Evaluación bolsas de palabras como característica estilométrica para atribución de autoría.

Evaluation bag of word as a stylometric feature for authorship attribution.

Favián Gutiérrez Constantino (1).
Estudiante. Tecnológico Nacional de México/Instituto Tecnológico de Tuxtla Gutiérrez, l16270780@tuxtla.tecnm.mx.

Luis Ángel Vázquez Moreno (2), estudiante, TecNM/ITTG, l16270839@tuxtla.tecnm.mx.

Germán Ríos Toledo (3)*, TecNM/ITTG, german.rt@tuxtla.tecnm.mx.

Héctor Guerra Crespo (4), TecNM/ITTG, hector.gc@tuxtla.tecnm.mx.

Aída Guillermina Cossío Martínez (5), TecNM/ITTG, aida.cm@tuxtla.tecnm.mx.

María Guadalupe Monjarás Velasco (6), TecNM/ITTG, maria.mv@tuxtla.tecnm.mx.

*corresponding author.

Artículo recibido en octubre 27, 2020; aceptado en noviembre 26, 2020.

Resumen.

La Atribución de Autoría es una disciplina del Procesamiento del Lenguaje Natural cuya finalidad es partiendo de un conjunto de autores conocidos, a quién de ellos pertenece un texto cuya autoría se desconoce. En este artículo se evaluaron la característica estilométrica conocida como bolsa de palabras para la representación de textos en Atribución de Autoría. Para este fin, se creó un corpus formado por 4 autores de novelas cuya lengua nativa es el español. Cada autor contó con una colección de 10 textos, cada texto contiene al menos 5,000 palabras. La Atribución de Autoría se abordó como un problema de clasificación con Aprendizaje Automático Supervisado y la métrica Exactitud. Los algoritmos de aprendizaje automático utilizados fueron Máquinas de Soporte Vectorial, Bosque Aleatorio, Multinomial Naive Bayes y Regresión Logística. Además, se evaluaron las técnicas de reducción de dimensiones Análisis de Componentes Principales y Análisis Semántico Latente. Los resultados de los experimentos mostraron que para algunos autores, la representación con bolsa de palabras logró una exactitud promedio superior al 80%.

Palabras claves: Atribución de Autoría, características estilométricas, algoritmos de aprendizajes automático supervisado.

Abstract.

Authorship Attribution is a discipline of Natural Language Processing whose purpose is based on a set of known authors, to whom belongs a text whose authorship is unknown. In this article, the stylometric characteristic known as bag of words for the representation of texts in Authorship Attribution was evaluated. For this purpose, a corpus made up of 4 novel authors whose native language is Spanish was created. Each author had a collection of 10 texts, each text contains at least 5,000 words. Authorship Attribution was addressed as a classification problem with Supervised

Machine Learning and the Accuracy metric. The machine learning algorithms used were Vector Support Machines, Random Forest, Multinomial Naive Bayes and Logistic Regression. In addition, the techniques of reduction of dimensions Principal Component Analysis and Latent Semantic Analysis were evaluated. The results of the experiments showed that for some authors, the representation with a bag of words achieved an average accuracy greater than 80%.

Keywords: Attribution of Authorship, stylometric characteristics, supervised automatic learning algorithms.

Introducción.

Stamatatos (2009) afirma que la atribución de autoría pretende construir métodos o modelos capaces de aprender el estilo de escritura de uno o más autores, para identificar automáticamente sus futuros documentos. Algunas de las aplicaciones relacionadas con Atribución de Autoría involucran: labores de inteligencia (atribución de mensajes o proclamaciones de terroristas), ley criminal (identificando escritores de mensajes intimidatorios, verificando la autenticidad de notas suicidas), ley civil (disputas de derecho de autor), ciencia forense computacional (identificar los autores de código fuente de software malicioso), además de las aplicaciones tradicionales en la investigación literaria (atribución anónima o trabajos en disputa de autores conocidos).

La estilometría es el análisis estadístico de textos escritos. Es decir, que hace referencia a un rasgo de la forma que un autor compone sus textos. Algunos ejemplos de características estilométricas son la distribución de frecuencia de longitud de palabra, longitud de oración, n-gramas de palabra, palabras con contenido semántico (sustantivos, verbos, adjetivos, adverbios), palabras funcionales (preposiciones, adverbios, artículos, pronombres, adjetivos), categorías gramaticales; errores de escritura, lemas, entre otras (Morales, 2007).

Uno de los estudios relacionados con la atribución de autoría fue el que realizaron (Sanderson, 2006), aplicaron la técnica de los núcleos de secuencia de palabras. Evaluaron un conjunto de textos relativamente cortos de 50 periodistas que cubrían más de un tema aplicando el enfoque de cadenas de Markov. Crearon textos de 312, 625, 1250, 2500 y 5000 palabras y de ellos obtuvieron 1750, 3500, 7000, 14000 y 28000 caracteres. Los investigadores indicaron que la cantidad de texto para el entrenamiento de los algoritmos tiene más influencia que la cantidad de textos de prueba. Además, concluyeron que se requieren aproximadamente entre 1250 y 5000 palabras en los textos de entrenamiento para obtener un rendimiento relativamente bueno.

Corney *et al*. (Corney, 2001) realizaron experimentos para identificar la autoría de correos electrónicos utilizando marcadores apropiados para este tipo de mensajes. Dichos mensajes contenían hasta 964 palabras, con una longitud promedio de 92 palabras. Utilizaron el algoritmo de aprendizaje automático supervisado SVM para discriminar entre las clases de autoría. Descubrieron que aproximadamente 20 mensajes con aproximadamente 100 palabras cada uno, deberían ser suficientes para discriminar la autoría en la mayoría de los casos. Mencionaron que el rendimiento del clasificador mejoró cuando agregaron un conjunto de características específicas de correo electrónico.

Luyckx y Daelemans (Luyckx, 2008) analizaron ensayos de un mismo tópico. Los ensayos contenían aproximadamente 1400 palabras y provenían de 145 estudiantes. Utilizaron palabras y n-gramas de etiquetas POS con el algoritmo de aprendizaje automático SVM. En sus conclusiones, argumentaron que su propuesta mostró solidez al tratar con datos limitados, pues de los 145 autores, casi el 50% de los textos fueron clasificados correctamente.

2. Metodología.

La metodología propuesta consta de 2 fases (véase Figura 1). La primera fase consiste en obtener las características estilométricas y la segunda en aplicar un enfoque de aprendizaje automático supervisado para clasificar documentos nuevos mediante las características seleccionadas en la etapa previa. A continuación, se describen detalladamente cada una de las actividades realizadas en cada etapa.

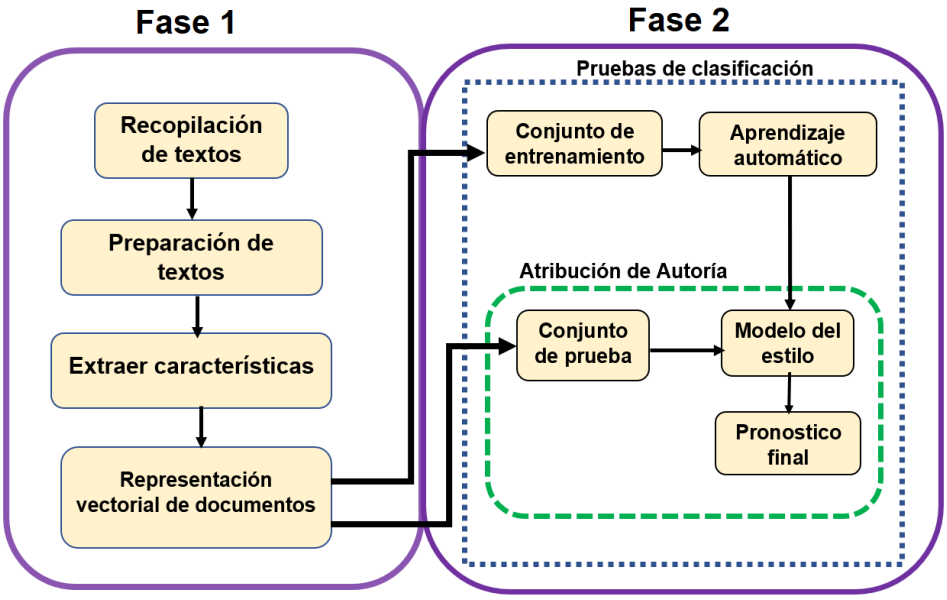

Figura 1. Diagrama a bloques de la propuesta.

Fase 1:

1. **Recopilación de textos:** en esta etapa se realizó una búsqueda de obras literarias del Siglo de Oro español, donde se encontró un corpus abierto de obras literarias en español, utilizadas en la revista "El análisis estilométrico aplicado a la literatura española: las novelas policiacas e históricas" (Rueda, 2016), en formato texto plano para estudios de estilometría[1]. A partir de esas novelas, se creó un corpus con 4 autores, con un promedio de 9 novelas cada uno.

2. **Preparación de textos:** en esta etapa se elimina de los textos toda la información que no es relevante, como por ejemplo la eliminación de los signos de puntuación y caracteres especiales. Para realizar estas acciones se utilizaron Expresiones Regulares (ER). Una ER es una secuencia de caracteres que forma un patrón de búsqueda, en este caso se utilizó un patrón para eliminar los caracteres especiales y para sustituir los números por una palabra clave. Además, se realizó la detección de Entidades Nombradas. Las entidades nombradas son palabras que hacen referencia a nombres de personas, lugares, organizaciones, etc (David Nadeau, 2007).

   Para detectarlas, se utilizó el Reconocedor de Entidades Nombradas (NER) de Spacy[2]. Cabe resaltar que las entidades nombradas en español solo existen 4 categorías (véase la Tabla 1).

Tabla 1. Categorías de entidades nombradas en español.

Tipo	Descripción
PER	Persona o familia nombrada
LOC	Nombre de la ubicación definida política o geográficamente (ciudades, provincias, países, regiones internacionales, masas de agua, montañas).
ORG	Entidad corporativa, gubernamental u otra organización designada.

[1] https://github.com/7PartidasDigital/NovelaBarroca
[2] https://spacy.io/models/es

MISC	Entidades diversas, por ejemplo, eventos, nacionalidades, productos u obras de arte.

Una vez detectada las entidades se procedió a remplazarla por sus respectivas etiquetas mediante un código desarrollado en Python. Después, se creó el diccionario de palabras. El diccionario está formado con todas las palabras diferentes que existen en todo el corpus. Para ello se utilizó el método *CountVectorizer()*, al cual convierte una colección de documentos de texto en un vector de recuentos de términos/tokens.

Posteriormente, las novelas fueron divididas en dos y tres partes, cada parte con la misma proporción en cuanto al número de palabras. Las novelas tienen distintos tamaños, pero en esta tarea se usaron las primeras 5000 palabras de cada una. Cada novela fue dividida en tres tamaños distintos: 5000, 2500 y 1666 palabras aproximadamente.

3. **Representación vectorial de documentos:** El Modelo Espacio Vectorial se utiliza para representar objetos por medio de sus características. En la práctica, dicho modelo es una matriz de dos dimensiones, donde las filas representan objetos, las columnas representan las características y las celdas contienen valores de distintos tipos (Toledo, 2019). Este modelo permite representar documentos por medio de cualquier característica estilométrica. Se crearon tres tipos distintos de matrices:
    - Frecuencia de palabras: indican las veces en que cada palabra es utilizada por un autor. Comúnmente, las dimensiones se ordenan de mayor a menor frecuencia
    - Binarias: representan la presencia o ausencia de cada palabra del diccionario en un texto determinado. Si la palabra está presente lo representa con el numero 1 y si no se encuentra con el 0.
    - Valor tf-idf: esta técnica se compone de dos términos.

        Término Frecuencia (***tf***): es la proporción de veces que la palabra aparece en un documento en comparación con el número total de palabras en ese documento. Aumenta a medida que aumenta el número de apariciones de esa palabra en el documento.

        Frecuencia de datos inversa (***idf***): se utiliza para calcular el peso de las palabras raras en todos los documentos del corpus. Las palabras que aparecen raramente en el corpus tienen una puntuación alta en las IDF.

      La medida tf-idf permite expresar la importancia de una palabra en un documento que forma parte de una colección de documentos. Si la palabra es muy común y aparece en muchos documentos, este número se acercará a 0. De lo contrario, se acercará a 1.

Fase 2:

Las distintas matrices se dividieron en conjuntos de entrenamiento y prueba. Comúnmente, la proporción de los conjuntos es de 80% de los datos para la etapa de entrenamiento y 20% para las pruebas. Posteriormente se evaluaron algunos algoritmos de aprendizaje automático supervisado, implementados en Scikit-Learn[3], una de las bibliotecas de aprendizaje automático más populares de Python. Scikit-Learn conteniene herramientas para visualización de datos, agrupamiento, reducción de dimensiones, selección de modelos, entre otras. Los algoritmos de aprendizaje automático supervisados utilizados fueron: Máquina de Soporte Vectorial (SVC), Naive Bayes (NB), Bosque Aleatorio (RF) y Vecinos más Cercanos (KNN). Para evaluar los algoritmos de clasificación existen algunas métricas como precisión, recall, accuracy y f1-score; se eligió la exactitud (accuracy) debido a que cada clase (autor), tiene el mismo número de ejemplos. Las clases están balanceadas. Con esta métrica se conoce la proporción de predicciones correctas que ha hecho el modelo del total de instancias de prueba. Las métricas junto con la matriz de confusión permiten conocer el desempeño de un modelo de aprendizaje. La matriz de confusión es una matriz cuadrada donde las filas se nombran según las clases reales y las columnas según las clases previstas por el modelo. La matriz muestra de forma explícita cuándo una clase es confundida con otra, lo que permite trabajar de forma separada con distintos tipos de error. La Tabla 2 muestra la representación de una matriz de confusión (Jiménez, 2010).

[3] https://scikit-learn.org/stable/

Tabla 2. Representación de una matriz de confusión.

Categoría	Texto positivo (Modelo)	Texto negativo (Modelo)
Texto positivo(real)	Tp	Tp
Texto positivo(real)	Fn	Tn

Donde los verdaderos positivos (Tp) es el número de elementos a los que se les asignó la clase en cuestión y realmente pertenecían a ella, Verdadero Negativo (Tn) son los números de elementos a los que no se les asignó la clase en cuestión y realmente no pertenecían a ella, Falso Positivo (Fp) representa el número de elementos a los que se les asignó la clase pero realmente no pertenecían a ella, Falso Negativo (Fn) es el número de elementos a los que no se les asignó la clase pero realmente pertenecían a ella y la diagonal principal contiene la suma de todas las predicciones correctas, representan el número de puntos para los cuales la etiqueta predicha es igual a la etiqueta verdadera, mientras que los elementos fuera de la diagonal son aquellos que están mal etiquetados por el clasificador. Cuanto más altos sean los valores diagonales de la matriz de confusión, mejor, lo que indica muchas predicciones correctas.

La implementación del algoritmo SVM (SVC, Support Vector Classifier) está basada en Libsvm. El algoritmo SVM se entrenó usando los hiperparámetros C, gamma y kernel. El parámetro C, común a todos los kernels (lineal, polinómica, RBF y sigmoide), intercambia errores de clasificación de los ejemplos de entrenamiento contra la simplicidad de la superficie de decisión. El parámetro gamma define cuánta influencia tiene un solo ejemplo de entrenamiento. La elección correcta de C, gamma y kernel es crítica para el rendimiento del SVM, para nuestros experimentos evaluamos con la variante de SVM (LinearSVC), cuyo uso se recomienda para grandes conjuntos de datos. Para determinar los valores apropiados de C, gamma y kernel se utilizó la estrategia GridSearch. Que consiste en realizar un ajuste de hiperparámetros para determinar los valores óptimos para un modelo dado. Los valores apropiados fueron C= 1, gamma= 'scale', kernel= 'linear'.

El algoritmo MNB implementa el algoritmo Naive Bayes para datos distribuidos multinomialmente, y es una de las dos variantes clásicas de Naive Bayes utilizadas en la clasificación de texto, donde los datos se representan típicamente como conteos de vectores de palabras. La Ecuación 1 muestra la base matemática del algoritmo de Naive Bayes.

$$\theta_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \qquad (1)$$

Dónde $N_{yi} = \sum x \in T^{x_i}$ es el número de veces la característica i en una muestra de clase en el conjunto de entrenamiento T y $N_y = \sum_{i=1}^{n} N_{yi}$ es el recuento total de todas las características para la clase y. Si α=1 se denomina suavizado Laplace, Si α<1 se llama suavizado Lidstone. Dentro de método, por defecto el parámetro alpha=1. Los antecedentes de suavizado $\theta \geq 0$ da cuenta de las características que no están presentes en las muestras de aprendizaje y evita cero probabilidades en otros cálculos.

En el algoritmo de K-Vecinos más cercanos (KNN), es un algoritmo que no tiene una fase de entrenamiento especializada. Más bien, usa todos los datos para entrenar mientras clasifica un nuevo punto de datos o instancia. KNN es un algoritmo de aprendizaje no paramétrico, lo que significa que no asume nada sobre los datos subyacentes, simplemente calcula la distancia de un nuevo punto de datos a todos los demás puntos de datos de entrenamiento.

El algoritmo de Bosque aleatorio (Random Forest, RF) es un tipo de algoritmo basado en el aprendizaje conjunto. El aprendizaje en conjunto es un tipo de aprendizaje en el que se unen diferentes tipos de algoritmos o el mismo algoritmo varias veces para formar un modelo de predicción más potente. El algoritmo de bosque aleatorio combina múltiples algoritmos del mismo tipo, es decir, varios árboles de decisión, lo que resulta en un bosque de árboles, de ahí el nombre "Random Forest".

Para el entrenamiento de los algoritmos, se utilizó la estrategia de Validación Cruzada. La validación cruzada de (K-fold Cross Validation) divide los datos en un número K de carpetas y se asegura que cada carpeta se utilice como un conjunto de prueba en algún momento. Se utilizo el escenario de validación cruzada de 5 partes (K = 5), ya que es una cifra utilizada en diversos artículos que abordan el problema de Atribución de Autoría (Labbe, 2017) (Eugenia B. Bortolotto, 2020). En la primera iteración, la primera carpeta se usa para probar el modelo y el resto se usa para entrenar el modelo. En la segunda iteración, la segunda carpeta se utiliza como conjunto de prueba, mientras que el resto sirve como conjunto de entrenamiento. Este proceso se repite hasta que cada una de las 5 carpetas se haya utilizado como conjunto de prueba.

Es importante aclarar que el número total de prueba tomada para cada predicción, se realiza de manera aleatoria. En cada iteración el método no toma la misma cantidad de ejemplos por autor. Esta es la razón que en cada clasificador los autores tienen distintos números de ejemplos para pruebas.

Las Matrices o modelos obtenidos con bolsa de palabras tienen una gran cantidad de variables o dimensiones. La reducción de dimensionalidad es el proceso de reducir a una mínima cantidad, el número total de dimensiones que existen en modelo espacio vectorial. Para ello, se utilizaron dos de las técnicas más populares en esta categoría: Análisis de Componentes Principales (PCA, Principal Component Analysis) y Análisis Semántico Latente (LSA, Latent Semantic Analysis).

El Análisis de Componentes Principales reduce la dimensionalidad de un conjunto de datos transformando a un conjunto nuevo de variables denominados componentes principales, que no están correlacionados y que están ordenados de modo que los primeros conserven la mayor parte de la variación presente en todas las variables originales. En PCA, el algoritmo encuentra una representación de baja dimensión de los datos mientras retiene la mayor cantidad de variación posible (Lloret, 2015).

El Análisis Semántico Latente (LSA) es método para extraer y representar el significado de uso contextual de las palabras mediante cálculos estadísticos aplicados a un gran corpus de texto (Jaber, 2012). LSA utiliza el modelo de bolsa de palabras (**B**ag **O**f **W**ord), que da como resultado una matriz de documentos de términos (aparición de términos en un documento). Las filas representan términos y las columnas representan documentos, este transformador realiza una reducción de la dimensionalidad lineal mediante la descomposición de valores singulares truncados (SVD). Es muy similar a PCA, pero opera en vectores de muestra directamente, en lugar de en una matriz de covarianza. Esto significa que puede trabajar con matrices dispersas de manera eficiente. El número de componentes elegidos para PCA y LSA fue de dos dimensiones.

3. Experimentación y resultados.

Mediante la representación con bolsa de palabras, se obtuvieron las palabras más frecuentes utilizadas por el conjunto de autores utilizando un método llamado FreqDistVisualizer() implementado en NLTK[4] de Python. La

Figura ***2*** muestra las 50 palabras más frecuentes utilizadas por los cuatro autores. Se observa que estas pertenecen a la categoría de palabras vacías (Stop Words). También se muestra el vocabulario formado por 16,214 palabras de un total de 177,977. Además, se muestran 8,202 palabras que fueron utilizadas solo una vez (o dos veces) dentro todo el corpus de 38 textos. A este tipo de palabras se le conoce como hápax.

[4] https://www.nltk.org/

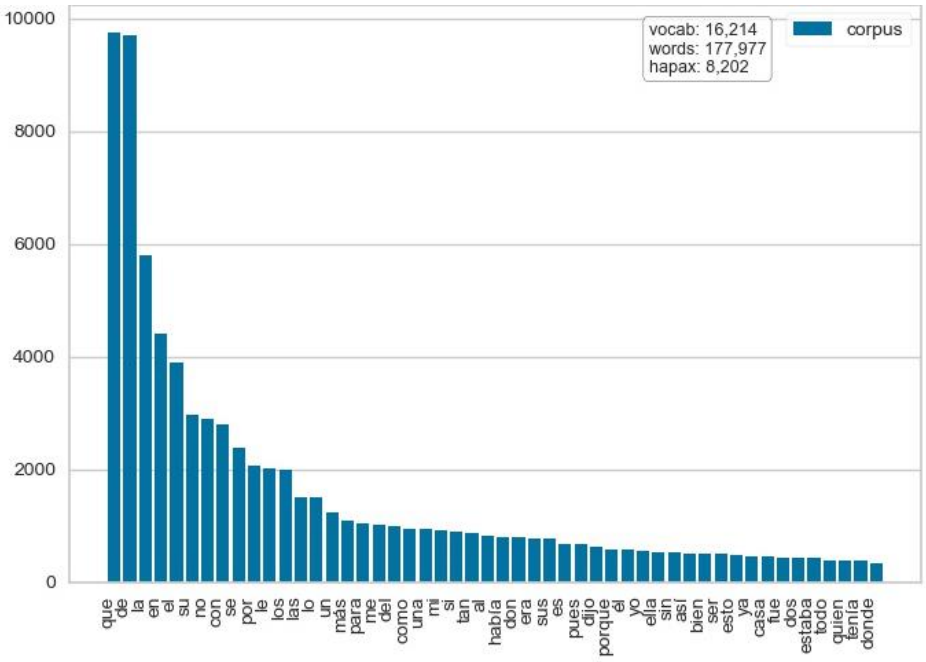

Figura 2. Palabras más frecuentes entre los 4 autores.

Parte 1.

La primera etapa de los experimentos consistió en evaluar las representaciones de matrices binarias, frecuencias y tf-idf en los textos de novelas completas, medias novelas y tercio de novelas.

La Tabla 3 muestra el número de ejemplos que fueron clasificados correctamente para cada uno de los autores. Se observa que en las matrices binarias los autores María de Zayas y Mariana de Carvajal son clasificados correctamente por los cuatro algoritmos de aprendizaje. Mientras que Alonso de Castillo Solorzano y Miguel de Cervantes Saavedra son reconocidos por tres de los cuatro clasificadores. En las matrices de frecuencia se destaca que el autor Mariana de Carvajal tuvo resultados correctos en el clasificador RF, mientras que el autor Miguel de Cervantes Saavedra tiene resultados positivos en todos los clasificadores. En la representación tf-idf se destaca el autor Castillo muestra mejores resultados, los cuatro clasificadores arrojan una mayor efectividad que los demás autores.

Tabla 3. Bolsa de palabras de novelas completas.

Binaria					
Autores	Prueba	KNN	RF	SVM	MNB
Castillo	3	0	3	3	2
Zayas	1	1	1	1	1
Carvajal	1	1	1	1	1
Cervantes	3	0	3	3	3
Frecuencia					
Autores	# de ejemplos	KNN	RF	SVM	MNB
Castillo	3	1	0	1	1
Zayas	2	2	2	2	1
Carvajal	1	0	1	0	0
Cervantes	2	2	2	2	2

tf-idf					
Autores	# de ejemplos	KNN	RF	SVM	MNB
Castillo	2	2	2	2	2
Zayas	3	3	3	0	1
Carvajal	1	1	1	0	1
Cervantes	2	0	2	1	2

La Tabla 4 muestra los resultados de los experimentos realizados con novelas divididas por la mitad, obteniendo un total de 16 textos de pruebas. El número de textos de prueba por autor automáticamente se genera de forma aleatoria. Se observa una mejora en la clasificación correcta de las novelas de cada autor. La representación de matriz de frecuencia muestra mejores resultados que el resto de las representaciones, cabe destacar que los autores Alonso de Castillo Solorzano y María de Zayas obtienen buenos resultados en las tres representaciones matriciales.

Tabla 4. Bolsa de palabras de Medias Novelas.

Binaria					
	# de ejemplos	KNN	RF	SVM	MNB
Castillo	2	1	2	2	2
Zayas	5	5	5	5	4
Carvajal	5	5	3	5	3
Cervantes	4	0	4	4	4
Frecuencia					
	# de ejemplos	KNN	RF	SVM	MNB
Castillo	4	4	4	4	3
Zayas	4	4	4	4	4
Carvajal	2	2	2	2	2
Cervantes	6	2	5	6	6
tf-idf					
	# de ejemplos	KNN	RF	SVM	MNB
Castillo	7	7	7	0	4
Zayas	1	1	1	1	1
Carvajal	4	2	3	0	4
Cervantes	4	4	4	3	3

La Tabla 5 muestra los resultados obtenidos de los experimentos realizados de las novelas divididas en 3 partes, tomando un total de 23 textos de prueba de forma aleatoria. De forma general, todos los clasificadores muestran mejores resultados que los experimentos anteriores. Po otra parte, los resultados son muy similares en las tres representaciones, destacando que el autor Mariana de Carvajal muestra ser muy consistente en las tres representaciones obteniendo un número mayor de aciertos respecto al resto de los autores. Se observa una tendencia: al utilizar bloques de medias novelas y tercia de novelas los resultados muestran un mejor número de acierto en cada clasificador.

Tabla 5. Bolsa de palabras de Tercia de Novelas.

Binaria					
	# de ejemplos	KNN	RF	SVM	MNB
Castillo	6	0	5	6	6
Zayas	7	4	6	7	6
Carvajal	5	5	5	5	5
Cervantes	5	0	5	5	5
Frecuencia					
	# de ejemplos	KNN	RF	SVM	MNB
Castillo	11	9	9	11	9
Zayas	3	3	3	3	3
Carvajal	2	2	2	2	2
Cervantes	7	6	6	7	7
tf-idf					
	# de ejemplos	KNN	RF	SVM	MNB
Castillo	6	6	6	6	6
Zayas	6	6	6	6	6
Carvajal	5	5	5	3	4
Cervantes	6	1	5	5	6

La Tabla 6 muestra la matriz de confusión que se obtuvo para el clasificador KNN en su representación frecuencial, utilizando medias de novelas. los autores María de Zayas y Mariana de Carvajal obtienen una mayor precisión que el resto de los autores, de acuerdo con la diagonal principal el clasificador alcanza un 61% de exactitud.

Tabla 6. Matriz de confusión de KNN con Frecuencia.

	Predicción			
Actual	Castillo	Zayas	Carvajal	Cervantes
Castillo	**2**	1	1	0
Zayas	0	**4**	0	0
Carvajal	0	0	**7**	0
Cervantes	1	2	4	**1**

En la

Figura ***3*** se muestra de forma general los promedios de exactitud obtenidos por los clasificadores en los distintos tamaños de textos y representaciones de matrices. Se observa que los clasificadores RF y MNB son persistentes con buenos resultados en cualquier bloque de novelas y representación. La representación de frecuencia las novelas divididas en dos y tres partes presentaron mejores resultados que las novelas completas, algo similar ocurre con matrices tf-idf, a excepción de RF donde las novelas completas fueron superiores.

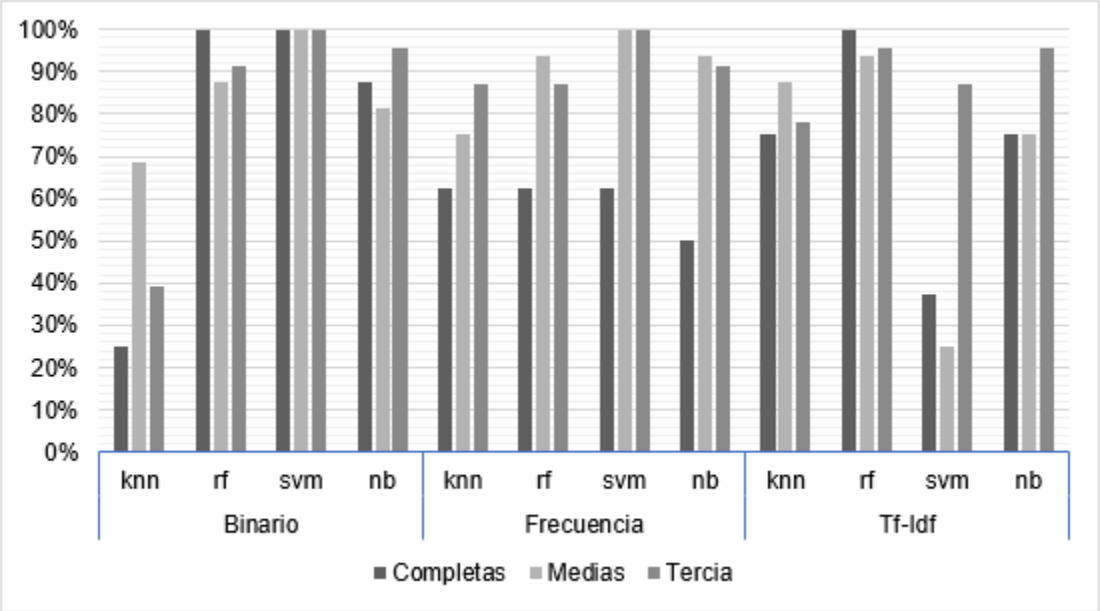

Figura 3. Resultados generales de la bolsa de palabras en sus distintos escenarios.

Parte 2:

Las matrices obtenidas con bolsa de palabras contaban con 14,909 dimensiones (características). Se procedió a aplicar las técnicas de reducción de dimensiones PCA y LSA para observar si la exactitud de los clasificadores mostraba una mejoría.

La Tabla 7 muestra los resultados de PCA y LSA utilizando únicamente 2 componentes (dimensiones) con respecto a bolsa de palabras. Cabe destacar que, las novelas completas mostraron una mejora de exactitud en el clasificador MNB y PCA, pasando de un máximo de 75% a un 88% de exactitud, mientras que en el clasificador KNN se mantuvo en 75%. Otro aspecto a destacar de las novelas completas es que en el clasificador SVM mostro una mejora de resultados tanto en algoritmos de PCA como en LSA, pasando de una exactitud de 25% al 50% en ambos casos. En novelas divididas en dos partes hay un caso en particular que destacar, el clasificador de SVM tuvo una mejora de exactitud pasando de 25% a un 50% en PCA, y en LSA alcanzó los 56%. Sin embargo, en novelas divididas en tres partes los algoritmos de reducción produjeron una disminución en la exactitud bastante notoria. En términos generales el algoritmo de reducción PCA obtuvo mejores resultados de exactitud con respecto al algoritmo de reducción de LSA.

Tabla 7. Bolsa de palabras vs (PCA y LSA) en matrices tf-idf.

Tamaño de texto	Bolsa de Palabras				PCA				LSA			
	KNN	RF	SVM	NB	KNN	RF	SVM	NB	KNN	RF	SVM	NB
Completas	**75%**	**100%**	38%	**75%**	**75%**	**88%**	50%	**88%**	**63%**	25%	50%	**63%**
Medias	**88%**	**94%**	25%	**75%**	**63%**	**63%**	50%	**69%**	56%	50%	56%	44%
Tercia	**78%**	**96%**	**87%**	**96%**	**74%**	**70%**	30%	**74%**	26%	52%	30%	39%

Parte 3.

Con la finalidad de anular la influencia del tópico de los textos se aplicó la sustitución de Entidades Nombradas junto con la eliminación de las palabras vacías utilizando matrices de tf-idf. La Tabla 8 muestra una comparación de los resultados obtenidos en bolsa de palabras respecto a la técnica utilizada para eliminar las palabras vacías junto con las

entidades nombradas, destacando que en las novelas completas hubo una disminución de exactitud en tres de los cuatro algoritmos y que solo KNN se mantuvo con 75%. Las medias novelas se destacó un aumento en el clasificador de SVM, pasando de un 25% a un 94% de exactitud. Mientras que en las novelas divididas en tres partes hubo un ligero aumento de exactitud en los clasificadores SVM y MNB, pasando de 87% a 96% y de un 96% a un 100% respectivamente. De manera general se observaron resultados notables en medias novelas y tercia de novelas dentro de sus respectivas categorías.

Tabla 8. Eliminación de palabras vacías y NER con tf-idf.

Tamaño de texto	Bolsa de palabras				Sin stopword y NER			
	knn	rf	svm	nb	knn	rf	svm	nb
Completas	**75%**	**100%**	38%	**75%**	**75%**	**88%**	38%	50%
Medias	**88%**	**94%**	25%	**75%**	56%	**69%**	**94%**	56%
Tercia	**78%**	**96%**	**87%**	**96%**	43%	**87%**	**96%**	**100%**

Conclusiones.

Cada estudio de atribución de autoría es único, ya múltiples factores influyen en la efectividad del método propuesto: el estilo de escritura utilizado en la redacción de una novela no es el mismo para la redacción de un correo electrónico o un poema. Por otro lado, las novelas, correos y poemas tienen diferente extensión en cuanto al número de palabras. Con base en los textos y métodos utilizados en esta investigación se concluye lo siguiente:

Respecto a los distintitos tamaños del texto la longitud máxima en promedio fue de 5,000 palabras y la cantidad mínima de palabra que se utilizó para los experimentos tienen un aproximado de 1,666. Es recomendable aplicar esta estrategia para aumentar el número de textos disponibles y determinar de forma experimental la mínima cantidad de palabras requeridas para un caso de estudio en particular. No es posible generalizar la cantidad mínima de texto en palabras requeridas que sea adecuada en todos los casos de atribución de autoría.

No es recomendable utilizar un solo algoritmo de clasificación ya que ese algoritmo puede tener buenos resultados con algún autor y resultar poco eficientes en otros. Se sugiere utilizar una combinación de dos o más clasificadores usando estrategias como el Boosting que consiste en crear una regla de predicción altamente precisa combinando muchas reglas relativamente débiles e imprecisas. La idea fundamental detrás de Boosting consiste en elegir conjuntos de entrenamiento para el algoritmo de aprendizaje base de tal manera que este obligue a inferir algo nuevo en los datos cada vez que se lo llame (Mayhua López, 2013). Los algoritmos de aprendizajes aquí evaluados mostraron resultados diferentes en los distintos escenarios de Novelas completas, medias novelas y tercio de novelas, usando las técnicas de Bolsa de Palabras y en la eliminación de las stopword o palabras vacías junto con las sustituciones de las entidades nombradas.

La reducción de dimensiones debe aplicarse con reservas, ya que no siempre se obtiene una mejora en la métrica aplicada. En ocasiones es mejor dejar las matrices tal y como se generaron con muchas dimensiones. Los algoritmos Análisis de Componentes Principales (PCA) y Análisis Semántico Latente (LSA) no eliminan características. En su lugar, estos algoritmos transforman el conjunto de características a uno nuevo de menores dimensiones. Estas técnicas adolecen del mismo problema de la selección por frecuencia: establecer cuál es el número apropiado de dimensiones del nuevo modelo. La reducción excesiva puede disminuir el desempeño de un modelo de aprendizaje. Con base en los resultados obtenidos, se recomienda que, se deben considerar todas las características del modelo y utilizar la reducción de dimensiones para fines de representación gráfica.

En un estudio de atribución de autoría tampoco es recomendable una sola característica estilométrica, como por ejemplo bolsa de palabras. En su lugar se sugiere extraer distintos tipos de características de los textos como por ejemplo etiquetas POS, lematizado, n-gramas de palabras, de caracteres, entre otros.

Por otro lado, las matrices de frecuencias aparentemente registran mejor el estilo de escritura de un autor, ya que muestran las palabras a las que el autor recurre con más o menos frecuencia. Dado que las palabras más utilizadas pueden tener mayor peso que las menos utilizadas es recomendable aplicar la normalización, que consiste en convertir todo el texto en el mismo caso (superior o inferior), eliminando la puntuación, convirtiendo los números a sus equivalentes de palabras, y así sucesivamente (Leskovec, 2014), de manera general la normalización pone todas las palabras en igualdad, y permite que el procesamiento proceda de manera uniforme.

Las entidades nombras hacen referencia a los nombres de personas, lugares, organizaciones, etc. La detección automática de este tipo de palabras es muy importante en el Procesamiento del Lenguaje Natural y en particular para la atribución de autoría. Las entidades nombras se detectan para reemplazarlas por sus respectivas etiquetas, logrando así disminuir el número de dimensiones y la influencia del tópico de documento.

Referencias bibliográficas.

Corney, M. W. (2001). Identifying the authors of suspect email.

David Nadeau, S. S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26.

Eugenia B. Bortolotto, G. S. (2020). Modelos y métodos estadísticos para la estimación de frío y calor en duraznos.

Jaber, T. A. (2012). Enhanced approach for latent semantic indexing using wavelet transform. IET Image Processing, 6(9), 1236-1245.

Jiménez, V. G. (2010). Distribuciones de clases no balanceadas: Métricas, Análisis de Complejidad y Algoritmos de Aprendizaje. Universitat Jaume I.

Labbe, S. A. (2017). Determinación de autoría por medio de. Tesis Doctoral. Pontificia Universidad Catolica De Valparaiso.

Leskovec, J. a. (2014). Mining of Massive Datasets Cambridge University Press.

Lloret, O. A. (2015). Estudio de la influencia de incorporar conocimiento léxico-semántico. Análisis de Componentes Principales para la generación de resúmenes multilingües.

Luyckx, K. a. (2008). Authorship attribution and verification with many authors and limited data. Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), 513--520.

Mayhua López. (2013). Elementos locales en conjuntos de clasificadores diseñados por" Boosting".

Morales, R. M. (2007). Clasificación Automática de Textos considerando el Estilo de Redacción. Instituto Nacional de Astrofísica, Óptica y Electrónica.

Rueda, J. M. (2016). El análisis estilométrico aplicado a la literatura española: las novelas policíacas e históricas. Caracteres: estudios culturales y críticos de la esfera digital 5.2, 196-245.

Sanderson, C. a. (2006). Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 482--491.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology, 60: 538–56.

Toledo, G. R. (2019). Detección automática de cambio de estilo de escritura utilizando aprendizaje automático. Tesis de Doctorado. Tecnológico Nacional de México, Cuernavaca, Morelos, México.

Información de los autores.

Favián Gutiérrez Constantino, estudiante de Ingeniería en Sistemas Computacionales en el Instituto Tecnológico de Tuxtla Gutiérrez, sus áreas de interés se relacionan con las ciencias de la matemática, ciencias de la computación, redes de computadoras y el análisis de datos.

Luis Ángel Vázquez Moreno, estudiante de ingeniería en sistemas computacionales en el Instituto Tecnológico de Tuxtla Gutiérrez, sus áreas de interés se relacionan con el desarrollo de software, aplicaciones móviles y análisis de datos.

Germán Ríos Toledo obtuvo el grado de Doctor en Ciencias de la Computación en 2019 por el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) en Cuernavaca, Morelos, México. Actualmente, es profesor de tiempo completo en el Departamento de Computación del Tecnológico Nacional de México (campus Tuxtla Gutiérrez, Chiapas) en la Ingeniería en Sistemas Computacionales y en la Maestría en Ciencias en Ingeniería Mecatrónica. Su área de especialización es el Procesamiento del Lenguaje Natural, particularmente en el uso de información sintáctica como una característica para el análisis de estilo de escritura. Otras áreas de su interés incluyen el procesamiento y análisis de imágenes, audio y video por medio de Algoritmos de Aprendizaje Automático y Aprendizaje Profundo. Miembro del Sistema Nacional de Investigadores, nivel C (2024-2021).

Héctor Guerra Crespo es egresado del I. T. de Mérida (Yucatán, México) de la carrera de Ingeniería en Sistemas Computacionales en 1994, es Doctor en Sistemas Computacionales por la Universidad del Sur (Chiapas, México) en 2011. Es profesor en el área de Ingeniería en Sistemas Computacionales del Instituto Tecnológico de Tuxtla Gutiérrez y en el área de Licenciatura en Sistemas Computacionales de la Universidad Autónoma de Chiapas, en ambas desde 1995. Es miembro del Claustro Doctoral "Doctorado en Ciencias de la Ingeniería" perteneciente al Programa Nacional de Posgrados de Calidad, I.T. de Tuxtla Gutiérrez, desde 2016. Miembro del Sistema Nacional de Investigadores, nivel C (2024-2021). www.hectorguerracrespo.com.

Aída Guillermina Cossío Martínez es Maestra en Ciencias en Administración por el Instituto Tecnológico de Tuxtla Gutiérrez en 2002. Es profesora de tiempo completo del área de Ingeniería en Sistemas Computacionales del Instituto Tecnológico de Tuxtla Gutiérrez, desde 1994. Se especializa en la formulación y evaluación de proyectos, así como el emprendimiento y desarrollo de planes de negocio, actualmente es perfil deseable y trabaja en la línea de investigación Tecnología de Información y Base de Datos.

María Guadalupe Monjarás Velasco, Obtuvo el grado de Doctor en Sistemas Computacionales en 2012, el grado de Maestra en Ciencias de la Computación con especialidad en Sistemas de Información y Bases de Datos en 2009, terminó la carrera de Ingeniería en sistemas Computacionales en el año 2006, actualmente es Jefa del Departamento de Sistemas y Computación del I. T. de Tuxtla Gutiérrez desde 2016, asesora proyectos en concursos académicos. Miembro del comité Académico Del ITTG.