad y la precisión en una variedad de tareas, así como en su potencial para impulsar la innovación y resolver problemas complejos a escala global. Sus aplicaciones son diversas y abarcan desde sistemas de recomendación en plataformas digitales hasta diagnósticos médicos asistidos por IA. En el campo de la psicología, la inteligencia artificial no tiene como objetivo primordial reemplazar las labores humanas, sino más bien posicionarse como una herramienta destinada a optimizar la recolección y el análisis de los datos [1]. En el área del marketing, la IA nos permite

anticipar las necesidades de los usuarios y ofrecer soluciones a las mismas, inclusive antes de que sean buscadas mediante el análisis predictivo; empleando el uso de chatbots y técnicas de Deep learning, se pueda obtener datos de varias fuentes de información para crear contenido de publicidad lo más personalizada posible [2].

Es común que muchas personas proporcionen información falsa sobre su nombre, edad, sexo y ubicación para ocultar su identidad real. Esto presenta un desafío importante para las autoridades encargadas de hacer cumplir la ley y los moderadores de las redes sociales, ya que atrapar a los depredadores en línea implica investigar una gran cantidad de perfiles y conversaciones, además de lidiar con el hecho de que estos depredadores suelen presentarse con una identidad falsa y actúan como jóvenes para ganarse la confianza de sus víctimas. Para identificar a estos impostores, resulta útil analizar su estilo de escritura utilizando el perfilado de autor. Por tanto, se vuelve esencial la implementación de sistemas automatizados eficaces para descubrir e inspeccionar identidades en diferentes tipos de texto en diversas plataformas. [3].

La tarea del perfilado de autor se basa en el análisis del contenido generado o compartido por los usuarios con el objetivo de determinar sus atributos demográficos como edad, género y rasgos de personalidad empleando el aprendizaje automático y procesamiento de lenguaje natural. Esta es una técnica crucial en diversas áreas, como la investigación académica, la seguridad cibernética, el análisis de redes sociales y la inteligencia de negocios. Con su implementación, distintas empresas y organizaciones pueden ajustar el contenido y las herramientas que proveen a cada usuario con el objetivo de mostrar temas de su interés en relación con programas sociales, mercadotecnia, entretenimiento, información educativa, promoción política, entre otros [4].

En la actualidad, el acceso a una gran cantidad de información a través de plataformas sociales ha transformado significativamente la manera en que comprendemos y analizamos los comportamientos sociales. Automatizar procesos mediante el uso de Inteligencia artificial permite la optimización de recursos y disminución de costos; pero sobre todo tiene un impacto transcendental para los usuarios. La naturaleza dinámica de dichas plataformas ofrece nuevas oportunidades para la obtención de información de manera más rápida y eficaz en comparación con los métodos tradicionales. Ante el volumen de datos generado por las plataformas sociales, la tradicional metodología de recolección de información mediante

encuestas individuales se muestra insuficiente e ineficiente. Este enfoque clásico no solo demanda recursos considerables en términos de tiempo y mano de obra, sino que también enfrenta desafíos significativos en cuanto a la precisión y la manipulación de los datos recopilados [3].

En contraste, el análisis del comportamiento de los usuarios en su entorno natural dentro de las plataformas sociales promete proporcionar resultados más efectivos y menos susceptibles a manipulaciones indebidas. Para desarrollar y clasificar un sistema automático de creación de perfiles de autor, es necesario contar con un conjunto de datos de referencia que abarque diferentes géneros, dado que la naturaleza del texto varía de una categoría de género a otra. Por ejemplo, los tweets y los mensajes en Facebook suelen ser breves e informales, mientras que los contenidos de los blogs suelen tener una extensión más considerable y un estilo más formal. Por lo tanto, para un adecuado entrenamiento del sistema de creación de perfiles de autor, se requieren recursos de referencia estándar que abarquen diversos tipos de géneros [3].

A través de los años, el perfilado de autor se ha convertido en una herramienta de gran ayuda para los sociólogos cuando tienen que realizar análisis sobre temas que disponen de información digital; también tiene un gran valor para los partidos políticos, dado que pueden conocer cuál es el perfil de sus votantes. De igual forma ofrece a las empresas una forma de conocer el perfil de los clientes que opinan de forma positiva y negativa acerca de sus productos [5]. Se les ha brindado a los clientes un espacio para evaluar cierto producto, y la mayoría de las reseñas obtenidas se realizan de forma anónima. En este contexto, se investigan las opiniones de estos consumidores considerando variables como su edad, género, idioma nativo, ocupación y características de personalidad. Con base en las puntuaciones obtenidas, las empresas intentan desarrollar nuevas estrategias comerciales para satisfacer mejor las necesidades de los clientes [3].

Capítulo 2

Trabajos previos

En este capítulo se presenta una recopilación y análisis de los trabajos desarrollados en el campo del perfilado de autor en textos escritos por usuarios de diversas redes sociales y contextos digitales. Se examinan los enfoques utilizados para identificar características de los autores,tales como nacionalidad, edad, género, ocupación y rasgos de personalidad. Además, se examinan las técnicas de procesamiento de lenguaje natural y de aprendizaje automático empelados para resolver la tarea.

2.1. Diversidad de enfoques para el perfilado de autor

2.1.1. Enfoque basado en estilo

La estilometría se define como una técnica que utiliza métodos computacionales para llevar a cabo un estudio cuantitativo del estilo literario. Se sustenta en la observación de que cada autor posee un estilo de escritura único, el cual refleja su personalidad y características lingüísticas distintivas. Esta disciplina categoriza los rasgos estilísticos en tres principales grupos: sintácticos, léxicos y estructurales. Su propósito radica en identificar patrones lingüísticos que no pueden ser percibidos mediante métodos convencionales, lo que permite analizar patrones de estilo en corpus [6].

Los estudios en este campo suelen implicar la extracción y el análisis de una variedad de características estilísticas a partir de conjuntos de datos textuales. Estas características se utilizan luego para entrenar modelos de aprendizaje automático, que pueden predecir atributos del autor con base en el estilo de escritura [6].

- **Características sintácticas:** se centran en los patrones y estructuras utilizados para construir oraciones en un texto. En el análisis de frases sintácticas, relaciones entre palabras, dependencias gramaticales y reglas de escritura, como la puntuación, las palabras funcionales y la estructura de las frases. También se considera la longitud de las oraciones, los signos de puntuación, los signos de interrogación, las oraciones de interrogación y las etiquetas de partes del discurso (POS) [7].
- **Características léxicas:** se refieren a aspectos relacionados con el contenido del texto y las unidades léxicas utilizadas por el escritor. Esto implica el análisis de características basadas en caracteres, la frecuencia de caracteres, minúsculas y mayúsculas, y características basadas en palabras, como el número total de palabras, la frecuencia de palabras cortas y largas, y la ocurrencia de palabras en mayúsculas, únicas y repetidas [7].
- **Características estructurales:** se refieren a la organización y disposición general del texto. Esto incluye medidas estadísticas como el número de palabras por oración, oraciones por párrafo, longitud de las palabras y frecuencia de n-gramas. Además de estas medidas estadísticas, las características estructurales también pueden incluir aspectos visuales y de formato del texto, como el tamaño, presencia de emojis y el color de la fuente, así como características propias del tipo de texto, enlaces URL, hashtags y menciones en el caso particular de Twitter [7].

En un estudio reciente [8], se realizó una clasificación por edad y género basada en rasgos estilísticos en reseñas de hotel del conjunto de datos de PAN 2014 en inglés. Se aplicaron procedimientos de preprocesamiento de datos con el objetivo de eliminar hashtags, enlaces URL, espacios en blanco, etiquetas HTML y realizar la tokenización de los textos. El estudio identificó 14 características estilísticas. Estas incluyen longitud de palabra, porcentaje de oraciones interrogativas, uso de puntos y coma, longitud de oraciones cortas y largas, cantidad de mayúsculas y la presencia de dígitos. Se evaluó el rendimiento de seis modelos de aprendizaje automático, entrenados mediante validación cruzada de 10 veces. Estos modelos

incluyen Máquina de Vectores de Soporte (MVS), Bosque Aleatorio (BA), Bayes Ingenuo (BI), Regresión Logística (RL), Árbol de Decisión (AD) y K-Vecinos Más Cercanos (K-VMC).

El artículo [9], describe un método para detectar la edad y el género en chats y blogs en inglés y español utilizando el conjunto de datos PAN 2018. Se elaboraron recursos léxicos, como diccionarios de emojis, abreviaturas, términos comunes en mensajes de texto y contracciones más comunes, que se utilizaron para reemplazar estos elementos en el conjunto de datos. Se emplearon herramientas como Tree-Tagger para el idioma español y Stanford POS-tagger para el idioma inglés para realizar el etiquetado de palabras. Se presentaron dos enfoques de evaluación utilizando el modelo de clasificación de Bosque Aleatorio (BA). En el primero, se clasificó inicialmente el género del autor, seguido de la clasificación por edad; luego, se invirtió el proceso. En el segundo enfoque, se empleó la edad como variable de clasificación en el modelo. Una vez identificada la edad del autor, se infería su género, basado en la edad determinada en la fase previa.

2.1.2. Enfoque basado en contenido

Este enfoque se centra en la extracción de palabras clave, la secuencia ordenada de palabras o caracteres, las características temáticas, las estructuras semánticas y la representación en bolsas de palabras (BoW). Tiene como objetivo la predicción de diversos atributos, tales como género, grupo de edad, país o región de origen, así como las preferencias de los autores de textos. Estas predicciones se fundamentan en el análisis del contenido textual, lo que permite inferir información significativa sobre los autores basándose únicamente en el contenido de sus escritos [6].

En el estudio [10], se desarrolló un sistema utilizando características basadas en contenido, para predecir la edad y el género de los autores a partir de sus textos en los conjuntos de datos de PAN 2014 y 2016. El proceso incluyó limpiar el conjunto de datos, eliminando puntuaciones, palabras irrelevantes y simplificando las palabras a su forma básica. Se identificaron las palabras más importantes basadas en la frecuencia de ocurrencia y se formaron los vectores de características. Se propuso un nuevo método para calcular la importancia de cada palabra en diferentes contextos: dentro de un documento específico, en documentos positivos y en documentos negativos. Se evaluaron dos algoritmos de aprendizaje automático Máquinas de Soporte Vectorial (SVM) y Bosques Aleatorios (RF).

En el estudio [11], se describe un enfoque basado en contenido para el perfilado de autor utilizando algoritmos de aprendizaje automático, tales como Bayes Ingenuo (BI), SMO (Sequential Minimal Optimization), Regresión Logística (RL), Bosque Aleatorio (RF) y J48. Para el entrenamiento de estos modelos se emplean características basadas en palabras y caracteres, así como n-gramas tradicionales de etiquetas de partes del discurso. Se utilizan los conjuntos de datos PAN 2014 y PAN 2016, etiquetados con rango de edad y género. Se evalúan diferentes enfoques, incluyendo uni-gramas de palabras y tri-gramas de caracteres, y se comparan con enfoques de referencia. Para la evaluación del rendimiento se utiliza validación cruzada de 10 veces en el entorno de WEKA. Además, se emplea el método de Ganancia de Información en la selección de características para determinar los atributos más relevantes.

2.1.3. Enfoque hibrido

Combina características estilísticas y de contenido. Aprovechando tanto la información lingüística como temática presente en los textos analizados [6]. El trabajo [12] realiza una clasificación por género de los autores en conjuntos de datos de Twitter utilizando características estilísticas, N-gramas y características basadas en contenido. Se emplearon dos algoritmos de clasificación: Bosque Aleatorio (RF) y Bayes Ingenuo Multinomial (BIM), para el análisis del conjunto de datos PAN 2019. El modelo de Bolsa de Características (BoF) fue seleccionado como la metodología para la vectorización de los documentos. Se llevaron a cabo técnicas de preprocesamiento en el conjunto de entrenamiento, que incluyeron la eliminación de palabras sin poder discriminatorio y llevarlas a su forma básica. Se identificaron cuatro tipos principales de características a partir del conjunto de datos: características basadas en palabras, basadas en caracteres, estructurales y sintácticas.

2.2. Conjuntos de datos

Para la tarea de Perfilado de Autor en redes sociales, existen diversas fuentes para la recopilación y análisis de datos como Twitter, Blogs, Facebook e Instagram. Entre los conjuntos de datos empleados para esta tarea, se destacan aquellos proporcionados por PAN at CLEF. Estos conjuntos abarcan textos en múltiples idiomas, desde inglés hasta árabe, y han

sido empleados por numerosos investigadores para el desarrollo de tareas relacionadas con el perfilado de autor, la ética computacional y la atribución de autoría [13].

La Iniciativa CLEF (Conference and Labs of the Evaluation Forum), se enfoca en la evaluación y desarrollo de metodologías para sistemas de información utilizando datos multilingües y multimodales, PAN se distingue por centrarse en la estilometría y el análisis forense de textos digitales. Desde su integración en CLEF en 2010, los laboratorios de evaluación PAN han proporcionado un entorno propicio para la investigación en este ámbito específico [14].

PAN facilita la colaboración y el intercambio de conocimientos entre investigadores interesados en el análisis de textos digitales, promoviendo avances significativos en áreas como la identificación de autores y la detección de plagio en entornos digitales. Esta iniciativa no solo fomenta el desarrollo de nuevas metodologías, sino también se ha encargado de proporcionar conjuntos de datos y evaluar el rendimiento de enfoques y algoritmos pertenecientes al campo del perfilado de autor [15].

Durante el período comprendido entre 2013 y 2023, PAN ha organizado tareas anuales de perfilado de autor, ofreciendo conjuntos de datos para el entrenamiento de modelos de aprendizaje automático. La tabla 2.1 proporciona un resumen detallado de estas tareas, abordando fuentes de datos, idiomas, atributos específicos y tamaños de los conjuntos de datos [13].

Tabla 2.1: Evolución de tareas y conjuntos de datos en el Perfilado de Autor (2013-2023)

Año	**Tarea**	**Fuente**	**Idiomas**	**Atributos**	**Tamaño**
2013	Identificar edad y género	Blogs	Inglés, español	Edad, género	713 MB
2014	Evaluar cómo los enfoques de detección se comportan con diferentes categorías de textos	Redes sociales, blogs, Twitter y reseñas de hoteles	Inglés y español	Edad y género	205 MB
2015	Identificar edad, género y cinco rasgos de personalidad	Twitter	Inglés, español, italiano y holandés	Edad, género y personalidad	2 MB
2016	Entrenamiento y evaluación de modelos con distintas categorías de texto	Twitter, redes sociales, blogs, ensayos y reseñas de hotel	Inglés, español y holandés	Edad y género	2 MB
2017	Identificar género y variedad lingüística	Twitter	Árabe, inglés, portugués y español	Género y variedad lingüística	254 MB
2018	Identificar el género desde una perspectiva multimodal, analizando texto e imágenes	Twitter	Árabe, inglés, español	Género	7 GB
2019	Determinar si el autor es bot o humano y perfilar género	Twitter	Inglés y español	Género	38 MB
2020	Perfilar a difusores de noticias falsas	Twitter, webs de fact-checking como PolitiFact o Snopes	Inglés y español	Contenido con noticias falsas en tweets	7 GB
2021	Perfilar difusores de discursos de odio	Twitter	Inglés y español	Contenido de odio en tweets	3 MB
2022	Perfilar ironía y difusores de estereotipos hacia mujeres o la comunidad LGTB	Twitter	Inglés	Contenido irónico y presencia de estereotipos en tweets	6 MB
2023	Perfilar influencers de criptomonedas	Twitter	Inglés	Influencia, intereses e intenciones en tweets sobre criptomonedas	202 KB

El estudio [16] detalla el proceso de creación del Corpus BT-AP-19. Este procedimiento involucró la recolección manual de identificadores de Twitter junto con datos demográficos, y el uso de la API de Twitter para obtener los tweets de usuarios que proporcionaron su identificación. Además, se empleó un formulario de Google para la recopilación de datos demográficos y se desarrolló un software para la extracción automática de tweets. El Corpus BT-AP-19 consta de 339 perfiles de autores y un total de 41,864 tweets, con un promedio de 124 tweets por perfil. Asimismo, se presenta un análisis de las cinco palabras más frecuentes en el corpus, tanto en inglés como en Roman-Urdu. Se ofrece una distribución de perfiles según diversos atributos demográficos, como edad, género, nivel educativo, ubicación geográfica, idioma y afiliación política. Además, se llevaron a cabo pruebas utilizando cuatro modelos de aprendizaje profundo: Red Neuronal Convolucional (CNN), Memoria a Largo Plazo y Corto Plazo (LSTM), Memoria a Largo Plazo y Corto Plazo Bidireccional (Bi-LSTM) y Unidades Recurrentes con Puerta (GRU).

Debido a la falta de un conjunto de datos público, el estudio [17] llevó a cabo la construcción de un conjunto de datos de Twitter etiquetado con el género a partir de los tweets. Se utilizó un conjunto de datos existente de Twitter con 20,000 filas para identificar los ID de usuario y sus respectivos géneros. Se excluyó del conjunto a aquellos usuarios cuya identificación como masculino o femenino no era clara. Posteriormente, se realizó un preprocesamiento de los tweets con el objetivo de eliminar elementos no deseados, como URL, hashtags y emojis. Además, se realizó un filtrado para asegurar que los tweets estuvieran escritos en inglés, y por longitud de tweet de 5 a 50 palabras. Para reducir la dimensionalidad de los datos y graficarlos, se aplicó el Análisis de Componentes Principales (PCA).

Este artículo [18] propone un sistema para inferir el género de los autores de textos escritos en idioma egipcio. Se emplea un conjunto de datos creado con textos de Twitter y etiquetados con el género. Se desarrolla una solución de clasificación de texto que utiliza un vector de características mixtas, el cual incluye emojis, sufijos femeninos y palabras funcionales, además de un vector de n-gramas de características. Se realizan experimentos con diversos clasificadores, entre los cuales se incluyen Bosque Aleatorio, Naive Bayes Multinomial, Naive Bayes Bernoulli, Regresión Logística y Gradiente Estocástico Descendente.

2.3. Sistemas similares

No se encontró algún sitio web que realice la tarea de predicción de género y variedad linguística al mismo tiempo, sin embargo, existen plataformas en línea que ofrecen el servicio de la predicción de género.

2.3.1. uClassify

Esta plataforma en línea ofrece una variedad de servicios en relación al análisis de textos. Cuenta con clasificador de género, análisis de sentimientos, clasificación de temas. Su funcionamiento consiste en ingresar un texto y como resultado se visualiza el porcentaje de probabilidad de que el autor sea hombre o mujer. Además ofrece una herramienta para clasificar usuarios de acuerdo al URL de su perfil, sin embargo, esta opción siempre marca error. [19].

Este sistema fue entrenado con 11,000 blogs, de los cuales 5,500 estuvieron orientados al género femenino y 5,500 al género masculino. Según los datos de la página, la precisión de los clasificadores de esta plataforma varía entre el 70 % y 80 %, dependiendo de factores como el contenido y la longitud del texto [20]. Sin embargo, las pruebas que se realizaron en esta página, orientadas a la predicción del género masculino, siempre fueron erróneas.

En la Figura 2.1 se muestra el resultado obtenido después de realizar una prueba conformada por 100 tweets escritos por un hombre.

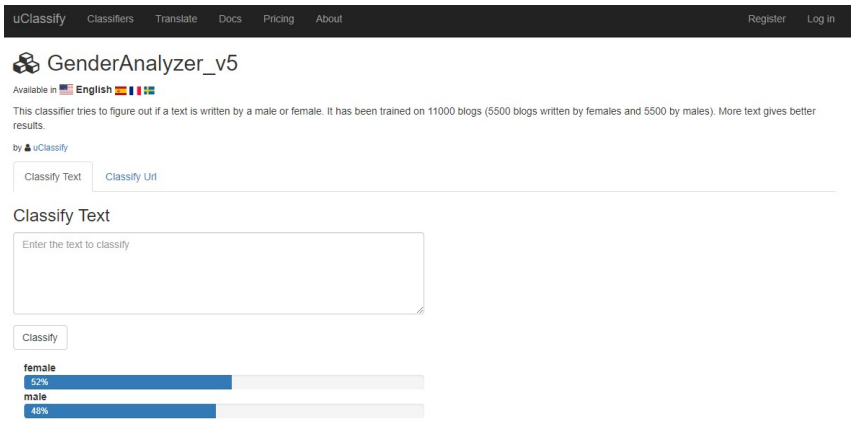

Figura 2.1: Sitio web uClassify

2.3.2. Readable

La plataforma Readable le brinda a sus usarios un analizador de textos que permite detectar si el autor del texto es hombre o mujer. Su función consiste en analizar un texto y compararlo con los datos de un corpus conocido por medio del análisis de la frecuencia de palabras [21].

En la Figura 2.2 se visualiza una prueba conformada por el análisis de 100 tweets de un autor de género masculino.

Los datos proporcionados por la plataforma indican que la precisión del sistema es del 70 %, sin embargo, los resultados obtenidos en todas las pruebas orientadas al género masculino, fueron erróneos.

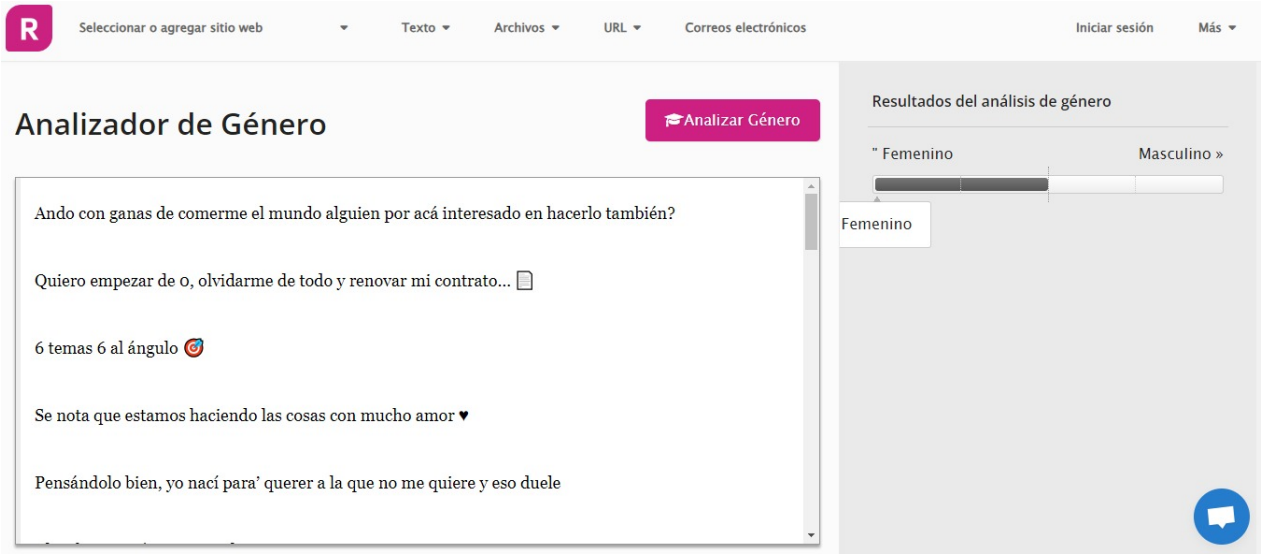

Figura 2.2: Sitio web predictor de género en base a textos

Capítulo 3

Marco teórico

En este capítulo se describen conceptos relacionados con la tarea del perfilado de autor mediante el uso de algoritmos de aprendizaje automático. Se describen las principales representaciones de un texto dado, las características generales de los clasificadores y las medidas de evaluación empleadas para medir los resultados obtenidos.

3.1. Aprendizaje Automático

El aprendizaje automático, conocido también como *machine learning*, es un campo de la inteligencia artificial que se dedica al desarrollo de sistemas que pueden aprender, ajustarse y optimizarse de manera independiente basándose en experiencias previas. Una característica clave del aprendizaje automático es su enfoque en la estadística, lo que permite a los sistemas examinar los datos proporcionados, encontrar patrones entre ellos y convertirlos en información o conocimiento. [22].

En la clasificación de textos, existen dos enfoques predominantes. El primero, basado en la ingeniería del conocimiento, emplea reglas de clasificación para integrar el conocimiento de expertos. El segundo, centrado en el aprendizaje automático (ML: *Machine Learning*), utiliza procesos inductivos para crear un clasificador mediante el aprendizaje a partir de ejemplos ya etiquetados.

Se han observado varios resultados respecto al desempeño entre la ingeniería del conocimiento y los sistemas de aprendizaje automático en la gestión documental. Normalmente, la ingeniería del conocimiento supera a los sistemas de aprendizaje automático, pero actualmente esta diferencia se está reduciendo progresivamente debido a un incremento en las investigaciones enfocadas en el aprendizaje automático. Esta tendencia se debe en parte a la complejidad involucrada en la creación y el mantenimiento de reglas de codificación del conocimiento, mientras que el aprendizaje automático requiere un conjunto de ejemplos clasificados manualmente, lo cual es menos costoso. La ingeniería del conocimiento se enfoca en desarrollar reglas de clasificación de manera manual, donde un experto en el campo define condiciones para categorizar un documento específico, lo cual puede ser un proceso que consume mucho tiempo y esfuerzo humano. En contraste, los sistemas de aprendizaje automático pueden asignar etiquetas al contenido de manera automática o semiautomática. Estos sistemas utilizan algoritmos para observar cómo se etiquetan los objetos y proponen alternativas para etiquetas existentes o nuevas para contenido no etiquetado. Los algoritmos de clasificación aprenden a partir de ejemplos utilizando datos que han sido organizados en diferentes categorías de forma manual o mediante algún proceso automatizado. [23].

Un clasificador, al recibir un conjunto de atributos de un objeto, busca asignarle una etiqueta. Para ello, se basa en el conocimiento obtenido a partir de ejemplos previos de cómo se han etiquetado otros objetos. Estos ejemplos, conocidos como datos de entrenamiento, proporcionan la información que el clasificador utiliza para tomar decisiones sobre objetos no analizados anteriormente. La categorización, por otro lado, se enfoca en asignar una categoría a un objeto. En el contexto específico de la categorización de documentos, este proceso implica asignar una categoría a un texto utilizando características comunes. Aunque en esta etapa las categorías suelen estar basadas en el tema del documento, también existen aplicaciones que categorizan documentos mediante el análisis de sentimientos. Por ejemplo, podríamos tener categorías como: positivo o negativo en una revisión de producto, o identificar las emociones en un correo electrónico o una solicitud de soporte al cliente. [23].

La extracción de datos ofrece instrumentos potentes para identificar patrones ocultos y conexiones en conjuntos de datos organizados. Este procedimiento parte de la premisa de que los datos a utilizar ya están guardados en un formato estructurado. Por ello, su

preprocesamiento se centra principalmente en la depuración y normalización de los datos. [23].

El proceso de *machine learning* comprende varias fases, incluyendo la recolección y el procesamiento preliminar de datos. Es crucial que los datos obtenidos sean limpiados y manipulados para asegurar la uniformidad del conjunto de datos de entrada. Estos datos se dirigen a la fase de entrenamiento, donde se seleccionan etiquetas en función de características relevantes y se transforman los datos al formato necesario para el algoritmo de entrenamiento. Luego se elige un modelo y se realizan pruebas con los datos procesados para mejorar la precisión del algoritmo. Después de que el modelo ha sido entrenado adecuadamente, se avanza a la fase de evaluación, utilizando datos que no han sido etiquetados previamente. Dependiendo de los resultados obtenidos, se decide si es necesario regresar a fases anteriores para obtener más datos, preparar los datos de forma diferente, seleccionar distintos algoritmos o ajustar otros parámetros, o si el modelo muestra un desempeño satisfactorio y está listo para su implementación o pruebas en un entorno de producción.

Figura 3.1: Proceso del aprendizaje automático.

El ámbito del machine learning se fundamenta en principios y conceptos derivados de diversas áreas, como la inteligencia artificial, la estadística, la teoría de la información y la complejidad computacional. Los algoritmos de aprendizaje han mostrado su eficacia en múltiples campos de aplicación, como la extracción de patrones en vastas bases de datos, permitiendo descubrir regularidades implícitas de manera automatizada. Además, resultan

útiles en áreas poco comprendidas donde los humanos no poseen el conocimiento necesario para desarrollar algoritmos efectivos, así como en dominios que requieren una adaptación dinámica de programas para responder a cambios en el entorno. Por otro lado, las técnicas de aprendizaje son ampliamente empleadas para la clasificación de textos, definiéndose como procesos inductivos que construyen clasificadores automáticamente a partir de un conjunto de documentos previamente clasificados. [22].

Las técnicas de aprendizaje se clasifican en supervisadas y no supervisadas. En las supervisadas, el propósito es as aprender el mapeo de las respuestas correctas para los datos de entrada proporcionados, utilizando un conjunto de datos de entrenamiento compuesto por pares de entrada y salida correcta. En contraste, las técnicas de aprendizaje no supervisado no necesitan un conjunto de datos previamente etiquetados, y su meta es descubrir patrones significativos observando la distribución y la estructura de los datos proporcionados. [22]. Ejemplos de técnicas de aprendizaje supervisado incluyen árboles de decisión y máquinas de vectores de soporte, entre otros. Mientras tanto, ejemplos de técnicas de aprendizaje no supervisado incluyen técnicas de *clustering*.

De una forma más abstracta, si vemos el aprendizaje como un proceso de usar la experiencia para obtener conocimientos, en este contexto, la experiencia obtenida tiene como finalidad predecir la información ausente en los datos de prueba. En estas situaciones, podemos considerar al entorno como un instructor que guía al estudiante al darle información adicional (etiquetas).

En el aprendizaje supervisado para la clasificación de textos, se empieza con un conjunto de ejemplos etiquetados que se presentan al algoritmo para desarrollar un clasificador. Luego, se emplean ejemplos sin etiquetar para medir la eficacia del clasificador. [24].

Por otro lado, en el aprendizaje no supervisado, no se diferencia entre los datos de entrenamiento y los datos de prueba. El modelo maneja los datos de entrada con el propósito de generar algún tipo de resumen o versión comprimida de esos datos. Un ejemplo común de esta tarea es la agrupación de un conjunto de datos en subconjuntos de elementos semejantes. [25].

Los algoritmos de aprendizaje no supervisado se utilizan en contextos donde los datos no tienen etiquetas y no hay un valor objetivo establecido. Esta naturaleza de no supervisión implica que no necesitan la intervención humana para detectar patrones o agrupaciones ocultas dentro

de la estructura de los datos. Su habilidad para identificar similitudes o diferencias entre los datos los hace perfectos para aplicaciones como el análisis exploratorio de datos, estrategias de negocios, segmentación de clientes y reconocimiento de imágenes. [22].

3.2. Procesamiento de Lenguaje Natural (PLN)

El lenguaje se define como el método a través del cual los seres humanos se comunican y expresan sus ideas, opiniones y pensamientos. Utiliza medios como la escritura y el habla para facilitar una interacción efectiva. El lenguaje natural se refiere a la comunicación entre dos individuos, ya sea a través de idiomas comunes, lenguaje de señas, códigos o claves establecidas.

El procesamiento del lenguaje natural (PLN) implica la capacidad de una máquina para manejar información expresada mediante el lenguaje humano. En PLN, las computadoras examinan, interpretan y dan sentido al lenguaje humano, permitiendo su aplicación práctica. [26]. Se podría decir que el PLN emplea una forma de comunicación natural para interactuar directamente con una computadora. Se desarrollan modelos lingüísticos computacionales minuciosamente para permitir la creación de programas capaces de ejecutar diversas tareas o solicitudes utilizando el lenguaje humano. [27].

El campo del procesamiento del lenguaje natural ha experimentado un notable crecimiento en los últimos años. Sus áreas de estudio abarcan la recuperación y extracción de información, la minería de datos, la traducción automática, el análisis de sentimientos, la generación de resúmenes automáticos, los sistemas de búsqueda de respuestas, entre otros.

El PLN tiene aplicaciones significativas en la actualidad, como lo son los chatbots, cuya dinámica es entablar una conversación con el usuario utilizando su mismo lenguaje. Existen diversos tipos de chatbots según su función, los chatbots conversacionales como Alexa y Siri, a pesar de ser complejos, son más amigables con el usuario ya que le permite enviar cualquier pregunta o respuesta. Similares a los chatbots conversacionales, los chatbots contextuales son capaces de comprender las intenciones del usuario y dar una respuesta expresándose de la misma forma que lo haría una persona, lo cual permite tener una conversación fluida. A diferencia de los tipos de chatbots anteriores, los chatbots en regla, siguen una estructura predefinida durante la interacción con el usuario, son utilizados en la tarea de soporte al

cliente ya que ofrece un menú de opciones disponibles para el usuario que dirigen el rumbo de la conversación. [22]

Además de ello, el PLN también tiene aplicación en la clasificación de documentos y mensajes, análisis de sentimientos y opiniones, detección de similitudes o anomalías en los textos, búsqueda avanzada de información, detección de entidades (personas, lugares, etc.) [28]. Por otro lado, cuenta con beneficios como agilizar y optimizar tareas que anteriormente se realizaban manualmente, reconocimiento automático de patrones, recomendaciones de productos y contenido agradable para el usuario en base a los gustos compartidos en sus redes sociales, entre otros. [22]

Cuando se habla de un grupo de objetos, la tarea de clasificarlos consiste en asignarlos a un conjunto predefinido de categorías. La categorización automática de documentos es una forma de clasificación de patrones, lo cual es esencial para garantizar una gestión eficaz de sistemas de información de texto. Esta técnica se emplea en diversas aplicaciones como el proceso de analizar y estructurar documentos de texto para facilitar la búsqueda de información relevante dentro de ellos y así emplearlos para la filtración de correos no deseados, la entrega personalizada de contenido y la identificación del género de textos, entre otros usos [23].

El análisis de sentimientos en textos implica la identificación y extracción de información subjetiva, también conocida como minería de opiniones. Este proceso generalmente implica el uso de herramientas de procesamiento del lenguaje natural (PLN) y software de análisis de textos para automatizarlo. Técnicas avanzadas permiten realizar un análisis gramatical y descomponer la oración. [23]

El análisis de sentimientos responde a la necesidad de identificar y resumir opiniones en grandes cantidades de texto con fines de mercadeo y gestión de imagen. Sin embargo, los algoritmos heurísticos presentan el problema de la dificultad para identificar manualmente todos los patrones que expresan sentimientos. Por ello, la siguiente fase de investigación se enfoca en la vasta información disponible en Internet, como comentarios de diversas fuentes. Se utilizan reglas gramaticales, similares a las de los compiladores, para extraer inferencias. El motor de reglas se aplica repetidamente para transformar el texto etiquetado en oraciones que definen la relación entre una palabra y una categoría gramatical con un sentimiento evaluado. Para la implementación, se emplean herramientas de etiquetado y una base de datos que contiene claves/frases con valoraciones de la polaridad emocional. Se

requiere una cantidad considerable de ejemplos para que el clasificador comprenda cómo las características se relacionan con las categorías. La cantidad de muestras necesarias dependerá de la complejidad de la tarea de clasificación, incluyendo el número de clases, características y la dimensionalidad de las reglas de clasificación. [23]

3.2.1. Modelo de Bolsa de Palabras

El modelo de Bolsa de Palabras o BoW (Bag of Words) es una técnica simple de representación de documentos que consiste en la extracción de las palabras que aparecen en un texto, sin tener en cuenta sus características gramaticales o el orden en el que aparecen. De este modo, se obtiene un vector de características en el que cada posición representa una palabra y que tiene como valores el número de veces que aparece dicha palabra en el documento. [5]

El modelo Bolsa de Palabras, inicialmente desarrollado en el campo del análisis de lenguaje, se diseñó para clasificar automáticamente el contenido de textos escritos, facilitando la identificación del tema de los documentos sin necesidad de intervención humana. En su forma más elemental, BoW mide la cantidad de veces que aparecen las palabras en un conjunto de documentos, sin tener en cuenta su secuencia ni las normas gramaticales o sintácticas del idioma. [29]

Cada documento se describe mediante un histograma que muestra la frecuencia de las palabras que contiene, donde cada barra del histograma representa la cantidad de veces que aparece una palabra específica, formando así la "bolsa de palabras"del documento. La idea principal es que documentos que sean parecidos tendrán bolsas de palabras similares. Por ejemplo, se espera que dos textos relacionados con política tengan una alta frecuencia de términos como presidente, debate, y política. De manera similar, los textos sobre biología tendrán frecuencias elevadas de palabras como vida, células y fotosíntesis. Cuanto mayor sea el porcentaje de palabras comunes entre dos documentos, mayor será la similitud entre sus histogramas. Además, este método permite clasificar la relevancia de un texto para responder a determinadas consultas. [29]

Es importante destacar que no todas las palabras de un corpus son igualmente relevantes para revelar el contenido de los documentos. Es por ello que antes del cálculo de los histogramas, la metodología BoW incluye la creación de un diccionario de términos relevantes, eliminando

palabras demasiado frecuentes o poco frecuentes. Dado que una colección de documentos suele incluir textos de diferentes longitudes, se lleva a cabo un proceso de normalización para convertir la representación en una función de densidad de probabilidad. Esto garantiza que cada barra del histograma represente la contribución proporcional de una palabra al contenido del documento. [29]

Una vez calculados los histogramas de todo el corpus, estos pueden ser organizados por categorías temáticas, ordenados según su grado de similitud, o examinados a través de la comparación de la frecuencia de ciertos términos. Para que este método sea funcional, es esencial desarrollar un sistema automático de identificación, que se divide en dos etapas: la primera etapa consiste en la extracción de histogramas de una extensa colección de textos, y la segunda etapa se centra en la aplicación de métodos de consulta basados en el aprendizaje automático adquirido en la etapa anterior. [29]

3.3. Métodos Supervisados

Una vez que son extraídas las características formales del texto, el siguiente paso es utilizarlas para su clasificación. Para ello, existen una serie de algoritmos capaces de aprender los patrones existentes en los conjuntos de datos y utilizarlos para etiquetar o clasificar datos nuevos. [5]

Los algoritmos utilizados en el aprendizaje supervisado operan con bases de datos que han sido previamente categorizadas en términos de entrada y salida. Su propósito principal es asignar categorías a los datos o estimar valores con un grado específico de exactitud. Para alcanzar este objetivo, desarrollan un modelo donde los datos de entrada se alinean con la salida esperada. En el transcurso de este procedimiento, el método ajusta sus parámetros basándose en los errores cometidos en iteraciones anteriores, afinando su capacidad para adaptarse a la nueva información de entrada. [22]

3.3.1. Árboles de Decisión

Los Árboles de Decisión son un tipo de algoritmo de aprendizaje supervisado que se basa en la creación de un modelo visual de predicción en forma de árbol y permite visualizar los

atributos que están aportando información relevante en los modelos de aprendizaje y observar las jerarquías de importancia entre los atributos. [31]

En este sentido, cada nodo interno del árbol representa una característica del conjunto de datos y cada rama representa una condición sobre los valores de dicha característica. Así, para clasificar un nuevo dato, se recorre el árbol desde la raíz hasta una de sus hojas, la cual representa la clase a la que pertenece. [5]

Los árboles de decisión, que presentan una estructura parecida a la de los sistemas fundamentados en reglas, son bien conocidos por su eficiencia y flexibilidad en la representación y clasificación de condiciones sucesivas para abordar diferentes problemas. Su uso se destaca como uno de los enfoques de clasificación más comunes y ampliamente implementados en la práctica. [30]

El conocimiento obtenido a través del proceso inductivo se representa como un árbol, donde el nodo principal, conocido como la raíz, actúa como el punto inicial para la clasificación. Los nodos internos representan las preguntas sobre los atributos específicos del problema. Cada posible respuesta a estas preguntas se divide en nodos secundarios, y las ramas que emergen de estos nodos reflejan los distintos valores que puede tener el atributo. Finalmente, los nodos terminales, o nodos hoja, muestran las decisiones finales, las cuales corresponden a las variables de clase del problema a resolver. [31]

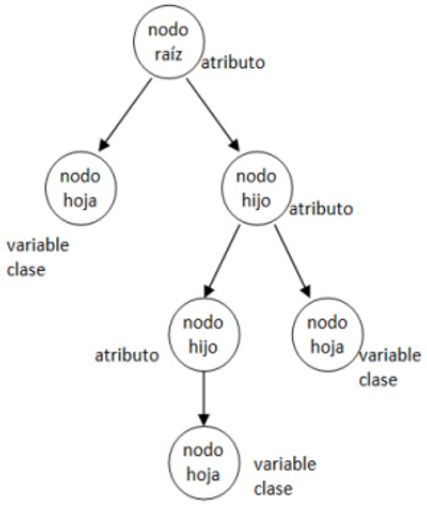

Figura 3.2: Estructura de un árbol de decisión.

La creación de un árbol de decisión se fundamenta en una narración detallada del problema, brindando una visualización gráfica del proceso de decisión que detalla las variables analizadas, las acciones a realizar y la secuencia de decisiones. En cada aplicación de este modelo, se sigue un camino específico según el valor de la variable en cuestión, lo que resulta en una respuesta determinista basada en los datos ingresados. Además, es relevante mencionar que las variables pueden ser discretas o continuas, lo que proporciona versatilidad al modelo. [31].

El algoritmo para la creación de árboles de decisión se divide en dos fases principales: la formación del árbol y la categorización. En la fase inicial, el árbol se construye utilizando un conjunto de datos de entrenamiento, dividiendo repetidamente el conjunto en grupos menores en función de los valores de los atributos elegidos. Este proceso comienza con la generación del nodo raíz, la selección de un atributo para evaluar y la partición del conjunto de entrenamiento en segmentos. Cada partición da lugar a un nuevo nodo, lo que lleva a la expansión continua del árbol. Cuando un nodo contiene elementos de una única categoría, se convierte en una hoja y se le asigna la etiqueta correspondiente a esa categoría. En la segunda fase del algoritmo, se realiza la categorización de nuevos elementos, que son evaluados mediante el árbol construido en la fase anterior. A partir del nodo raíz, se sigue un recorrido específico a través del árbol, guiado por las decisiones tomadas en cada nodo interno. Finalmente, se alcanza una hoja que determina la asignación del elemento a una categoría particular, ofreciendo así una solución al problema planteado. [31]

3.3.2. Máquinas de Soporte Vectorial (SVM)

La máquina de soporte vectorial se fundamenta en algoritmos que aprenden de un conjunto de datos o muestras aleatorias de un sistema a clasificar y, en ocasiones, a predecir su comportamiento futuro, ya sea a corto, mediano o largo plazo. Las máquinas de soporte vectorial utilizan un subconjunto de puntos de entrenamiento llamados vectores de soporte y son efectivas en casos donde el número de dimensiones es mayor al número de ejemplos. Este algoritmo tiene ciertas limitaciones, por ejemplo, si el número de características es mucho mayor al número de ejemplos, se debe elegir una función de núcleo y término de regularización para evitar un sobre ajuste. Dicho problema se presenta frecuentemente en el análisis de historiales muy extensos extraídos de redes sociales [32].

Las SVMs son especialmente adecuadas para la clasificación de dos clases, es decir, para la clasificación binaria. Durante este proceso, se pueden distinguir dos etapas principales: el aprendizaje automático y el reconocimiento. En la etapa de aprendizaje automático, se elige un conjunto de datos de entrenamiento, se extraen los atributos y características del espacio de entrada, y se entrena el clasificador [33].

Su capacidad de clasificación superior se evidencia en la creación de un plano que separa y clasifica los distintos datos proporcionados al algoritmo. Si los datos no pueden ser clasificados de esta manera y quedan datos sin clasificar en la frontera de clasificación, la SVM puede proyectar los datos en un plano de un espacio de dimensión finita y buscar un hiperplano que divida y clasifique correctamente los datos. Con el uso de ciertas técnicas como el truco del kernel, este algoritmo es capaz de maximizar el margen de dicho plano entre los datos de las diferentes clases, ya sean linealmente separables o no [5], esta distancia máxima se conoce como margen funcional. Un margen más amplio genera un menor error de generalización del clasificación. Los puntos de datos que se encuentren más cercanos al margen funcional, son denominados vectores de soporte [33].

El uso de un kernel permite transformar los datos a un espacio de dimensiones superiores con la finalidad de facilitar la separación de las clases. Las funciones utilizadas para definir los vectores de soporte y los hiperplanos pueden ser de diversos tipos, como lineales, polinómicas, y de base radial [33].

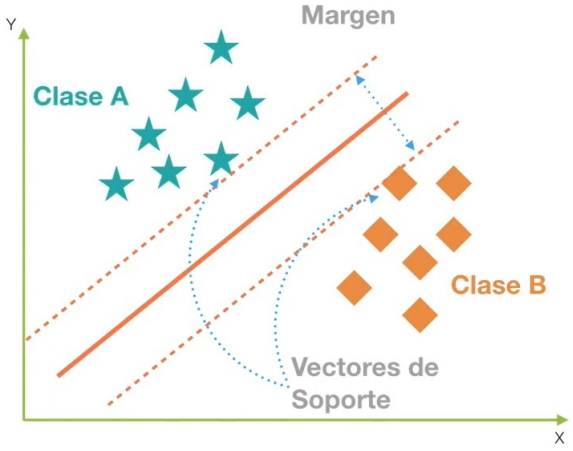

Figura 3.3: Kernel lineal

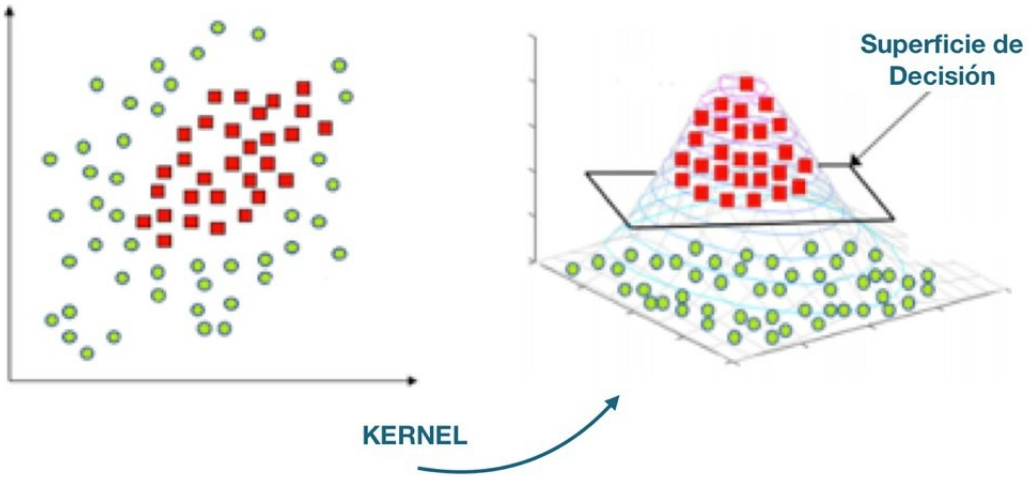

Figura 3.4: Kernel lineal en 3D

3.3.3. Multinomial Naive Bayes

Naive Bayes representa un conjunto de métodos de clasificación basados en el teorema de Bayes, una fórmula estadística que permite calcular la probabilidad condicional de un evento dado que otro evento ha ocurrido. Esta probabilidad condicional se utiliza en la clasificación para determinar la probabilidad de que un dato pertenezca a una clase específica, dado un conjunto de características observadas. Una característica clave de estos métodos es la hipótesis de independencia entre las características de los datos, conocida como la "hipótesis ingenua". Esto significa que cada atributo contribuye de manera independiente a la probabilidad de que un dato pertenezca a una clase, sin interacción con otros atributos. [38].

En el ámbito de la categorización de textos, estos métodos asumen que la probabilidad de que una palabra se presente en un texto es independiente de la presencia de otras palabras en el mismo texto. Esto implica que el modelo trata cada palabra como una característica individual y no considera la estructura gramatical o el significado semántico de las oraciones. [38]. Existen dos tipos principales de clasificadores Naive Bayes utilizados en el procesamiento de textos:

El modelo Naive Bayes Bernoulli Multivariado describe un documento como un vector binario, usando 1 para indicar la presencia de un término y 0 para su ausencia. Este método es eficaz cuando solo se tiene en cuenta si los términos están presentes o no.

Por otro lado, el enfoque Multinomial Naive Bayes no solo considera la presencia de términos, sino también la frecuencia con la que aparecen en el documento. De este modo, cada término aporta de manera proporcional a la probabilidad total de que un documento se clasifique en una categoría particular. [38].

3.4. Métricas de evaluación

Es importante conocer el rendimiento de cada modelo de aprendizaje supervisado y para ello se utilizan diversas medidas de evaluación. Sin embargo, antes de aplicarle alguna métrica de evaluación al modelo, es importante elaborar una matriz de confusión para poder agilizar el proceso ya que cada métrica utiliza los números que se encuentran contenidos dentro de la matriz [6].

Tabla 3.1: Matriz de confusión para evaluación de modelos.

	Predicción: 1	**Predicción: 0**
Real: 1	VP	FN
Real: 0	FP	VN

Cada predicción puede ser uno de cuatro resultados, basado en cómo coincide con el valor real:

- Verdadero Positivo (VP): Predicción Verdadero y Verdadero en realidad.
- Verdadero Negativo (VN): Predicción Falso y Falso en realidad.
- Falso Positivo (FP): Predicción de verdadero y falso en la realidad.
- Falso Negativo (FN): Predicción de falso y verdadero en la realidad.

Con los datos correspondientes obtenidos en la matriz de confusión, se emplean diversas métricas de evaluación como lo son la exactitud, precisión, el recuerdo y F1.

La exactitud (*accuracy*) representa la proporción del número de ejemplos en el conjunto de evaluación que son clasificados correctamente por el modelo (ver ec. 3.1).

La exactitud es utilizada cuando las clases tienen la misma importancia para la clasificación.

$$Exactitud = \frac{VP + VN}{VP + VN + FN + FP} \tag{3.1}$$

La precisión (ec. 3.2) representa la proporción del número de factores positivos predichos correctamente:

$$Precision = \frac{VP}{VP + FP} \tag{3.2}$$

El recuerdo (recall) es la proporción de los factores positivos que lograron ser recuperados y se presenta en la ecuación 3.3:

$$Recuerdo = \frac{VP}{VP + FN} \tag{3.3}$$

La ecuación 3.4 proporciona una puntuación más realista ya que considera tanto a la precisión como al recuerdo:

$$F1 = \frac{2 \cdot \text{precisión} \cdot \text{recuerdo}}{\text{precisión} + \text{recuerdo}} \tag{3.4}$$

3.5. Python

Python es un lenguaje de programación de alto nivel. Creado e implementado a principios de 1990 por Guido Van Rossum en el Centro de Matemáticas e Informática (CWI) en los Países Bajos. Desde entonces, ha experimentado un crecimiento significativo y se ha convertido en uno de los lenguajes de programación más populares y ampliamente utilizados en el mundo [39].

Tabla 3.2: Comparación de características entre Python, Java y C#

Paradigma	**Python**	**Java**	**C#**
POO	POO	POO	POO
Ejecución	Interpretado: código leído y ejecutado línea a línea por un intérprete.	Compilado e interpretado: código se compila a bytecode y luego interpretado por la Máquina Virtual de Java.	Compilado: código se compila en un lenguaje intermedio (IL) y ejecutado por el Common Language Runtime (CLR) de .NET.
Tipado	Dinámico y fuerte	Estático y fuerte	Estático y fuerte
Multiplataforma	Puede ejecutarse en diversos sistemas operativos.	Ejecutable en una variedad de arquitecturas de hardware y software, siempre que se cuente con la JVM.	Código generado puede interactuar con otros lenguajes .NET.
Sintaxis	Simple y concisa	Rigurosa, requiere más líneas de código	Intuitiva y estructurada, aunque puede requerir más líneas de código.
Bibliotecas	Amplio acceso a bibliotecas y frameworks de código abierto.	Herramientas disponibles, aunque su implementación puede ser compleja.	Adaptaciones disponibles, pero con recursos limitados en comparación con las versiones originales en Python o Java.

3.5.1. Bibliotecas y módulos de Python para PNL

Python es una herramienta flexible y aplicable a diversos campos del desarrollo de software, como el procesamiento del lenguaje natural, la visión computacional, la interfaz gráfica de usuario, el desarrollo de videojuegos, la creación de aplicaciones web, la robótica, la ciencia de datos y la inteligencia artificial. En el documento, se describen áreas de aplicación particulares junto con las bibliotecas principales utilizadas en cada una. [40].

A continuación, se proporciona una descripción de algunas de las bibliotecas de Python empleadas en el desarrollo de aplicaciones dentro del campo del procesamiento de lenguaje natural. [40].

1. **Natural Language Toolkit (NLTK):** Es una biblioteca que proporciona acceso a 50 conjuntos de datos (corpus) y recursos léxicos, como WordNet, una base de datos en idioma inglés. También, incluye bibliotecas y módulos para realizar tareas como clasificación de texto, tokenización, derivación, etiquetado, análisis sintáctico y razonamiento semántico, bibliotecas de PNL de potencia industrial, y un foro de discusión activo. Es compatible con diversos sistemas operativos como Windows, Mac OS X y Linux. Además, es gratuito y de código abierto [40].

2. **Scikit-learn:** Es una biblioteca de código abierto utilizada para realizar aprendizaje automático, cuenta con algunos módulos para la evaluación de los modelos, como acuraccy_score y f1_score. Tambipen cuenta con diversos módulos para la tarea de clasificación de las clases, entre ellos:

   `-CounVectorizer`: Convierte una colección de cdocumentos de texto en una matriz de recuento de tokens. [41]

   `-svm:` Este módulo implementa el algoritmo de máquinas de soporte vectorial y cuenta con la clase SVC para realizar una clasifiación binaria. [42]

   `-MultinomialNB:` Implementa el algoritmo de Bayes para datos distribuidos multinomialmente [43]

   `-tree:` Es una clase con la capacidad de clasificar múltiples clases en un conjunto de datos. Toma como parámetros de entrada dos matrices, una que contiene las muestras de entrenamiento y otra que contiene las etiquetas de las clases para las muestras. [44]

Capítulo 4

Metodología propuesta

Este capítulo describe las diferentes fases involucradas en la metodología para la creación de perfiles de autor. Esta metodología se basa en la utilización de modelos de aprendizaje supervisado y técnicas de Procesamiento de Lenguaje Natural (PLN). Se realiza el análisis de textos en español obtenidos de plataformas de redes sociales, particularmente tweets. El propósito principal es identificar tanto el género como la variante lingüística del autor, considerando diferentes países: México, Perú, Argentina, España, Chile, Colombia y Venezuela.

4.1. Descripción de la metodología

El modelo propuesto para la clasificación de autores por género y variedad lingüística (México, Perú, Argentina, España, Chile, Colombia y Venezuela) consta de cuatro etapas principales: búsqueda y selección de la base de datos (PAN 2017), preprocesamiento, extracción de características útiles, entrenamiento y evaluación de los modelos de aprendizaje supervisado (Bayes Ingenuo Multinomial (BIM), Árboles de Decisión (AD) y Máquinas de Vectores de Soporte (MVS)) y la clasificación de Tweets que no pertenecen a la base de datos.

Para este estudio, el primer paso fue la búsqueda y selección de una base de datos existente que cumpliera con las características necesarias para el perfilado de autor. La base

de datos seleccionada incluye una variedad representativa de textos correspondientes a diferentes géneros y variedades lingüísticas de autores provenientes de México, Perú, Argentina, España, Chile, Colombia y Venezuela. Los textos en la base de datos fueron previamente etiquetados con información relevante sobre el autor, incluyendo el género (masculino o femenino) y la variedad lingüística.

El siguiente paso es el preprocesamiento de los textos (tweets). Este proceso consistió en la eliminación de etiquetas XML y la tokenización de los documentos del conjunto de datos, con el fin de limpiar y preparar los textos para su análisis posterior. En la etapa de extracción de características, se generó la matriz de frecuencia de términos, la cual sirve como entrada para los modelos de clasificación.

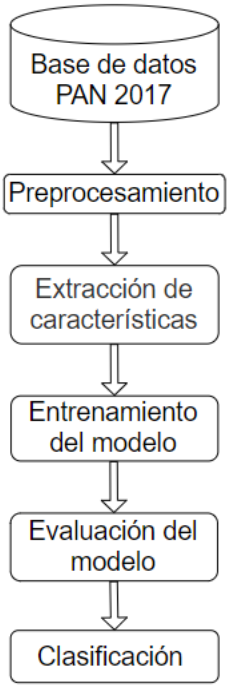

Figura 4.1: Metodología de clasificación por género y variedad lingüística.

Posteriormente, se entrenaron varios modelos de aprendizaje supervisado utilizando las características extraídas. Los modelos utilizados incluyen Bayes Ingenuo Multinomial (BIM), Árboles de Decisión (AD) y Máquinas de Vectores de Soporte (MVS). Cada uno de estos modelos fue entrenado y posteriormente evaluado en cuanto a su desempeño en

la clasificación de los autores. Se compararon los resultados para determinar cuál modelo ofrece la mayor precisión y puntuación F1 en la tarea de clasificación. Finalmente se llevaron a cabo pruebas adicionales utilizando textos que no pertenecen al conjunto de datos original. Estas pruebas permitieron evaluar la capacidad de generalización del modelo y su exactitud en escenarios reales.

4.2. Descripción y estructura del conjunto de datos

En este proyecto de titulación, empleamos el conjunto de datos PAN 2017, creado específicamente para la quinta tarea de perfilado de autor. Esta tarea se centró en identificar el género y la variedad lingüística de autores en Twitter. El conjunto de datos se estructuró de tal manera que cada combinación de género y variedad lingüística contenga la misma cantidad de tweets, totalizando 500 tweets por combinación. Cada autor en el conjunto de datos contribuye con un total de 100 tweets y 1600 palabras . Además, el conjunto de datos se divide en dos conjuntos: un conjunto de entrenamiento y un conjunto de prueba. El primero se utiliza para entrenar modelos de aprendizaje automático, mientras que el segundo se emplea para evaluar el rendimiento de estos modelos. Esta división se realiza en una proporción del 60 % para entrenamiento y del 40 % para prueba. En total, se utilizan los tweets de 300 autores para el entrenamiento de modelos, mientras que los tweets de otros 200 autores se reservan para la evaluación de estos modelos [47]. La Tabla 4.1 detalla los idiomas, las variantes lingüísticas y el número total de autores en cada subgrupo.

Tabla 4.1: Autores por cada idioma

Idioma	**Variedad lingüística**	**Autores**
Árabe	Egipto, Golfo, Levantino, Magrebí	4000
Inglés	Australia, Canadá, Gran Bretaña, Irlanda, Nueva Zelanda, Estados Unidos	6000
Español	Argentina, Chile, Colombia, México, Perú, España, Venezuela	7000
Portugués	Brasil, Portugal	2000

A continuación, la Tabla 4.2 muestra las distribuciones de datos para género y variedad lingüística del idioma Español, indicando la cantidad de usuarios y el total de tweets asociados a cada combinación de país y género en el conjunto de datos.

Tabla 4.2: Descripción de la base de datos en español

País	**Género**	**Usuarios (Entrenamiento)**	**Usuarios (Prueba)**	**Total de Tweets**
México	Femenino	300	200	500
	Masculino	300	200	500
Perú	Femenino	300	200	500
	Masculino	300	200	500
Argentina	Femenino	300	200	500
	Masculino	300	200	500
España	Femenino	300	200	500
	Masculino	300	200	500
Chile	Femenino	300	200	500
	Masculino	300	200	500
Colombia	Femenino	300	200	500
	Masculino	300	200	500
Venezuela	Femenino	300	200	500
	Masculino	300	200	500
Total		4200	2800	7000

4.2.1. Lenguaje XML

El lenguaje de marcado es una forma de representar información usando etiquetas o marcas que indican la estructura de los datos. A diferencia de los lenguajes de programación, los lenguajes de marcado no incluyen instrucciones ni acciones directas. Más bien, se componen de reglas que organizan la información para darle una estructura coherente y facilitar su procesamiento automático. En resumen, los lenguajes de marcado se refieren a la información añadida a los datos de un documento mediante marcas. [58].

La variedad de lenguajes de marcado conlleva una diversidad de estructuras y formatos, cada lenguaje posee su propia organización y composición, esto quiere decir que no existe un formato estándar o una estructura universal [58].

El avance del desarrollo web ha llevado a un notable incremento en la utilización de estos lenguajes, ya que son cruciales en varias áreas, como la configuración de aplicaciones y la creación de interfaces gráficas de usuario. Los lenguajes de marcado más frecuentes tienen como objetivo principal la representación de datos, como en el caso de HTML, o el almacenamiento e intercambio de información, como ocurre con XML. [59].

El Lenguaje de Marcado Extensible (XML) actúa como un metalenguaje, lo que implica que tiene la capacidad de definir normas para marcar diversas secciones de un texto, tales como palabras, frases, cifras o ecuaciones. Por ejemplo, en un documento, se pueden señalar componentes bibliográficos como el título, autor, secciones, palabras clave, tablas, ilustraciones, citas y referencias. En lugar de centrarse en el formato del texto, como el tamaño o el tipo de letra, este lenguaje se enfoca en el contenido del documento, facilitando la etiquetación precisa de la información. [48].

La esencia de que un sistema automatizado pueda identificar qué documento podría ser beneficioso para un usuario radica en enseñarle a buscar artículos particulares que incluyan etiquetas determinadas. Estas etiquetas orientan a las bases de datos sobre la ubicación de diferentes secciones del contenido: el resumen en un área y la referencia bibliográfica en otra. Este tipo de datos está diseñado para ser interpretado por las máquinas que examinan y manejan la información, no por los lectores. [48]

Un archivo XML se encarga de organizar la información que contiene. Posee uno o varios elementos que inician con una etiqueta de apertura, la cual consiste en un nombre envuelto entre los símbolos <y >. Los datos situados entre estas etiquetas son comprensibles para los humanos. [49] Los elementos concluyen con una etiqueta de cierre que utiliza el mismo nombre que la etiqueta de apertura, pero rodeado por </ y >. [48]

En el siguiente ejemplo se puede observar el lenguaje de marcado XML y la estructura de su información en base a las etiquetas correspondientes:

```
<?xml versión “1.0”?>
<!DOCTYPE MENSAJE SYSTEM “mensaje.dtd”>
<mensaje>
<remitente>
    <nombre> Lizbeth Córdova </nombre>
    <email>lizcor1701@gmail.com</email>
</remitente>
<destinatario>
   <nombre> Fernando Murillo</nombre>
   <email>jfermurillo@gmail.com</email>
</destinatario>
<asunto>Posada 2024</asunto>
<texto>
 <parrafo>
    Hola Fer, ¿podrías confirmarme tu
    asistencia a la posada antes del fin de semana?
 </parrafo>
</texto>
</mensaje>
```

Figura 4.2: Código xml.

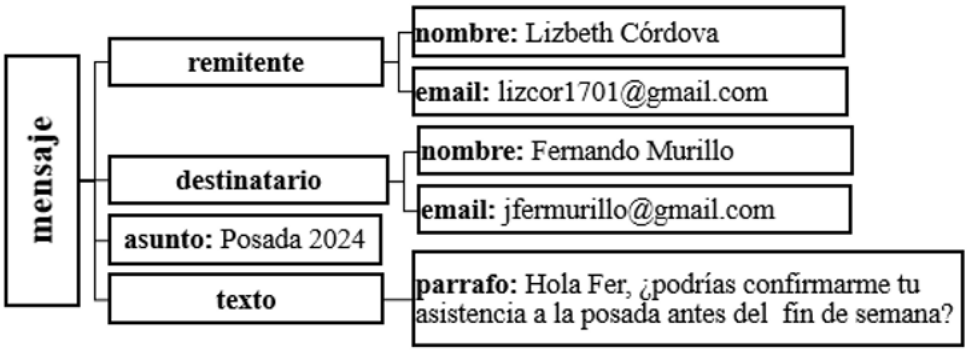

Figura 4.3: Estructura del código xml.

4.2.2. Etiquetas CDATA

Una etiqueta CDATA (Datos de Carácter) con formato <![CDATA[datos]]>, permite especificar datos utilizando cualquier tipo de carácter, sea especial o no. Dentro de una sección CDATA podemos poner cualquier cosa, sin embargo, existe una excepción, la cadena con la que termina el bloque CDATA: "]] >". Esta cadena no puede utilizarse dentro de las secciones CDATA [49].

Los bloques CDATA le indica al procesador que ignore su contenido y lo considere como información externa al XML, lo cual resulta útil cuando se desea incluir código fuente de algún lenguaje de programación en el documento. Este código puede contener caracteres que un analizador XML interpretaría como parte del marcado, como < o & [50].

Si se desea extraer el contenido de una etiqueta CDATA en un archivo XML, se puede lograr procesando el archivo XML mediante la implementación de una biblioteca o herramienta que permita la manipulación de documentos XML.

4.2.3. Formato de la base de datos

Los tweets emitidos por cada usuario se guardan en un archivo XML. Estos archivos llevan el nombre de la identificación del usuario seguido de la extensión ".xml". La disposición de los datos en cada archivo XML se visualiza en la [Figura]. Cada elemento dentro de la jerarquía de documentos corresponde a un tweet que ha sido compartido por el usuario.

```
<author lang="es">
 <documents>
  <document><![CDATA[Solo 1 juego Roger C'MON]]></document>
  <document><![CDATA[El primer set para #sumajestad CMON!]]></document>
      .
      .
      .
      .
 </documents>
</author>
```

Figura 4.4: Ejemplo de archivo XML con tweets en español.

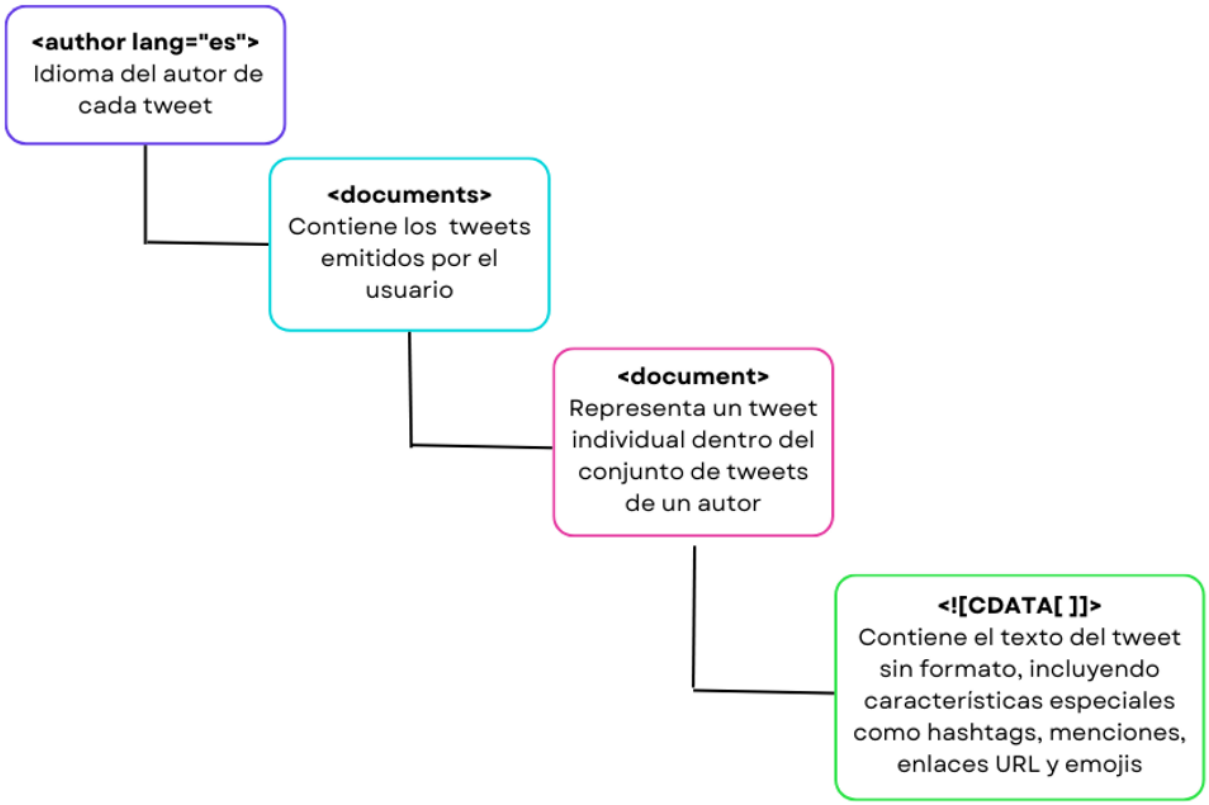

Figura 4.5: Estructura del archivo XML del conjunto de datos.

4.3. Preprocesamiento de texto

El preprocesamiento de texto es una técnica utilizada en la conversión de datos textuales inconsistentes en un formato coherente y comprensible para sistemas de análisis y algoritmos de aprendizaje automático. Este proceso implica una serie de operaciones que transforman los datos brutos en datos estructurados y limpios, listos para ser utilizados en análisis posteriores y modelado predictivo [52].

En este proyecto, se aplicaron técnicas de preprocesamiento de texto, que incluyen:

- **Conversión a minúsculas:** Este procedimiento evita la consideración de diversas variaciones de una misma palabra, como aquellas que difieren únicamente en la capitalización , las cuales pueden generar resultados distintos. Esta diferencia implica que palabras idénticas pero escritas con diferente capitalización se interpretan como entidades separadas. Por ejemplo, después de la conversión a minúsculas, tanto *India* como *india* se consideran la misma palabra.
- **Eliminación de puntuación:** La puntuación, como comas, puntos, signos de interrogación, etc., puede introducir ruido en el texto, especialmente en documentos no estructurados como artículos, comentarios o redes sociales.
- **Eliminación de palabras vacías:** Se refiere a términos que carecen de un significado específico para el análisis texto y, por lo tanto, se eliminan. Estas palabras suelen ser funcionales y gramaticales, como conectores (y, o), preposiciones (de, en), artículos (el, la) y pronombres (él, ella). Al remover estas palabras, se extraen términos que definen el contenido
- **Tokenización:** Es el proceso de fragmentar una secuencia de texto en atributos únicos, tales como palabras, frases o símbolos, denominados tokens, que sirven como características para el análisis. [52].

A continuación se propone un ejemplo utilizando un archivo XML del conjunto de datos, que contiene los tweets de un usuario .

```
<author lang="es">
    <documents>
        <document><![CDATA[Maduro: Las redes sociales promueven la violencia y el culto a las armas]]></document>
        <document><![CDATA[El sueldo de muchos, en dos billetes.]]></document>
        <document><![CDATA[El billete de 100 tiene mas vidas que un gato.]]></document>
    </documents>
</author>
```

Figura 4.6: Tweets de un usuario.

Tabla 4.3: Preprocesamiento de tweets de un usuario.

Etapa	**Texto inicial**	**Texto después del preprocesamiento**
Conversión a minúsculas	Maduro: Las redes sociales promueven la violencia y el culto a las armas. El sueldo de muchos, en dos billetes. El billete de 100 tiene más vidas que un gato.	maduro: las redes sociales promueven la violencia y el culto a las armas. el sueldo de muchos, en dos billetes. el billete de 100 tiene más vidas que un gato.
Eliminación de puntuación	maduro: las redes sociales promueven la violencia y el culto a las armas. el sueldo de muchos, en dos billetes. el billete de 100 tiene más vidas que un gato.	maduro las redes sociales promueven la violencia y el culto a las armas el sueldo de muchos en dos billetes el billete de tiene más vidas que un gato
Eliminación de palabras vacías	maduro las redes sociales promueven la violencia y el culto a las armas el sueldo de muchos en dos billetes el billete de 100 tiene más vidas que un gato	maduro redes sociales promueven violencia culto armas sueldo billetes billete 100 tiene vidas gato
Tokenización	maduro redes sociales promueven violencia culto armas sueldo billetes billete 100 tiene vidas gato	‘maduro’, ‘redes’, ‘sociales’, ‘promueven’, ‘violencia’, ‘culto’, ‘armas’, ‘sueldo’, ‘billetes’, ‘billete’, ‘100’, ‘tiene’, ‘vidas’, ‘gato’

4.4. Extracción de características

Los algoritmos de aprendizaje automático (ML) utilizados en la construcción de modelos de clasificación tienen limitaciones para procesar textos en su formato original. Por lo que, se requiere la implementación de un procedimiento de indexación para convertir el contenido textual en vectores de atributos numéricos con dimensiones constantes. Este procedimiento debe aplicarse al conjunto de datos de entrenamiento, validación y prueba [53].

4.4.1. Modelo de bolsa de palabras

El esquema de representación de documentos empleado en este trabajo se fundamenta en el modelo de bolsa de palabras (BoW), una técnica aplicada en el análisis de texto que traduce la aparición de palabras en un documento en un vector de características. Cada vector posee un componente para cada término presente en el vocabulario del corpus. Este enfoque descarta la estructura sintáctica y gramatical del texto, enfocándose exclusivamente en la existencia de términos individuales. [56].

El modelo de bolsa de palabras es una técnica para extraer características de un texto. La idea básica detrás de este método es contar cuántas veces aparece cada palabra en un documento y luego utilizar estos recuentos como características para representar el texto. Cada documento se proyecta como un vector en un espacio dimensional, donde cada componente del vector representa la frecuencia de aparición de un término específico dentro del corpus. [58]

El método de bolsa de términos lleva a cabo una vectorización a través del conteo. Esta técnica se fundamenta en representar un texto mediante el número de veces que aparecen palabras individuales en un documento específico. Los conteos obtenidos se utilizan luego para medir la similitud entre documentos en diferentes contextos, como la búsqueda de información y la categorización de textos. [59]

La representación de documentos en el enfoque de bolsa de palabras se efectúa a través de una matriz de términos en la que la secuencia de las palabras no es relevante. Las filas corresponden a los documentos (número de tweets por usuario) y las columnas a

las palabras o términos del vocabulario. El valor de cada celda en la matriz refleja la cantidad de veces que cada palabra se presenta en cada documento. [59].

Tabla 4.4: Matriz de términos.

Documentos	**Término 1**	**Término 2**	**Término 3**	**Término 4**	**Término N**
Documento 1	1	1	1	2	0
Documento 2	0	3	0	1	1
Documento 3	1	0	0	0	1

4.4.2. Modelo de espacio vectorial

En el modelo de espacio vectorial los objetos son representados por sus características y los valores asociados a estas. El espacio en el que se representan los objetos tiene N dimensiones, donde cada dimensión representa una característica del objeto. El número de dimensiones en en el espacio es igual al número de características en el modelo. Una dimensión es un eje que permite marcar los valores de una característica especifica del objeto. [54]

En este modelo, cada objeto se representa como un vector en un espacio de N dimensiones, donde N es el número de características distintivas que definen el objeto. Cada dimensión del espacio corresponde a una característica específica y, por lo tanto, los vectores representan las propiedades y atributos de los objetos en términos numéricos [54].

Cuando se realiza la representación de un objeto en este espacio, se inicia su vector en el origen, donde todas las coordenadas son nulas, y luego se extiende en cada dimensión según el valor asociado a cada característica. Este proceso de asignación de valores a las dimensiones genera una ubicación única en el espacio vectorial que caracteriza al objeto en cuestión. El orden en que se presentan las características no altera la representación del objeto, dado que todas las dimensiones son consideradas de manera equiparable y el resultado final es independiente de la secuencia en que se definen las características [54].

4.4.3. Representación de documentos en el modelo de espacio vectorial

El modelo de espacio vectorial transforma datos textuales en vectores numéricos dentro de un espacio de múltiples dimensiones. Cada dimensión del vector representa distintos aspectos semánticos, sintácticos o estadísticos del texto original [54].

Con los vectores numéricos es posible realizar operaciones matemáticas, tales como la comparación o la agrupación de vectores según métricas como la distancia euclidiana y el coseno [56].

Estas técnicas permiten llevar a cabo tareas, como la recuperación de información, detección de similitudes entre documentos, agrupación de textos relacionados, clasificación de documentos en categorías específicas, minería de datos y análisis predictivo [55].

La representación de un texto como un vector de n dimensiones en el modelo de espacio vectorial es:

$$v_d = (w_{0,j}, w_{1,j}, \ldots, w_{n,j})$$

Donde n corresponde al número de características representativas del texto y w es el valor asociado a dicha característica.

4.5. Entrenamiento y evaluación del modelo

Una vez que se obtiene la representación vectorial de los documentos y se crea el vocabulario o diccionario de tokens, en nuestro caso palabras únicas. El siguiente paso es entrenar a los modelos de aprendizaje supervisado: Máquinas de Soporte Vectorial, Bayes Ingenuo Multinomial y Árboles de Decisión.

La Figura 4.7 se presenta el proceso de entrenamiento y prueba.

El modelo se entrenó con 3 métodos supervisados: Máquina de Soporte Vectorial, Árboles de decisión y Multinomial Naive Bayes. Posteriormente se implementaron 3 métricas de evaluación para reportar la precisión de los métodos en cada tarea.

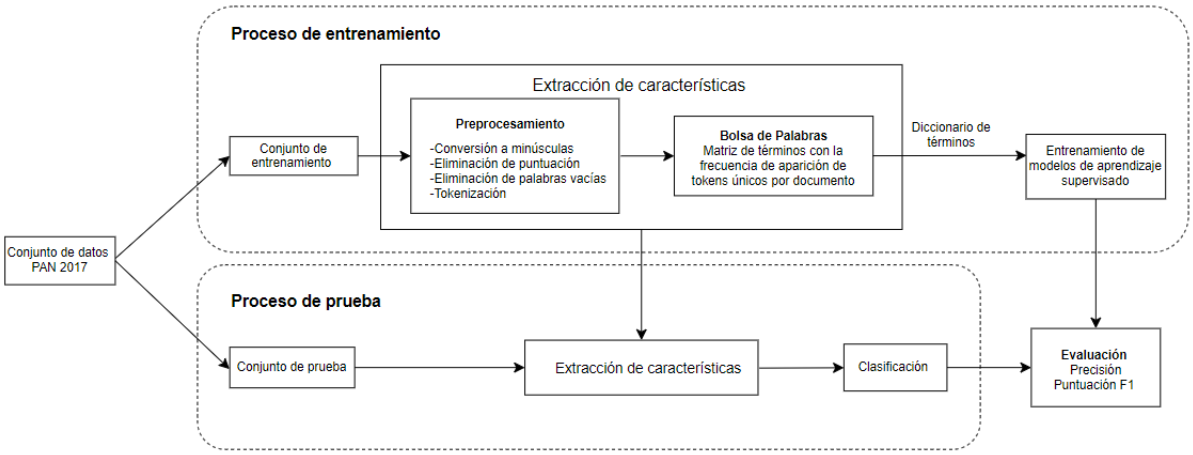

Figura 4.7: Entrenamiento, prueba y evaluación de los modelos de aprendizaje supervisado.

Capítulo 5

Experimentos y resultados

En este capítulo se presenta la descripción y resultados de los experimentos realizados con la base de datos de prueba de PAN at Clef 2017 y posteriormente, con tweets directamente extraídos del perfil de famosos de diversos países de habla hispana. Para la realización de los experimentos, se creó la siguiente interfaz gráfica en Python (Figura 5.1).

Un experimento consiste en ingresar los tweets en el espacio en blanco, ya sea con un formato libre o con un formato estructurado como el de nuestra base de datos (lenguaje xml con etiquetas cdata), al presionar el botón se obtiene la predicción del género y país de cada autor.

5.1. Evaluación de los modelos de clasificación

El presente análisis constituye una evaluación global del rendimiento de tres modelos de aprendizaje supervisado: Máquinas de Vectores de Soporte (SVM), Bayes Ingenuo Multinomial (MNB) y Árboles de Decisión (DT), aplicados en la clasificación por género y variedad lingüística del español. Utilizamos un conjunto de datos de prueba del PAN 2017 compuesto por 2800 muestras etiquetadas. Estas se distribuyen equitativamente entre género masculino y femenino, con 1400 muestras para cada categoría. Además, se dividen en muestras etiquetadas por variedad lingüística del español, con

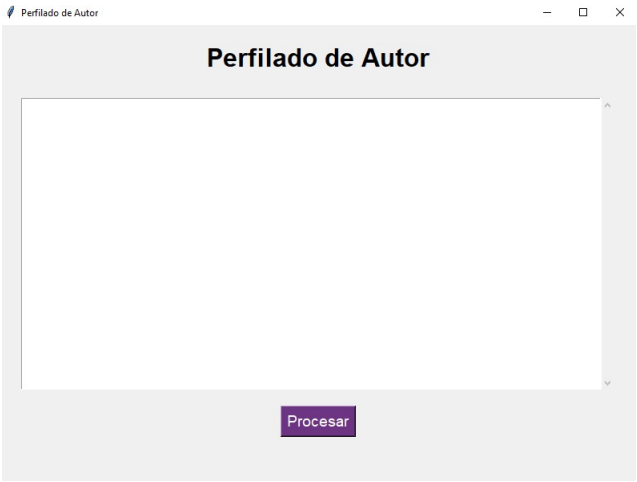

Figura 5.1: Interfaz gráfica

400 muestras para cada país de habla hispana, Argentina, Chile, Colombia, España, Venezuela, México y Perú. Cada modelo se evaluó utilizando dos métricas estándar en aprendizaje automático: precisión global (accuracy) y puntuación F1 (F1-score). En la clasificación por género, el modelo SVM destacó con una precisión y F1-score de 0.7689. En contraste, en la clasificación por variedad lingüística, el modelo MNB sobresalió significativamente con una precisión de 0.9114 y una puntuación F1 de 0.9115. La Tabla 5.1 muestra los resultados específicos de cada modelo en la tarea de clasificación por género, mientras que la Tabla 5.2 proporciona los resultados para la variedad lingüística.

Tabla 5.1: Evaluación de la clasificación por género

	Accuracy	**F1**
Máquina de Soporte Vectorial	0.7689	0.7689
Multinomial Naive Bayes	0.7192	0.7192
Arboles de decisión	0.6532	0.6531

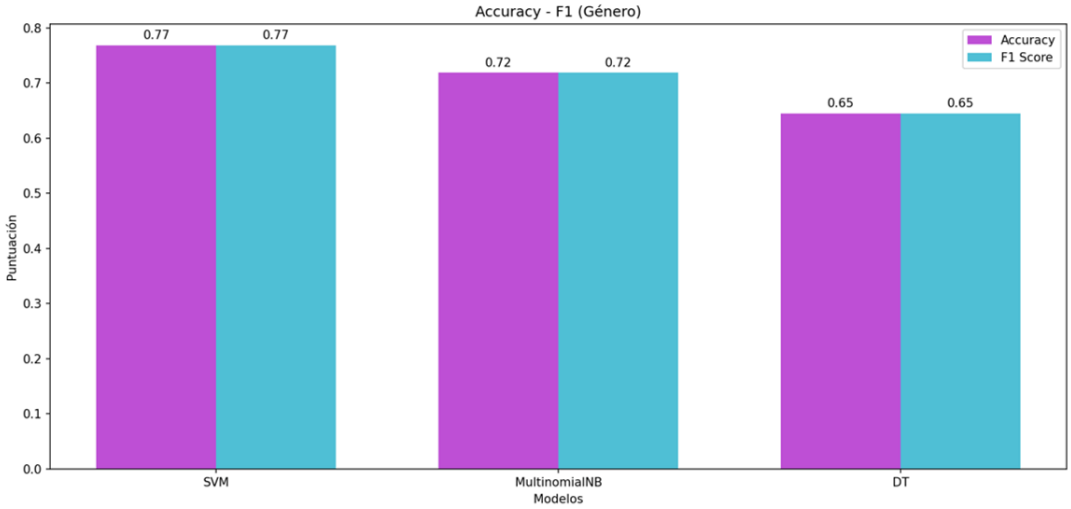

Figura 5.2: Exactitud y puntuación F1 para clasificación por género.

Tabla 5.2: Evaluación de la clasificación por país

	Accuracy	**F1**
Máquina de Soporte Vectorial	0.9085	0.9085
Multinomial Naive Bayes	0.9114	0.9114
Arboles de decisión	0.7810	0.7810

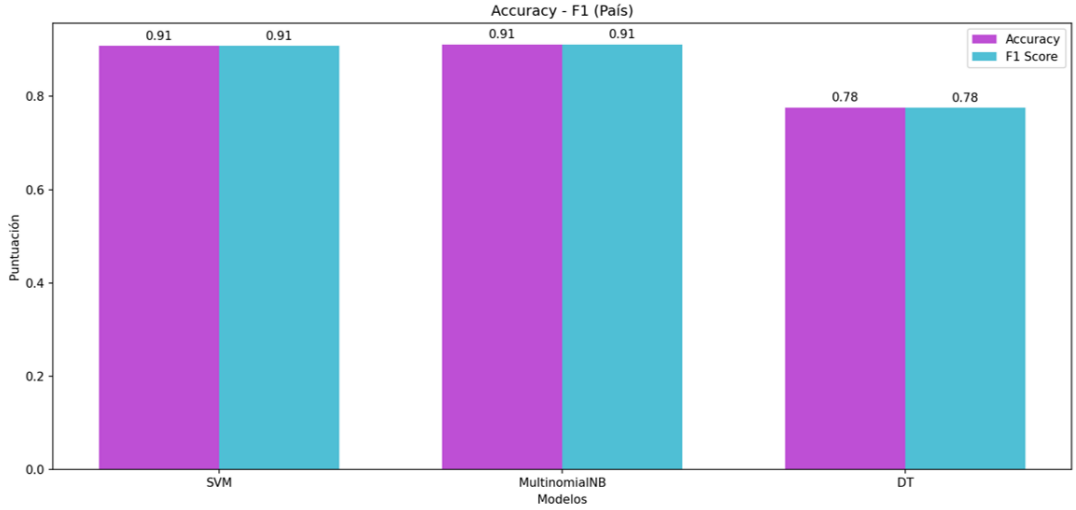

Figura 5.3: Exactitud y puntuación F1 para clasificación por país.

5.2. Reporte de clasificación

Esta sección presenta los reportes de clasificación de los modelos evaluados. Se han utilizado métricas como precisión, recall y F1-score para evaluar la capacidad de los modelos en la predicción de categorías específicas, tales como género y variedad lingüística. Los resultados detallan el rendimiento de cada modelo en la clasificación de estas clases individuales, presentando tablas con el recuento de muestras por cada categoría evaluada, acompañadas de las matrices de confusión.

5.2.1. Reporte de clasificación por género

Los resultados del rendimiento del clasificador SVM en la clasificación de género por clase indican que la precisión es del 77.1 % para el género femenino y del 76.7 % para el género masculino. En términos de recall, se obtuvo un 76.5 % para el género femenino y un 77.3 % para el género masculino, lo cual señala la capacidad del modelo para identificar correctamente las instancias positivas reales de cada género. En cuanto a los valores de F1-score, estos son del 76.8 % para el género femenino y del 77.0 % para el género masculino. Estos resultados reflejan un buen equilibrio entre la precisión en las predicciones positivas y la capacidad del modelo para identificar todas las instancias positivas reales en la clasificación de género.

En la Tabla 5.3 se presentan los resultados de las métricas de evaluación del método SVM en la predicción de género para cada clase.

Tabla 5.3: Resultados de SVM para la clasificación por género

Clase	Precisión	Recall	F1-score	No. muestras
Femenino	0.7710	0.765	0.7680	1400
Masculino	0.7668	0.7729	0.7698	1400

En la Figura 5.4 se presenta la matriz de confusión del clasificador SVM. Los resultados se desglosan de la siguiente manera:

– Verdaderos Negativos (N-N): El modelo acertó en 1071 casos al clasificar correctamente instancias masculinas como masculinas.

- Falsos Negativos (P-N): Se observaron 318 ocasiones en las que el modelo clasificó incorrectamente instancias femeninas como masculinas.
- Falsos Positivos (N-P): Con un total de 329 casos, el modelo cometió errores al clasificar instancias masculinas incorrectamente como femeninas.
- Verdaderos Positivos (P-P): Acertó en 1082 casos al clasificar correctamente instancias femeninas como femeninas.

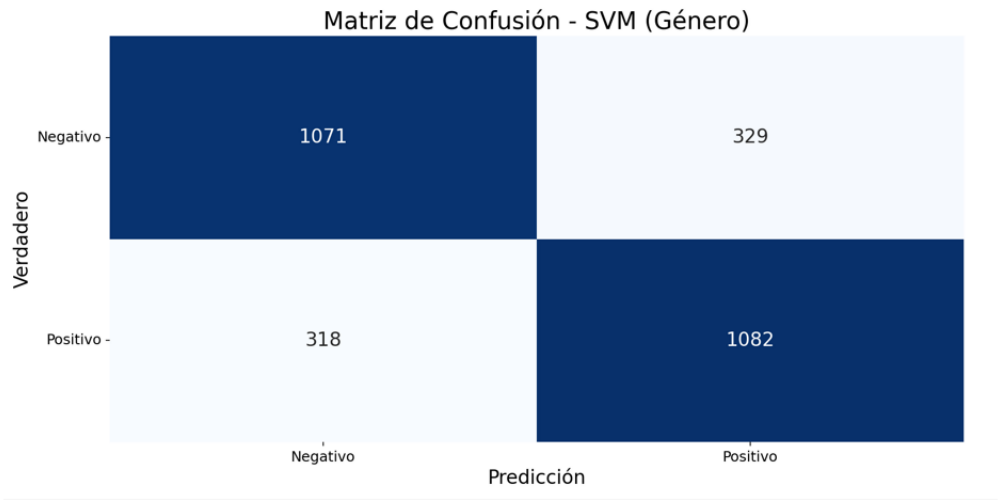

Figura 5.4: Matriz de confusión SVM para la clasificación por género.

Los resultados del rendimiento del clasificador Multinomial Naive Bayes (MNB) en la clasificación de género por clase indican que la precisión es del 71.87 % para el género femenino y del 71.99 % para el género masculino. En términos de recall, se obtuvo un 72.07 % para el género femenino y un 71.79 % para el género masculino. En cuanto a los valores de F1-score, estos son del 71.97 % para el género femenino y del 71.89 % para el género masculino.

En la Tabla 5.4 se presentan los resultados de las métricas de evaluación del método MultinomialNB en la predicción de género para cada clase.

En la Figura 5.5 se presenta la matriz de confusión del clasificador MultinomialNB. Los resultados se desglosan de la siguiente manera:

Tabla 5.4: Resultados de MultinomialNB para la clasificación por género

Clase	Precisión	Recall	F1-score	No. muestras
Femenino	0.7187	0.7207	0.7197	1400
Masculino	0.7199	0.7179	0.7189	1400

– Verdaderos Negativos (N-N): El modelo acertó en 1009 casos al clasificar correctamente instancias masculinas como masculinas.

– Falsos Negativos (P-N): Se observaron 395 ocasiones en las que el modelo clasificó incorrectamente instancias femeninas como masculinas.

– Falsos Positivos (N-P): Con un total de 391 casos, el modelo cometió errores al clasificar instancias masculinas incorrectamente como femeninas.

– Verdaderos Positivos (P-P): Acertó en 1005 casos al clasificar correctamente instancias femeninas como femeninas.

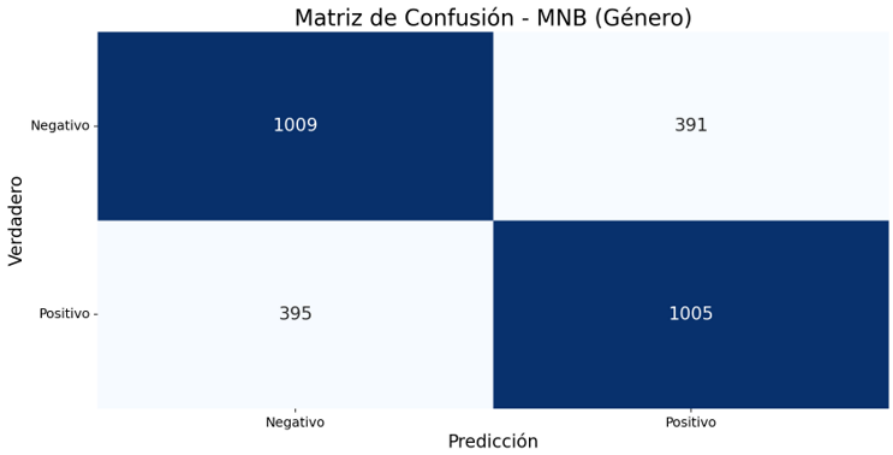

Figura 5.5: Matriz de confusión MultinomialNB para la clasificación por género.

Los resultados del rendimiento del clasificador Árboles de Decisión en la clasificación de género por clase indican que la precisión es del 64.44 % para el género femenino y del 64.63 % para el género masculino. En términos de recall, se obtuvo un 64.86 % para el género femenino y un 64.21 % para el género masculino. En cuanto a los valores de

F1-score, estos son del 64.65 % para el género femenino y del 64.42 % para el género masculino.

En la Tabla 5.5 se presentan los resultados de las métricas de evaluación del método de Árboles de decisión en la predicción de género para cada clase.

Tabla 5.5: Resultados de Árboles de decisión para la clasificación por género

Clase	**Precisión**	**Recall**	**F1-score**	**No. muestras**
Femenino	0.6444	0.6486	0.6465	1400
Masculino	0.6463	0.6421	0.6442	1400

En la Figura 5.6 se presenta la matriz de confusión del clasificador Árboles de Decisión para la clasificación por género. Los resultados se desglosan de la siguiente manera:

- Verdaderos Negativos (N-N): El modelo acertó en 908 casos al clasificar correctamente instancias masculinas como masculinas.
- Falsos Negativos (P-N): Se observaron 501 ocasiones en las que el modelo clasificó incorrectamente instancias femeninas como masculinas.
- Falsos Positivos (N-P): Con un total de 492 casos, el modelo cometió errores al clasificar instancias masculinas incorrectamente como femeninas.
- Verdaderos Positivos (P-P): Acertó en 899 casos al clasificar correctamente instancias femeninas como femeninas.

5.2.2. Reporte de clasificación para país

Los resultados del rendimiento del clasificador SVM en la clasificación de país por clase indican que, Argentina tiene la mayor precisión con un 93.48 %, seguida de Chile con un 93.11 % y España con un 92.25 %. Por otro lado, México tiene la menor precisión con un 85.95 %, lo que indica que se tiene mayor dificultad para distinguir registros de México en comparación con los de otros países. España presenta el recall más alto con un 95.25 %, lo que significa que el modelo es más efectivo en identificar correctamente los registros de España. En contraste, Venezuela tiene el recall más bajo con un 85.00 %, lo que sugiere que el modelo tiende a no identificar correctamente una mayor cantidad

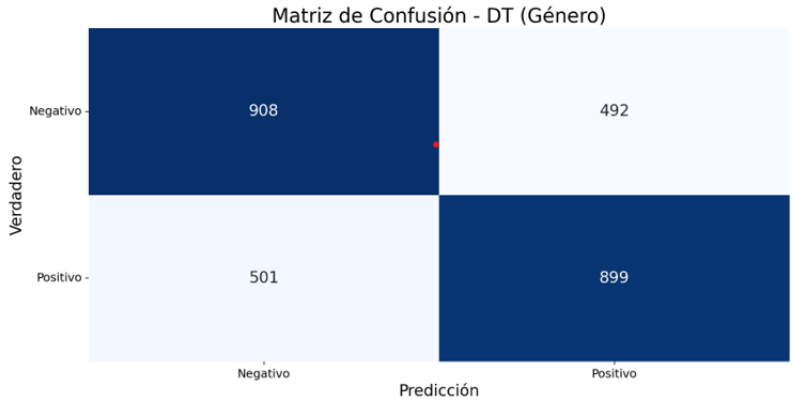

Figura 5.6: Matriz de confusión Árboles de Decisión para la clasificación por género.

de registros de Venezuela. España también se destaca con el F1-score más alto de 93.73 %. Por otro lado, México tiene el F1-score más bajo con un 88.75 %. En la Tabla 5.6 se presentan los resultados de las métricas de evaluación del método SVM en la predicción de país para cada clase.

Tabla 5.6: Clasificación por país SVM

Clase	**Precisión**	**Recall**	**F1**	**No.muestras**
Argentina	0.9348	0.9325	0.9337	400
Chile	0.9311	0.9125	0.9217	400
Colombia	0.9139	0.9025	0.9082	400
México	0.8595	0.9175	0.8875	400
Perú	0.8837	0.8925	0.8881	400
España	0.9225	0.9525	0.9373	400
Venezuela	0.9189	0.8500	0.8831	400

En la Figura 5.7 se presenta la matriz de confusión del clasificador SVM para la clasificación por país. A continuación, se detallan los resultados:

– **Argentina (Clase 0):** el modelo acertó en 373 instancias (verdaderos positivos), mientras que cometió errores en 27 ocasiones (falsos negativos), clasificando incorrectamente estas instancias como pertenecientes a otros países (7 como Chile, 1 como Colombia, 5 como México, 7 como Perú, 2 como España y 5 como Venezuela).

- **Chile (Clase 1):** se identificaron correctamente 365 instancias (verdaderos positivos) y se cometieron 35 errores (falsos negativos), principalmente confundiéndolas con Perú y México (8 como México, 10 como Perú, 4 como España y 7 como Venezuela).
- **Colombia (Clase 2):** se registraron 361 verdaderos positivos y 39 falsos negativos, con errores notables principalmente hacia Perú y México (9 como México, 14 como Perú, 3 como España y 6 como Venezuela).
- **México (Clase 3):** tiene 367 verdaderos positivos y 33 falsos negativos, con errores frecuentes hacia Colombia y Perú (8 como Colombia, 7 como Perú, 4 como España y 3 como Venezuela).
- **Perú (Clase 4):** se identificaron 357 verdaderos positivos y 43 falsos negativos, con errores hacia México y Venezuela (11 como México, 6 como España y 7 como Venezuela).
- **España (Clase 5):** se observaron 381 verdaderos positivos y 19 falsos negativos, con errores hacia México, Colombia y Venezuela (9 como México, 5 como Colombia y 2 como Venezuela).
- **Venezuela (Clase 6):** presenta 340 verdaderos positivos y 60 falsos negativos, con confusiones frecuentes con México y Perú (18 como México, 9 como Colombia y 13 como Perú).

Los resultados del rendimiento del clasificador Multinomial Naive Bayes (MNB) en la clasificación de país por clase indican que, Chile tiene la mayor precisión con un 98.14 %, seguida de Venezuela con un 93.33 % y Perú con un 93.02 %. Por otro lado, Colombia tiene la menor precisión con un 81.32 %, lo que indica que se tiene mayor dificultad para distinguir registros de Colombia en comparación con los de otros países. Colombia presenta el recall más alto con un 95.75 %, lo que significa que el modelo es más efectivo en identificar correctamente los registros de Colombia. En contraste, Venezuela tiene el recall más bajo con un 80.50 %, lo que sugiere que el modelo tiende a no identificar correctamente una mayor cantidad de registros de Venezuela. Chile también se destaca con el F1-score más alto de 95.10 %. Por otro lado, Colombia tiene el F1-score más bajo con un 87.94 %. En la Tabla 5.7 se presentan los resultados de las

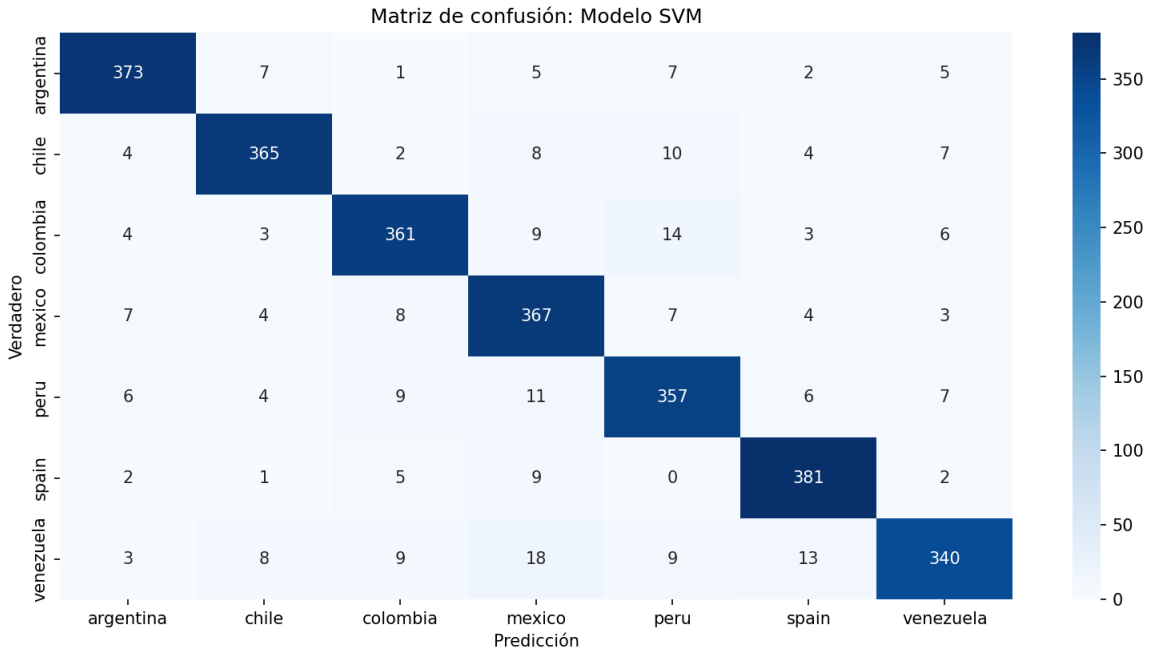

Figura 5.7: Matriz de confusión SVM para la clasificación por país

métricas de evaluación del método Multinomial Naive Bayes en la predicción de país para cada clase.

Tabla 5.7: Clasificación por país Multinomial Naive Bayes

Clase	**Precisión**	**Recall**	**F1**	**No.muestras**
Argentina	0.9136	0.9250	0.9193	400
Chile	0.9814	0.9225	0.9510	400
Colombia	0.8132	0.9575	0.8794	400
México	0.9039	0.9175	0.9107	400
Perú	0.9302	0.9000	0.9149	400
España	0.9293	0.9525	0.9407	400
Venezuela	0.9333	0.8050	0.8644	400

En la Figura 5.8 se presenta la matriz de confusión del clasificador Multinomial Naive Bayes para la clasificación por país. A continuación, se detallan los resultados:

– **Argentina (Clase 0):** el modelo acertó en 370 instancias (verdaderos positivos), mientras que cometió errores en 30 ocasiones (falsos negativos), clasificando

incorrectamente estas instancias como pertenecientes a otros países (3 como Chile, 4 como Colombia, 4 como México, 7 como Perú, 5 como España y 7 como Venezuela).

- **Chile (Clase 1):** se identificaron correctamente 369 instancias (verdaderos positivos) y se cometieron 31 errores (falsos negativos), principalmente confundiéndolas con Perú y México (6 como México, 10 como Perú, 3 como España y 5 como Venezuela).
- **Colombia (Clase 2):** se registraron 383 verdaderos positivos y 17 falsos negativos, con errores notables principalmente hacia México y Venezuela (3 como México, 1 como Venezuela).
- **México (Clase 3):** tiene 367 verdaderos positivos y 33 falsos negativos, con errores frecuentes hacia Perú y Venezuela (14 como Perú, 4 como España y 8 como Venezuela).
- **Perú (Clase 4):** se identificaron 360 verdaderos positivos y 40 falsos negativos, con errores hacia Chile y Venezuela (13 como Chile, 3 como España y 4 como Venezuela).
- **España (Clase 5):** se observaron 381 verdaderos positivos y 19 falsos negativos, con errores hacia Chile, Colombia y Venezuela (9 como Chile, 2 como Colombia y 4 como Venezuela).
- **Venezuela (Clase 6):** presenta 322 verdaderos positivos y 78 falsos negativos, con confusiones frecuentes con Colombia y México (53 como Colombia, 9 como México y 9 como Perú).

Los resultados del rendimiento del clasificador de Árboles de Decisión (Decision Tree) en la clasificación de país por clase indican que, Argentina tiene la mayor precisión con un 81.30 %, seguida de Chile con un 84.78 % y Venezuela con un 77.47 %. Por otro lado, Colombia tiene la menor precisión con un 73.28 %, lo que indica que se tiene mayor dificultad para distinguir registros de Colombia en comparación con los de otros países. Argentina presenta el recall más alto con un 84.75 %, lo que significa que el modelo es más efectivo en identificar correctamente los registros de Argentina. En contraste, Colombia tiene el recall más bajo con un 72.00 %, lo que sugiere que el modelo tiende

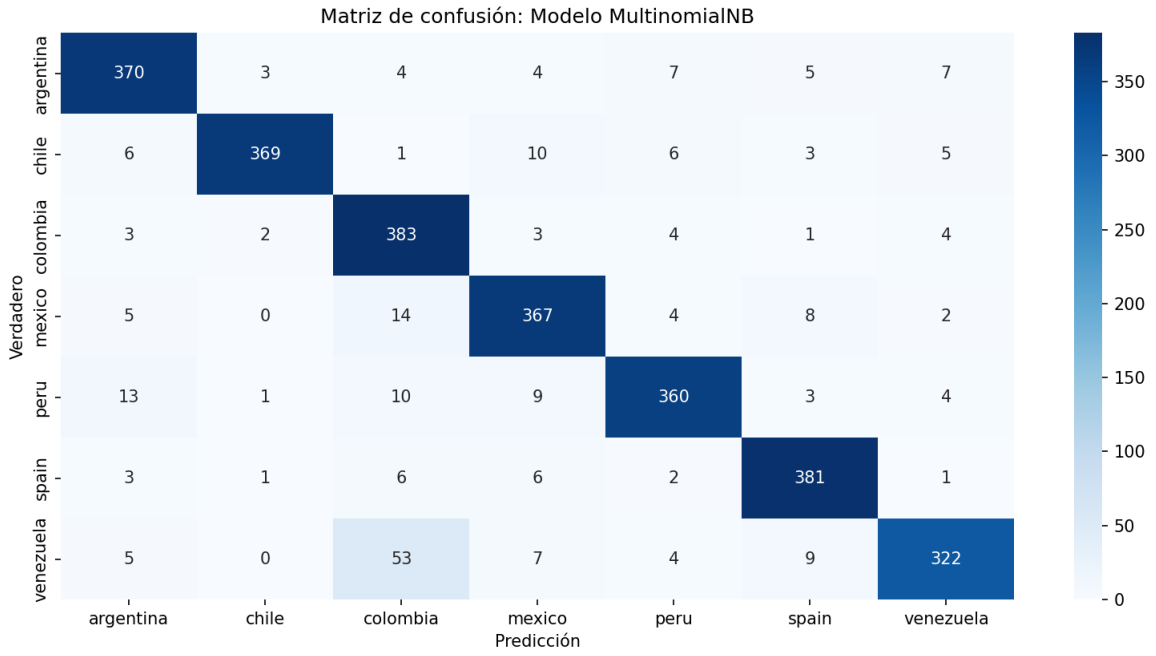

Figura 5.8: Matriz de confusión Multinomial Naive Bayes para la clasificación por país

a no identificar correctamente una mayor cantidad de registros de Colombia. Argentina también se destaca con el F1-score más alto de 82.99 %. Por otro lado, Colombia tiene el F1-score más bajo con un 72.64 %.

En la Tabla 5.8 se presentan los resultados de las métricas de evaluación del método de Árboles de Decisión para la tarea de predicción de país.

Tabla 5.8: Clasificación por país Decision Tree

Clase	**Precisión**	**Recall**	**F1**	**No.muestras**
Argentina	0.8130	0.8475	0.8299	400
Chile	0.8478	0.7800	0.8125	400
Colombia	0.7328	0.7200	0.7264	400
México	0.7220	0.8050	0.7612	400
Perú	0.8101	0.7250	0.7652	400
España	0.7441	0.7125	0.7280	400
Venezuela	0.7747	0.8425	0.8072	400

En la Figura 5.9 se presenta la matriz de confusión del clasificador Decision Tree para la clasificación por país. A continuación, se detallan los resultados:

- **Argentina (Clase 0):** el modelo acertó en 339 instancias (verdaderos positivos), mientras que cometió errores en 61 ocasiones (falsos negativos), clasificando incorrectamente estas instancias como pertenecientes a otros países (7 como Chile, 16 como Colombia, 12 como México, 9 como Perú, 11 como España y 6 como Venezuela).
- **Chile (Clase 1):** se identificaron correctamente 312 instancias (verdaderos positivos) y se cometieron 88 errores (falsos negativos), principalmente confundiéndolas con Perú y México (17 como México, 10 como Perú, 12 como España y 14 como Venezuela).
- **Colombia (Clase 2):** se registraron 288 verdaderos positivos y 112 falsos negativos, con errores notables principalmente hacia México y Venezuela (21 como México, 27 como Perú, 18 como España y 31 como Venezuela).
- **México (Clase 3):** tiene 322 verdaderos positivos y 78 falsos negativos, con errores frecuentes hacia Colombia y Perú (12 como Colombia, 24 como Perú, 20 como España y 15 como Venezuela).
- **Perú (Clase 4):** se identificaron 290 verdaderos positivos y 110 falsos negativos, con errores hacia Chile y Venezuela (13 como Chile, 8 como España y 14 como Venezuela).
- **España (Clase 5):** se observaron 285 verdaderos positivos y 115 falsos negativos, con errores hacia México, Colombia y Venezuela (14 como México, 14 como Colombia y 18 como Venezuela).
- **Venezuela (Clase 6):** presenta 337 verdaderos positivos y 63 falsos negativos, con confusiones frecuentes con Colombia y México (5 como Colombia, 9 como México y 15 como Perú).

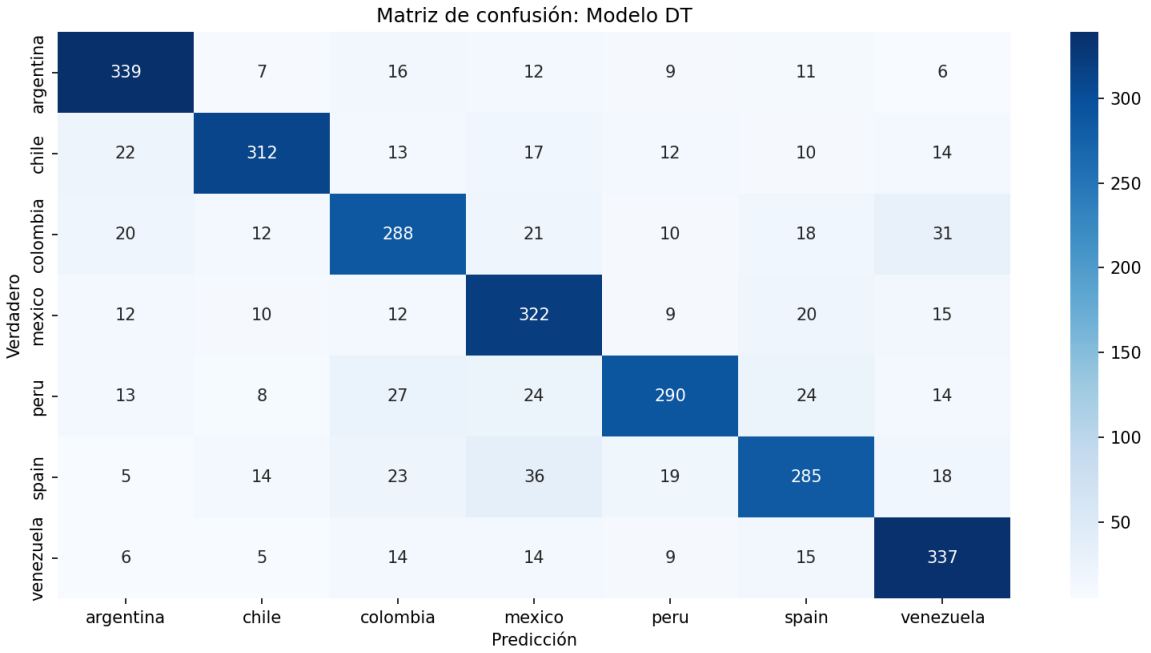

Figura 5.9: Matriz de confusión Decision Tree para la clasificación por país

5.3. Clasificación de tweets

Para la base de datos de estos experimentos, se visitó el perfil de cantantes, actores y periodistas de México, Argentina, Venezuela, Perú, Chile, España y Colombia. Por cada país se seleccionó a una mujer y un hombre. Posteriormente se creó un corpus por cada famoso, conformado por 100 tweets extraídos directamente de su perfil de X. En total se crearon 14 corpus, uno orientado al género femenino y otro al género masculino de cada país. En la Figura 5.3 se muestra uno de los experimentos orientados al género masculino. Los tweets se extrayeron del perfil del cantante Maluma (hombre colombiano).

5.3.1. Resultados generales de famosos

En la Tabla 5.9 se reporta el total de experimentos dividido en 3 secciones, los que tuvieron ambas predicciones correctas, los que fallaron en la predicción del género y los que fallaron en la predicción del país.

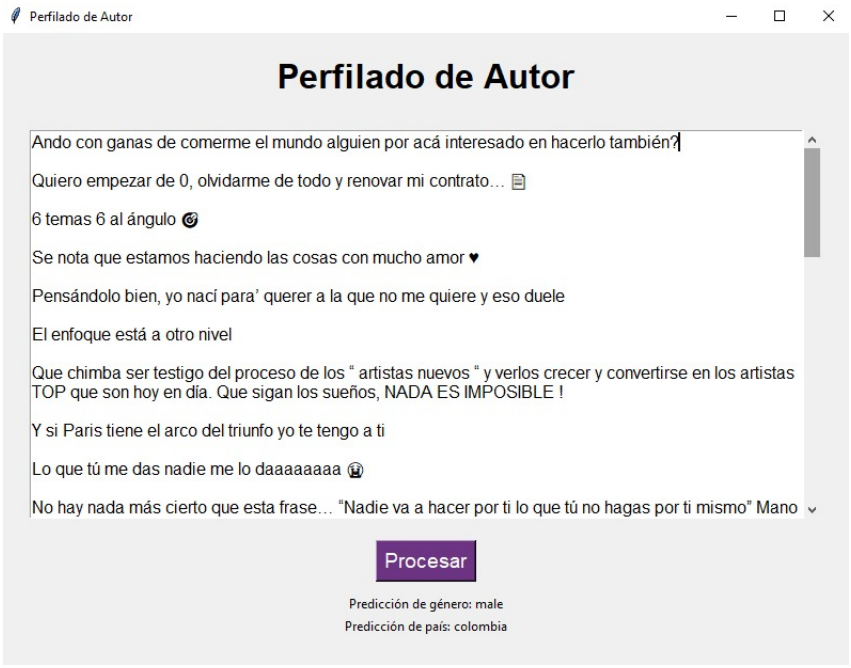

Figura 5.10: Experimento con tweets extraídos del perfil de un famoso

Tabla 5.9: Experimentos con tweets de famosos

Resultado de pruebas	**Cantidad**
Ambas clasificaciones correctas	10
Error en la clasificación de género	3
Error en la clasificación de país	1
Total	14

5.4. Costos de proyecto

En esta sección se presenta la Tabla 5.10 con los costos mensuales y anuales para la elaboración de este proyecto:

Tabla 5.10: Costos del proyecto

Gastos de operación	Descripción	Costo mensual	Costo anual
Sueldo	Desarrollador de Software Jr.	$15,000	$180,000
	Desarrollador de Software Jr.	$15,000	$180,000
Software y Hardware	Laptop Gateway	$	$11 899
	Windows 11	$	$7,000
	Laptop Lenovo	$	$9,514
	Windows 10	$	$2,700
Servicios	Luz	$507	$6,084
	Internet	$579	$6948
Total		$31,086	$404,125

5.5. Conclusiones

Al momento de evaluar los 3 modelos se determinó que el método de Máquinas de Soporte Vectorial obtuvo la mayor exactitud para la predicción de género respecto a los otros dos métodos evaluados.

Para la predicción de la variedad linguística, la mayor exactitud fue obtenida por el método de Multinomial Naive Bayes.

El sistema propuesto utilizó librerías preexistentes de PLN en Python y parámetros por defecto para los métodos supervisados, incluyendo un kernel lineal en SVM.

Los resultados muestran que es más difícil predecir el género masculino respecto al femenino, con una diferencia del 1 %. Entre los países, Chile fue el más fácil de predecir, mientras que México resultó ser el más difícil.

El sistema, centrado en el análisis del contenido de los tweets, demostró un tasa de error del 1.56 % para la predicción de género y del 1.05 % para la predicción de país.

Además, se encontró que la exactitud de los modelos mejora con un rango de 30 a 40 tweets (1680 - 2240 palabras).

5.6. Trabajos a futuro

- Se propone llevar el sistema a una plataforma web con el objetivo de que se encuentre al alcance de cualquier persona que posea un dispositivo conectado a internet.

-Entrenar el sistema para predecir una mayor variedad de países de habla hispana

Capítulo 6

Glosario

Corpus: Conjunto de textos recopilados, estructurados y utilizados en el ámbito del procesamiento de lenguaje natural.

Bibliografía

[1] Sánchez, A. L. (2021). El debate sobre la digitalización y la robotización del trabajo (humano) del futuro: automatización de sustitución, pragmatismo tecnológico, automatización de integración y heteromatización. Dialnet.

[2] Zúñiga, F. B., Poveda, D. A. M., Mora, D. P. M. (2023). La importancia de la inteligencia artificial en las comunicaciones en los procesos marketing. Vivat Academia (Alcalá de Henares), 19-39.

[3] Ameer, I., Sidorov, G., Nawab, R. M. A. (2019). Author profiling for age and gender using combinations of features of various types. Journal of intelligent fuzzy systems, 36(5), 4833–4843.

[4] Sánchez, J. C. A., Antonio, A. I. H., González, J. A. R., Belman, H. I. L., Santamaría, L. M. L., Carranza, J. C. G. (2021). Perfilado demográfico de celebridades en redes sociales. JÓVENES EN LA CIENCIA, 10.

[5] Rodríguez Bacelar, D. (2023). Perfilado automático de usuarios en corpus sociales sobre el movimiento Black Lives Matter.

[6] Ouni, S., Fkih, F., Omri, M. N. (2023). A survey of machine learning-based author profiling from texts analysis in social networks. Multimedia Tools and Applications, 82(24), 36653-36686.

[7] Monroy, A. P. L. (2012). Atribución de Autoría utilizando distintos tipos de características a través de una nueva representación (Doctoral dissertation, Instituto Nacional de Astrofísica, Óptica y Electrónica).

[8] Khan, M. H., Khan, B., Jan, S., Chughtai, M. I. Author's Age and Gender Prediction on Hotel Review Using Machine Learning Techniques.

[9] Silva, J., García, S., Binda, M. A., Gonzalez, F. M., Barrios, R., Castro, B. L., Castro, L. (2020). A method for detecting the profile of an author. Procedia Computer Science, 170, 959-964.

[10] Para, U. ., Patel, M. S. . (2023). A New Term Representation Method for Gender and Age Prediction . International Journal on Recent and Innovation Trends in Computing and Communication, 11(5s), 90–104. https://doi.org/10.17762/ijritcc.v11i5s.6633

[11] Ameer, I., Sidorov, G., Nawab, R. M. A. (2019). Author profiling for age and gender using combinations of features of various types. Journal of Intelligent Fuzzy Systems, 36(5), 4833-4843.

[12] Radha, D., Chandra Sekhar, P. (2019). Author profiling using stylistic and N-gram features. International Journal of Engineering and Advanced Technology, 9(1), 3044–3049. https://doi-org.bibliotecaipn.idm.oclc.org/10.35940/ijeat.A1621.109119

[13] Bevendorff, J., Borrego-Obrador, I., Chinea-Ríos, M., Franco-Salvador, M., Fröbe, M., Heini, A., ... Zangerle, E. (2023, September). Overview of PAN 2023: Authorship Verification, Multi-Author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection: Condensed Lab Overview. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 459-481). Cham: Springer Nature Switzerland.

[14] CLEF Initiative. (2023, September 18-21). CLEF 2023 Conference and Labs of the Evaluation Forum: Information access evaluation meets multilinguality, multimodality, and visualization. Retrieved May 30, 2024, from https://clef2023.clef-initiative.eu/

[15] PAN. (2024). PAN at CLEF 2024. Retrieved May 30, 2024, from https://pan.webis.de/

[16] Ashraf, M. A., Nawab, R. M. A., Nie, F. (2020). Author profiling on bi-lingual tweets. Journal of Intelligent Fuzzy Systems, 39(2), 2379-2389.

[17] Vashisth, P., Meehan, K. (2020, June). Gender classification using twitter text data. In 2020 31st Irish Signals and Systems Conference (ISSC) (pp. 1-6). IEEE.

[18] Hussein, S., Farouk, M., Hemayed, E. (2019). Gender identification of egyptian dialect in twitter. Egyptian Informatics Journal, 20(2), 109-116.

[19] GenderAnalyzer_v5 classifier. (n.d.). Uclassify.com. Retrieved June 23, 2024, from https://www.uclassify.com/browse/uclassify/genderanalyzer_v5?input=Text

[20] Classify. (2019, September 17). Estilometría. Retrieved from https://www.estilometria.com/herramienta/uclassify/

[21] Gender analyzer. (n.d.). Readable.com. Retrieved June 23, 2024, from https://app.readable.com/text/gender

[22] Figueroa Sacoto, S. S. (2021). Diseño y desarrollo de un chatbot usando redes neuronales recurrentes y procesamiento de lenguaje natural para tiendas virtuales en comercio electrónico (Bachelor's thesis).

[23] Hernández, M. B., & Gómez, J. M. (2013). Aplicaciones de Procesamiento de Lenguaje Natural. Revista Politécnica, 32. Recuperado a partir de https://revistapolitecnica.epn.edu.ec/ojs2/index.php/revista_politecnica2/article/view/32

[24] Viera, Á. F. G. (2017). Técnicas de aprendizaje de máquina utilizadas para la minería de texto. Investigación bibliotecológica, 31(71), 103-126.

[25] Ben-David, S. S.-S. A. (2014). Understanding Machine Learning from theory of algorithms. Cambridge University Press.

[26] Archiles, A. (2021, diciembre 7). Ejemplos de lenguaje natural. Una definición a partir de sus usos. Elipse.ai.

[27] Moreira, D., Cruz, I., Gonzalez, K., Quirumbay, A., & Magallan, C. (2021). Análisis del Estado Actual de Procesamiento de Lenguaje Natural. Revista Ibérica de Sistemas e Tecnologías de Informação, 42, 126–136.

[28] Procesamiento del Lenguaje Natural. (2020, octubre 27). Instituto de Ingeniería del Conocimiento. Retrieved from https://www.iic.uam.es/inteligencia-artificial/procesamiento-del-lenguaje-natural/

[29] Jiménez-Badillo, D., & Román-Rangel, E. (2017). Clasificación automática de fragmentos de vasijas arqueológicas mediante el modelo Bolsa de Palabras. Jiménez-Badillo (ed.), Arqueología Computacional. Nuevos enfoques para la documentación, análisis y difusión del patrimonio cultural. México: Instituto Nacional de Antropología e Historia, 111-126.

[30] Franca Tapia, K. E., Palacios Alvarado, S.U., Ramírez Cruz J. M. (2018). Compilación de corpus para la detección de noticias falsas en español. Instituto Politécnico Nacional.

[31] Martínez, R. E. B., Ramírez, N. C., Mesa, H. G. A., Suárez, I. R., Trejo, M. D. C. G., León, P. P., & Morales, S. L. B. (2009). Árboles de decisión como herramienta en el diagnóstico médico. Revista médica de la Universidad Veracruzana, 9(2), 19-24.

[32] Jimenez Villar, V. (2020). Aumento de datos para tareas relacionadas al perfilado de autor. Tesis de Maestría. Instituto Nacional de Astrofísica, Óptica y Electrónica.

[33] González, R., Barrientos, A., Toapanta, M., & Cerro, J. D. (2017). Aplicación de las Máquinas de Soporte Vectorial (SVM) al diagnóstico clínico de la Enfermedad de Párkinson y el Temblor Esencial. Revista Iberoamericana de Automática e Informática Industrial, 14(4), 394-405.

[34] Support vector machines. (n.d.). Scikit-Learn. Retrieved June 20, 2024, from https://scikit-learn.org/stable/modules/svm.html

[35] Macchine a vettori di supporto (SVM). (n.d.). Mathworks.com. Retrieved June 20, 2024, from https://it.mathworks.com/discovery/support-vector-machine.html

[36] Cano, G., García-Rodríguez, J., Orts, S., García-García, A., Peña-García, J., Pérez-Garrido, A., & Pérez-Sánchez, H. (2017). Predicción de solubilidad de fármacos usando máquinas de soporte vectorial sobre unidades de procesamiento gráfico. Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería, 33(1-2), 97-102.

[37] Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019, April). Comparison between multinomial and Bernoulli naïve Bayes for text classification. In 2019 International conference on automation, computational and technology management (ICACTM) (pp. 593-596). IEEE.

[38] Dhruv, A. J., Patel, R., & Doshi, N. (2021). Python: the most advanced programming language for computer science applications. Science and Technology Publications, Lda, 292-299.

[39] Saabith, A. S., Vinothraj, T., & Fareez, M. (2020). Popular python libraries and their application domains. International Journal of Advance Engineering and Research Development, 7(11), 5.

[40] NLTK: Natural Language Toolkit. (n.d.). Retrieved from https://www.nltk.org/

[41] CountVectorizer. (n.d.). Scikit-Learn. Retrieved June 23, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

[42] 1.4. Support vector machines. (n.d.). Scikit-Learn. Retrieved June 20, 2024, from https://scikit-learn.org/stable/modules/svm.html

[43] 1.9. Naive Bayes. (n.d.). Scikit-Learn. Retrieved June 23, 2024, from https://scikit-learn.org/stable/modules/naive_bayes.html

[44] 1.10. Decision Trees. (n.d.). Scikit-Learn. Retrieved June 23, 2024, from https://scikit-learn.org/stable/modules/tree.html

[45] Rangel, F., Rosso, P., Potthast, M., & Stein, B. (2017). Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. Working notes papers of the CLEF, 48.

[46] Nieto Ramírez, E. (2017). XML: un nuevo lenguaje, una nueva necesidad. Ginecología y obstetricia de México, 85(9), 0-0.

[47] Romero, A. R. (2000). Introducción a XML en Castellano. IV Simpósium Internacional de Telemática.

[48] Antonio de la Rosa. Introducción a XML para Documentalistas [en línea]. "Hipertext.net", núm. 1, 2003.

[49] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150.

[50] Mamgain, S., Balabantaray, R. C., & Das, A. K. (2019, December). Author profiling: Prediction of gender and language variety from document. In 2019 International Conference on Information Technology (ICIT) (pp. 473-477). IEEE.

[51] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.

[52] Sidorov, G. (2013). Construcción no lineal de n-gramas en la lingüística computacional. Sociedad mexicana de inteligencia artificial.

[53] Shahmirzadi, O., Lugowski, A., & Younge, K. (2019, December). Text similarity in vector space models: a comparative study. In 2019 18th IEEE international conference on machine learning and applications (ICMLA) (pp. 659-666). IEEE.

[54] Atkinson-Abutridy, J. (2022). Text Analytics: An Introduction to the Science and Applications of Unstructured Information Analysis. Chapman and Hall/CRC.

[55] Deng, X., Li, Y., Weng, J., & Zhang, J. (2019). Feature selection for text classification: A review. Multimedia Tools and Applications, 78(3), 3797-3816.

[56] Hackeling, G. (2017). Mastering Machine Learning with scikit-learn. Packt Publishing Ltd.

[57] Wang, J., & Dong, Y. (2020). Measurement of text similarity: a survey. Information, 11(9), 421.

[58] Qader, W. A., Ameen, M. M., & Ahmed, B. I. (2019, June). An overview of bag of words; importance, implementation, applications, and challenges. In 2019 International Engineering Conference (IEC) (pp. 200-204). IEEE.

[59] Paniagua Martín, F. (2021). Lenguajes de marcas y sistemas de gestión de información. Ediciones Paraninfo, SA.

Capítulo 7

Anexo

Bibliotecas complementarias para este proyecto:

xml.etree.ElementTree: Se implementó esta biblioteca para analizar y manipular documentos xml.

glob: Se utiliza para buscar archivos que coincidan con un patrón específico.

string: Proporciona funciones que ayudan a la manipulación de cadenas de texto.

joblib: Permite guardar y cargar modelos de aprendizaje automático como los son modelos entrenados y vectores de características.

tkinter: Proporciona herramientas para la creación de la interfaz gráfica