IDEAS EN CIENCIAS DE LA INGENIERÍA
Facultad de Ingeniería, UAEMEX

Detección del género de un autor en textos en español

Author gender detection in Spanish text

Priego-Sánchez, B.[1]*, Pinto, D.[2], Honorato, H.[1] (*abps@azc.uam.mx)

Recibido: 1 de marzo de 2023 Aceptado: 10 de julio de 2023

Resumen

Al navegar a través de la red, día a día, se tiene acceso a una gran cantidad de información; sin embargo, datos sobre el autor de estos contenidos no está a la vista o de manera explícita. Por lo que, al procesar e interpretar esta información es posible obtener características que identifiquen al autor como identificación del género, el rango de edad, la variante del lenguaje, entre otras características más. En este artículo, se presenta un trabajo en progreso que permite analizar y procesar textos de comentarios en español, con el objetivo de identificar el género (hombre o mujer) del autor. El proceso de análisis y procesamiento del sistema toma en cuenta las palabras, los emoticones de los textos y la utilización de algoritmos de aprendizaje automático. Los algoritmos empleados son Naïve Bayes y Árboles de decisión (J48), obteniendo el mejor resultado de precisión del 69.81 % con J48 para la identificación de género.

Palabras clave: *Perfilado de autor, Identificación de género, Análisis de textos en español.*

Abstract

Nowadays, when browsing through the network you have access to a lot of information; however, data about the author of these contents is not visible or explicitly. Therefore, when processing and interpreting this information, it is possible to obtain characteristics that identify the author as the gender identification, the age range, the variant language, among other characteristics. In this paper, we present a work in progress that allows to analyze and process texts of comments in Spanish, with the aim to identify author gender (man or woman). The process of analysis and processing of the system

[1]Universidad Autónoma Metropolitana unidad Azcapotzalco, Departamento de Sistemas, México.
[2]Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, México.

considers the words, text emoticons and the use of machine learning algorithms. The algorithms used are Naïve Bayes and Decision Trees (J48), obtaining the best precision result of 69.81 % with J48 for author gender identification.

Keywords: *Author profiling, Gender identification, Analysis of texts in Spanish.*

Introducción

Al redactar el contenido de un texto ya sea para una página web o para un posteo en una red social, se hace uso de diferentes palabras, términos y signos de puntuación que revelan patrones que permiten descubrir detalles del autor como el sexo, la edad, el origen geográfico, su orientación sexual, lengua materna, profesión, clase social, formación académica y otras características demográficas [1]. Sin embargo, se debe tener en cuenta dos aspectos relevantes que surgen: en primer lugar, la información a la que se tiene acceso a través de internet es inmensa y en ocasiones, la información en páginas como redes sociales están plagadas de texto que no cuentan con una buena ortografía e incluso se hacen uso de palabras y emoticones fabricados; en segundo lugar, el texto escrito por un supuesto autor puede ser plagio o incluso simplemente un texto escrito con palabras escuchadas en el transcurso del día. El Procesamiento de Lenguaje Natural (PLN) tendrá un papel importante al momento de analizar un texto en el idioma español, ya que la idea del análisis de texto surge meramente de esta área [2].

En este artículo se presenta un sistema capaz de analizar y procesar textos en español con la finalidad de reconocer el género de su autor. El análisis y procesamiento que realiza el sistema se basa en dos características: el conteo de las palabras y el conteo de emoticones que corresponden a hombres o mujeres. Para determinar qué palabra o emoticón corresponde a un hombre, a una mujer o a ambos, el sistema realiza dos tipos de búsquedas (igualdad y similitud) en diferentes conjuntos de textos que están conformados por las palabras y emoticones de mayor uso por hombres y mujeres que son extraídas de: libros de texto, de un corpus comentarios, páginas de internet y preguntas hacia distintas personas. El sistema representa dos opciones para introducir el texto a analizar (texto plano y adjuntar archivo) y el resultado se visualiza con el porcentaje calculado para cada género. A partir de los textos de entrada el sistema realiza un análisis basado en la construcción de vectores de característica empleando los algoritmos de aprendizaje automático: Naïve Bayes y Árboles de decisión (J48) que ofrece el software Weka, con la finalidad de determinar el género del autor del texto.

Un proceso de evaluación de esta tarea de identificación automática del género del autor del texto fue realizado, el cual consistió en tomar en cuenta 8,160 textos como conjunto de prueba y tratar de determinar el género del autor a partir de los textos de prueba. Los mejores resultados fueron obtenidos con el algoritmo Árbol de decisión (J48) logrando identificar 4,345 textos correctamente, alrededor de un 53.24 % de precisión.

En la literatura están reportados trabajos con diversos enfoques y aplican diferentes técnicas relacionadas al perfilado de autor; a continuación, se describen algunos trabajos relacionados con el presentado en este artículo. En los trabajos [3–6], toman en cuenta un conjunto de comentarios de la red social Twitter y microblogs, se analizan y procesan para la extracción de ciertas palabras y éstas se guardan en bases de datos; la finalidad es analizar y procesar un comentario de entrada para poder clasificarlo de acuerdo con las palabras que lo conforman. Realizando un análisis y procesamiento de textos para la extracción de palabras y guardarlas y ocuparlas para procesar diversos textos de entrada.

En [7] se implementan y comparan algunos algoritmos de aprendizaje automático para la clasificación emocional de noticias seguras e inseguras a partir de títulos de noticias en español, utilizando una diversidad de palabras extraídas de diversos textos con el objetivo de construir un vector de características a partir ciertas características léxicas-sintácticas y hacer uso de los algoritmos supervisados Naïve Bayes y Árboles de decisión (J48) del software de Weka para realizar predicciones; además, sólo se hace uso de las palabras y está orientado a la clasificación de noticias como buenas o malas. En [8] se investigan, analizan y comparan técnicas de procesamiento de lenguaje natural para la clasificación de documentos a partir de la identificación y extracción de información subjetiva como opiniones, sentimientos y emociones. Las principales afinidades que se tienen con este trabajo, son la identificación y extracción de información subjetiva en textos y la aplicación del proceso de clasificación supervisado; sin embargo, las principales diferencias radican en que no se hace uso del análisis de los emoticones y se basa en el estudio de diversas técnicas para la implementación de un análisis de sentimientos.

Los algoritmos de aprendizaje automático han sido ampliamente utilizados en tareas de clasificación de textos por tema, autor o polaridad como se ha analizado en la literatura. Por ello, este artículo se centra en dos de los algoritmos más utilizados en la literatura para resolver el problema de la identificación automática del género del autor de comentarios en español, estos algoritmos son Naïve Bayes y Árboles de decisión (J48). Si bien han sido utilizados en diversas tareas, con lo que respecta a abordar textos en español, el uso de estos algoritmos y diversos recursos son escasos para dicho lenguaje.

El resto del artículo se organiza como sigue. La Sección 2 presenta la propuesta de solución para la identificación automática del género de autor a partir de textos en español. En la Sección 3, se presentan los resultados de los experimentos realizados. Finalmente, en la Sección 4 se presentan las conclusiones y el trabajo a futuro.

METODOLOGÍA

En este apartado, se presenta la arquitectura de solución propuesta para implementar un sistema de detección de género de autor a partir de comentarios en español. Este sistema está basado en el análisis y procesamiento de palabras y emoticones contenidos en los textos. La Figura 1 muestra la arquitectura que consta de cuatro módulos descritos a continuación.

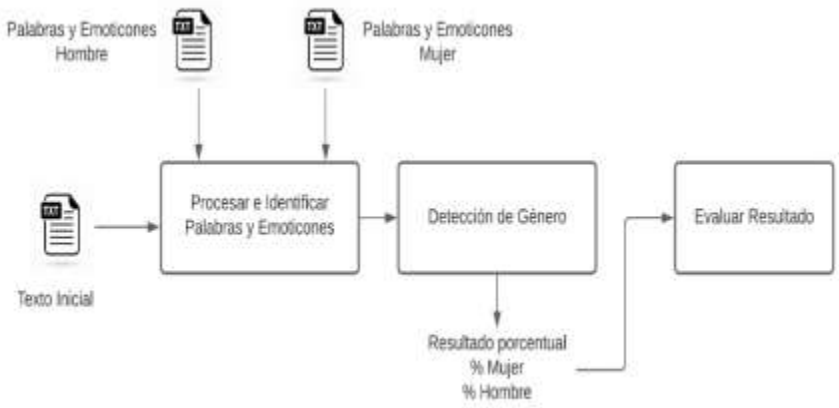


FIGURA 1. *Arquitectura de solución para la identificación de género en un texto en español.*

Selección de palabras y emoticones

El objetivo de este módulo, se centra en recabar las palabras y emoticones que son representativas y tiene un grado de relevancia en los textos escritos por hombres y mujeres. Para el funcionamiento de este módulo, se llevó a cabo la división de éste en cinco subetapas centradas en los textos: extracción, recolección, limpieza, pre-procesamiento y generar los listados que permiten el correcto funcionamiento de dicho objetivo y permiten que el resultado sea utilizado en el módulo siguiente.

Extracción: En esta etapa se consideran dos fuentes de donde se extrajeron las palabras. En primer lugar, dos libros (textos en formato plano), bajo los nombres de: "El laberinto de la soledad", del autor Octavio Paz, y "La última sirena", de la autora Eva Millet, de los cuales se optó por borrar su índice, agradecimientos y prefacio. En segundo lugar, se considera un corpus de comentarios escritos por hombres y mujeres utilizados en la tarea de perfilado de autor del PAN[1] [9] en la edición del 2013. Este corpus cuenta con 75,900 archivos en formato XML, clasificados por un ID y el género del autor. La obtención de los textos se ha basado en un método que extrae el texto delimitado por las etiquetas de marcado

[1]PAN es una serie de eventos científicos y tareas compartidas sobre texto digital forense y estilometría.

(<TAG></TAG>), convierte el texto en minúscula y, finalmente, generando archivos de textos plano con los textos correspondientes a cada género. Se optó por utilizar dos tipos de contextos y textos diferentes debido a que esto permite la comparación de la metodología en diversos tipos de escritura (formal e informal) en el lenguaje español; considerando textos bien escritos (libros literarios) y textos con detalles ortográficos e informalidad de escritura y llegando a la conclusión de que las técnicas empeladas en este artículo se pueden emplear para cualquier tipo de escritura y contexto.

Recolección: Esta tarea recaba los emoticones de mayor uso por hombres y mujeres en dos formatos de texto plano; para poder realizarlo, fue necesaria la construcción de un lexicón con emoticones asociados a hombres o mujeres. La recolección se realiza a través del informe presentado en [10]; documento en el cual, mediante preguntas sobre el uso de emoticones a 10 hombres y 10 mujeres, se obtienen subconjuntos de emoticones que caracterizan a cada perfil.

Limpieza: En esta etapa se realiza un filtrado y eliminación de elementos no relevantes de en los textos de hombres y mujeres que se obtuvieron en la etapa de "Extracción". Este filtrado consiste en eliminar correos, caracteres especiales, etiquetas de marcado que pudiesen haber quedado dentro de los textos y reemplazar puntuación, así como caracteres especiales por espacios en blanco, teniendo como principal objetivo un texto estándar en formato y texto.

Pre-procesamiento: Esta tarea se dividió en cuatro subtareas más: normalización de los textos, *tokenización*, identificación y eliminación de palabras no relevantes (palabras que no aportan en la tarea de identificación del género) y lematización. Para llevar a cabo cada una de estas subtareas, se tomaron en cuenta los textos generales de hombres y mujeres después de haber pasado por la tarea de "Limpieza".

a) *Normalización*: el proceso de normalización consiste en transformar sólo los textos de los libros (primer conjunto de datos) a minúsculas.

b) *Tokenización*: este proceso se ha llevado a cabo mediante el uso de expresiones regulares que consideran los espacios en blanco, valores numéricos, caracteres especiales y saltos de línea como marcadores de separación de palabras. El resultado de tokenizar los textos son listas de palabras representativas de cada uno de los grupos de textos.

c) *Identificación y eliminación de palabras no relevantes*: Esta tarea es considerada una de las de mayor importancia, ya que consistió en recabar un conjunto de palabras que no ayudan a la identificación de género. En primer lugar, se reunió una gran variedad de *stopwords* que se obtuvieron a través del uso de la librería de Python y de la literatura tomadas de [11–14] En segundo lugar, se contempló un conjunto de nombres comunes de hombres y mujeres tomados de [15]. Finalmente, tras un análisis, se tomó en cuenta una gran variedad de palabras que se consideró que no son de

gran aporte para la identificación de género. Una vez identificadas, todas las palabras anteriormente descritas, se eliminan de las listas de palabras representativas; así como la presencia de cantidades numéricas.

d) *Lematización*: El objetivo principal de esta etapa fue lematizar las listas de palabras que son representativas de los textos después de concluir la tarea de "Tokenización", así como también lematizar la lista de palabras no útiles.

Obtención: Esta última tarea se basa en obtener las palabras y emoticones finales que son útiles para la identificación de género. Para cumplir este objetivo se ha dividido en tres métodos: conteo, estructuración y selección.

a) *Conteo*: El objetivo de este método es realizar la tarea de conteo de las palabras y de los emoticones finales de las listas de palabras que se obtienen al finalizar las tareas de "Recolección" y "Pre-Procesamiento". Los conteos se pueden visualizar en las Tablas 1 y 2.

b) *Estructuración*: Una vez que se obtuvieron los conteos finales de las palabras y emoticones, se dio a la tarea de crear diez listas generales en texto plano, con el objetivo de que en los experimentos se observara la relación entre los archivos y la división. En primer lugar, se crearon cuatro archivos por cada género (Hombre y Mujer), los cuales se conforman del 10 %, 20 %, 30 % y 100 % del total de palabras no repetidas y mezclando las palabras obtenidas de cada género. En segundo lugar, se creó una lista de palabras por cada género (Hombre y Mujer) conformada por el 100 % de emoticones reunidos sin repeticiones. La estructuración de cada uno de los archivos se visualiza en la Tabla 3.

c) *Selección*: La finalidad de este método es seleccionar las palabras para conformar los 8 archivos generales que se muestran en la Tabla 3. La selección de las palabras se basó en los siguientes criterios: 80 % de palabras con mayor frecuencia, 17 % de palabras con frecuencia media y el 3 % de palabras con baja frecuencia; se reportan estos criterios debido a que son en los que se obtienen los mejores resultados. Cabe resaltar que estos porcentajes se aplicaron al total de palabras de cada uno de las listas especificadas en la Tabla 1.

TABLA 1. *Cantidades de palabras funcionales con y sin repetición de las listas de textos.*

Archivo	**Nro. de palabras finales con repeticiones**	**Nro. de palabras finales sin repeticiones**
Laberinto_de_la_soledad.txt (1)	23,686	5,000
La_ultima_sirena.txt (2)	53,133	6,686
Comentarios_Mujer.txt (3)	2,909,077	147,877
Comentarios_Hombre.txt (4)	3,387,956	171,459

TABLA 2. *Cantidades finales de las listas de emoticones.*

Archivo	**Nro. de emoticones finales con repeticiones**	**Nro. de emoticones finales sin repeticiones**
Emoticones_Mujer.txt	100	86
Emoticones_Hombre.txt	80	64
Total	180	150

Procesamiento e identificación de palabras y emoticones

Este módulo se encarga de procesar un texto de entrada con la finalidad de identificar las palabras y emoticones que corresponden a cada género (Hombre y Mujer). Esta identificación involucra los procedimientos de limpieza, normalización, *tokenización*, *lematización* y eliminación de palabras no relevantes, descritos anteriormente. La necesidad de procesar los textos en el sistema tiene como propósito obtener las palabras útiles y los emoticones de cualquier texto plano de entrada. El objetivo final de este módulo es, una vez obtenidas las palabras útiles y los emoticones de un texto de entrada, buscar dentro de las bases de datos cada una de las palabras útiles y los emoticones identificados. Estas bases de datos son expuestas en la Tabla 3 y corresponden a la selección de las listas de palabras y emoticones. Los procesos de búsqueda que se realizan en este módulo son: búsqueda por igualdad y búsqueda por similitud.

TABLA 3. *Estructura preliminar de las listas generales de palabras de hombres y mujeres.*

Archivo	**(1)**	**(4)**	**(2)**	**(3)**	**Total palabras**
Palabras_Mujer_10 %			669	14,788	15,457
Palabras_Mujer_20 %			1,337	29,576	30,913
Palabras_Mujer_30 %			2,006	44,364	46,370
Palabras_Mujer_100 %			6,686	147,877	154,563
Palabras_Hombre_10 %	500	17,146			17,646
Palabras_Hombre_20 %	1,000	34,292			35,292
Palabras_Hombre_30 %	1,500	51,438			52,938
Palabras_Hombre_100 %	5,000	331,021			336,021

Búsqueda por igualdad: Este tipo de búsqueda se basa en encontrar una palabra o emoticón en el texto, los tres posibles resultados de esta búsqueda son:

1. Si se encuentra la palabra o emoticón dentro de la base de datos de uno de los géneros (Hombre o Mujer), se aumenta en uno el contador para ese género,

2. Si la palabra o emoticón se encuentra en las dos bases tanto de hombres y de mujeres, se aumenta en uno los contadores de ambos géneros.

3. Si una palabra o emoticón no se encuentra en ninguna base de datos, la palabra se guarda en una lista de palabras no encontradas y el emoticón es desechado.

Búsqueda por similitud: La búsqueda por similitud se encarga de localizar la palabra con mayor similitud léxica a la palabra del listado. Este proceso de búsqueda se implementa utilizando tres elementos: palabras no encontradas, el algoritmo “distancia de Levenshtein” y el uso del cálculo de afinidad (ver Ecuación 1). Para determinar la similitud de las palabras contenidas en las bases de datos de hombres o mujeres y los textos, se ha propuesto una serie de condiciones, donde la más utilizada en este trabajo, es si el cálculo final de afinidad es mayor a 50, es considerada como similar.

En la Ecuación 1, AF es definido como la “afinidad”; D hace referencia al número de diferencias a nivel de letras que arroja como resultado el algoritmo “distancia de Levenshtein”; y L es la longitud mayor de las dos palabras que se comparan.

$$AF = (1 - \frac{D}{L}) \times 100 \tag{1}$$

Detección de género

El módulo de detección de género se encarga de mostrar una respuesta final por parte del sistema; es decir, toma en cuenta los resultados de los conteos de las palabras y emoticones que se obtuvieron tanto para hombres y mujeres, y expone aquellos resultados con mayor coincidencia hacia cada género. Este resultado, representa la identificación automática del género de un autor, se calcula de acuerdo con la Ecuación 2 y es aplicado tanto para la identificación de hombres como para la de mujeres.

$$RG = \text{Redondeo} \left(\frac{(PP + EP) \times 100}{PF + EF} \right) \tag{2}$$

En la Ecuación 2, RG representa el resultado final para el género (Hombre o Mujer) evaluado, PP hace referencia a la cantidad de palabras que se contabilizaron para un género (Hombre o Mujer) en específico, EP es la cantidad de emoticones que se contabilizaron para un género, PF y EF hacen alusión al total de palabras útiles y al total de emoticones que se encontraron dentro del texto. Esta ecuación es utilizada para la identificación de género, es importante resaltar que dentro de las cantidades que constituyen a PF y EF, se consideraron las repeticiones.

Los resultados que se obtienen son:

1. "Hombre o Mujer", si al realizar los cálculos tanto para hombre y mujer, uno resulta ser mayor que otro.
2. "Empate", si al realizar los cálculos tanto para hombre y mujer, los resultados son iguales.
3. "Texto no funcional", si la suma de PF + EF es cero.

Entrenamiento y pruebas

Este módulo tiene como objetivo realizar un segundo análisis para la detección automática de género de autor y poder evaluar los resultados arrojados por el sistema. Los algoritmos empleados, para la identificación de género de autor, son Naïve Bayes y Árboles de decisión (J48) y en éstos se aplicaron tanto para el conjunto de datos como para el conjunto de prueba.

El conjunto de entrenamiento se conformó por 75,902 textos, contemplando los textos de los dos libros y los 75,900 textos de comentarios. El conjunto de pruebas se conforma de 8,160 textos separados

del conjunto de comentarios; dado que se contaba con un conjunto de pruebas, fue posible entrenar con el conjunto para este fin. A partir de estos textos (entrenamiento y pruebas) se obtienen vectores de características, las cuales se describen a continuación:

1. El número total de palabras antes de los procesos limpieza, normalización, tokenización, lematización y eliminación de palabras no relevantes.

2. El número total de palabras después de los procesos limpieza, normalización, tokenización, lematización y eliminación de palabras no relevantes.

3. El número total de palabras identificadas de mujer después de los procesos descritos en el punto 1, 2, usando todos los listados.

4. El número total de palabras de hombre después de los procesos descritos en el punto 1, 2, usando todos los listados.

5. El número total de emoticones encontrados.

6. El número total de emoticones de mujer identificados usando la base del 100 %.

7. El número total de emoticones de hombre identificados usando la base del 100 %.

8. El género del autor en el archivo de entrenamiento y el signo de interrogación de cierre para el archivo de pruebas.

Entrenamiento para Naïve Bayes y Árboles de decisión (J48). Los clasificadores Naïve Bayes y Árboles de decisión (J48) fueron entrenados con estos vectores de características de los textos para tal fin, utilizando la herramienta WEKA. Las predicciones de género del autor se llevaron a cabo utilizando los métodos: Use training set, Cross-validation y Percentage-Split (70-30 % y 80-20 %, respectivamente); donde principalmente se usó una validación cruzada de 10 elementos y se cambiaron las propiedades de ejecución del algoritmo, cambiando las propiedades más importantes. El objetivo principal del entrenamiento de los clasificadores Naïve Bayes y Árboles de decisión (J48), fue realizar una comparación entre ambos algoritmos de clasificación y determinar el algoritmo con mejor exactitud para esta tarea; es decir, comparar el algoritmo que tiene el porcentaje más alto de instancias clasificadas correctamente.

Resultados

Con la intención de exponer los resultados que se obtuvieron a lo largo del desarrollo de este trabajo, esta sección se dividió en dos partes: resultados del sistema y resultados de las predicciones.

Resultados del sistema

En el módulo de "Selección de palabras y emoticones" se hizo mención de la creación de listas de palabras de mayor uso por hombres y mujeres. Además, se consideraron dos listas que contienen la cantidad total de emoticones reunidos con mayor incidencia en textos de hombres y mujeres. Por parte del sistema se utilizaron cada una de las bases (10 %, 20 %, 30 % y 100 % palabras) y el total de emoticones, para realizar la detección de los 8,160 textos de prueba. Los resultados se exponen en la Tabla 4.

TABLA 4. *Resultados de detección de género de autor considerando 100 % de emoticones.*

Resultado	**10 % de palabras**	**20 % de palabras**	**30 % de palabras**	**100 % de palabras**
Detecciones Correctas	2,192	1,968	1,935	1,676
Detecciones Incorrectas	1,559	1,400	1,269	5,175
Empates	4,334	4,717	4,881	-
Textos no funcionales	75	75	75	75
Total	8,160	8,160	8,160	8,160

Resultados de las predicciones Tras entrenar los algoritmos de aprendizaje automático utilizando los métodos de entrenamiento: *Use training set, Cross-validation (10 elementos) y Porcentaje-Split* (70-30 y 80-20 %, respectivamente); se eligió el clasificador "Árbol de decisión (J48)" junto con el método de entrenamiento "Use training set" para realizar las predicciones de género, ya que resultó ser la opción con mayor porcentaje de instancias clasificadas correctamente. En la Tabla 5 se muestran los resultados en porcentaje de las instancias correctamente clasificadas y en la Tabla 6 el resultado de las predicciones.

TABLA 5. *Resultados de los entrenamientos de clasificación.*

Clasificador	Use training set	Cross-validation	Percentage-split
Naïve Bayes	50.51 %	50.51 %	50.67 %
J48	69.81 %	66.64 %	66.34 %

TABLA 6. *Resultados de comparación de respuestas correctas e incorrectas (predicciones).*

Resultado	Predicciones
Predicciones de género correctas	4,345
Predicciones de género incorrectas	3,815
Total	8,160

CONCLUSIONES

En el presente trabajo se ha desarrollado un sistema fundamental para la detección de género de autor a través del análisis y procesamiento de palabras y emoticones presentes en texto en español. También, se construyó un lexicón basado en las frecuencias de las palabras y emoticones más utilizados por cada género. Si bien, el sistema realiza un análisis básico y procesamiento adecuadamente, al inspeccionar los resultados arrojados por el sistema, se nota una disminución en las detecciones automática de géneros de los textos. Esto se debe a que al considerar un mayor número de palabras tanto para hombres y mujeres, mayor es el número de palabras que comparten, lo que ocasionó fallas en la detección del género.

Los resultados experimentales comprobaron que en cada uno de los conjuntos de datos (10 %, 20 %, 30 % y 100 %) había una gran cantidad de palabras que compartían ambos géneros, afectando la predicción del género de los autores de los textos usando los algoritmos de aprendizaje automático. Aun cuando los resultados obtenidos son alentadores, estos se pueden mejorar para lograr una cantidad superior a 6,000 predicciones correctas de los 8,160 textos usados en las pruebas.

Los resultados se vieron afectados por el uso de textos con emoticones fabricados, textos compuestos de palabras con ruido, direcciones electrónicas, nombres de personas o textos demasiado extensos, entre otros factores. Si bien, el sistema no da una excelente detección del género, éste puede ser utilizado y empleado por analistas de textos para apoyarse en la predicción e identificación del género de la persona que escribe un texto, comentario o libro. Como trabajo a futuro, es posible seleccionar una mejor cantidad de palabras de hombres y mujeres basándose en mejores libros descriptivos. Finalmente, incorporar

un mayor número de características para la detección de género, tales como: características sintácticas, morfológicas y semánticas; sin embargo, es importante considerar que esto no garantiza el mejoramiento de la precisión, ya que vectores más grandes no garantizan un mejor desempeño.

REFERENCIAS

[1] P. Josep, "Los textos escritos delatan el sexo," 2016, consultado el 12 de febrero de 2023. [Online]. Available: https://www.lavanguardia.com/cultura/20161017/411054860677/linguistica-estu-dio-determina-sexo-autor-articulo-upf.html

[2] E. García Menier, "Análisis de textos por computadora," *Boletín de Lingüística*, vol. 18, no. 25, pp. 121–134, 2006.

[3] V. H. Rojas, "Sistema para la clasificación de opiniones generadas en twitter usando redes neuronales artificiales," División de Ciencias Básicas e Ingeniería, Universidad Autónoma Metropolitana Azcapotzalco, México, Tech. Rep., 2017.

[4] L. E. P. Terán, "Minería de opiniones sobre textos en twitter," División de Ciencias Básicas e Ingeniería, Universidad Autónoma Metropolitana Azcapotzalco, México, Tech. Rep., 2018.

[5] H. Librado, "Análisis automático de opiniones de productos en redes sociales," Master's thesis, Centro de Investigación en Computación, Instituto Politécnico Nacional, México, 2016.

[6] S. Delgado, "Modelo computacional de minería de microblogs para el análisis del comportamiento del consumidor de telefonía celular," Master's thesis, Escuela de Posgrado, Pontificia Universidad Católica del Perú, Perú, 2016.

[7] G. Villarruel, "Clasificación emocional de noticias en español utilizando algoritmos de aprendizaje automático," División de Ciencias Básicas e Ingeniería, Universidad Autónoma Metropolitana Azcapotzalco, México, Tech. Rep., 2018.

[8] L. Dubiau, "Procesamiento de lenguaje natural en sistemas de análisis de sentimientos," Master's thesis, Facultad de Ingeniería, Universidad de Buenos Aires, Argentina, 2013.

[9] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches, "Overview of the author profiling task at pan 2013," in *CLEF 2013*, R. N. Pamela Forner and D. Tufis, Eds., 2013.

[10] J. Hernández, "Primer informe sobre el uso de emojis, un idioma internacional," Blog, 2018, consultado 24 de febrero de 2022. [Online]. Available: https://blog.hostalia.com/informes/primer-informe-sobre-el-uso-de-emojis-un-idioma-internacional/

[11] "Full-text search – stopwords," 2017, consultado el 12 de febrero de 2023. [Online]. Available: https://wiki.easyvista.com/xwiki/bin/view/Documentation/Full+text+search+-+Stop+Words

[12] Savand, "stop-words," GitHub, 2020, consultado el 12 de febrero del 2023. [Online]. Available: https://github.com/Alir3z4/stop-words/blob/master/spanish.txt

[13] "The list of stop words," Countwordsfree Text Processing, Tools, 2009, consultado el 12 de febrero del 2023. [Online]. Available: https://countwords-free.com/stopwords/spanish

[14] D. E. Avendaño, "Stopwords spanish," Facultad de Ciencias de la Computación, BUAP, 2009, consultado el 24 de febrero de 2023. [Online]. Available: http://dpinto.cs.buap.mx/pln/course2009/Stopwords_Spanish.txt

[15] Olea, "Nombres propios y apellidos," GitHub, 2020. [Online]. Available: https://github.com/olea/lemarios/blob/master/nombres-propios-es.txt