Atribución de autoría mediante clasificación automática en corpus multicontexto.

Omar González Brito [A]., Aide Guadalupe Mares González [B], Laura Cleofas-Sánchez [C]

[A] Tecnológico de Estudios Superiores de Tianguistenco, Carretera Tenango, Santiago-La Marquesa 22, 52650 Santiago Tilapa, Tianguistenco, Estado de México, omar_g@test.edu.mx, Autor A.
[b] Tecnológico de Estudios Superiores de Tianguistenco, Carretera Tenango, Santiago-La Marquesa 22, 52650 Santiago Tilapa, Estado de México, aide_201823061@test.edu.mx, Autor B.
[c] Tecnológico de Estudios Superiores de Tianguistenco, Carretera Tenango, Santiago-La Marquesa 22, 52650 Santiago Tilapa, Tianguistenco, Estado de México, laura_cs@test.edu.mx, Autor C.

Resumen

Los estudios de atribución de autoría han implementado diversas técnicas de inteligencia artificial y métodos de procesamiento del lenguaje natural para obtener el estilo de escritura de un autor y así poder determinar la atribución de autoría de obras literarias, estudios científicos, notas periodísticas, revistas, blog, entre otros [1][2][3]. Sin embargo, aunque las técnicas aplicadas a esta tarea han sido diversas, todos los estudios tienen algo en común: el corpus, la revisión de la literatura muestra que han utilizado corpus de un solo contexto, es decir, conformados por noticias, tweets, correos electrónicos, o textos de foros, entre otros. De lo anterior, se puede observar que un área de oportunidad en la tarea de atribución de autoría es el uso de corpus multicontexto, en la presente investigación se desarrolla un corpus multicontexto y un sistema de clasificación automática de textos, utilizando como método de aprendizaje supervisado máquina de soporte vectorial y regresión logística. Se analizaron diferentes contextos como Blogs, Periódicos y Twitter los cuales forman parte de los contenidos corpus, se presentaron diversos resultados, el más significativo supera el 70% de exactitud. Los resultados obtenidos en la presente investigación demuestran que es posible determinar la atribución de autoría con diferente contexto siendo esta un área de oportunidad en la tarea de atribución de autoría.
.

Palabras clave—Atribución de autoría, Aprendizaje automático, Corpus multicontexto, Clasificación automática de textos

Abstract

Authorship attribution studies have implemented various artificial intelligence techniques and natural language processing methods to obtain an author's writing style and thus determine the authorship attribution of literary works, scientific studies, journalistic notes, magazines, blogs, among others [1][2][3]. However, although the techniques applied to this task have been diverse, all the studies have something in common: the corpus, the review of the literature shows that they have used corpora from a single context, that is, made up of news, tweets, emails, etc. emails, or forum texts, among others. From the above, an area of opportunity in the task of authorship attribution is the use of a multi-context corpus. In the present investigation, a multi-context corpus and an automatic text classification system are developed using supervised machine learning as a method. vector support and logistic regression. Different contexts such as Blogs, Newspaper and Twitter which are part of the corpus contents were analyzed, various results were presented, the most significant exceeds 70% accuracy. The results obtained in this research show that it is possible to determine the attribution of authorship with different contexts, this being an area of opportunity in the task of attributing authorship.

Keywords— *Authorship Attribution, Machine Learning, Multi-Context Corpus, Automatic classification of text*

1. INTRODUCCIÓN

La atribución de autoría se ha convertido en una tarea objetivo para la inteligencia artificial, específicamente para las áreas como la lingüística computacional y forense [4] algunos de los primeros problemas relacionados con atribución de autoría son: La obra del Quijote de Avellana, las obras de Shakespeare, y las obras de Nicolás Maquiavelo, estas han sido cuestionadas en su autoría [5].
Los autores al escribir dejan patrones de escritura propios como una huella digital, a través de esta es posible adjudicar hábitos personales subconscientes, que permite identificar el estilo de escritura de cada uno [6]. La atribución de autoría requiere del análisis de características textuales para poder determinar la autoría de un documento [7][8].

Las investigaciones de atribución de autoría abarcan una gran variedad de contextos para su estudio. Sin embargo, cada una de estas considera un solo contexto, por ejemplo; Escalante utiliza un corpus que está comprendido por documentos de 10 autores que se encuentran dentro de la categoría de noticias corporativas e industriales [8] por otra parte, [7] analiza la atribución de autoría de textos cortos, por medio de un corpus de 600 tweets en español, en otra investigación presentada por [9] analizó la atribución de autoría mediante el corpus de correo electrónico Enron. En estos estudios se implementan diferentes técnicas de inteligencia artificial y métodos del procesamiento de lenguaje natural analizando un corpus de un solo contexto. En la presente investigación se desarrolló un corpus multi contexto que considera contextos tales como blogs, notas periodísticas, y tweets, este se analizó mediante el desarrollo de un método de clasificación automático que implementa como método de aprendizaje supervisado máquina de soporte vectorial y regresión logística.

2. TRABAJOS RELACIONADOS

La revisión de la literatura muestra que los trabajos realizados utilizan corpus de un solo contexto, el contenido de estos corpus va desde notas periodísticas, tweets, blog, correos electrónicos, ejemplos de estos trabajos se describen continuación: en [10] utilizaron un corpus compuesto por

artículos de notas periodísticas, denominado URDU compuesto por 4800 documentos escritos por 12 columnistas , con 400 documentos por autor, formando un total de 5, 631,850 palabras que permiten capturar el estilo de escritura de cada uno de los autores, implementaron un enfoque de Latent Dirichlet Allocation (LDA) de n-gramas hasta nivel 5, combinado con una medida de similitud de coseno, esto les permitió obtener una precisión de 93.17%. En otra investigación presentada por [11] emplean 3 lematizadores (GOLD, Khoga y Light 10) para un corpus en idioma árabe, conformado por 2400 artículos periodísticos escritos por 97 autores diferentes, los resultados obtenidos muestran un 67% de precisión para el lematizador Light 10, 64% para Khoga, 61% para Gold y 78% para conjunto de datos sin lematizar, de lo anterior los autores infieren que la lematización influye para la determinación de autoría.

Otro contexto que ha sido utilizado para el análisis de atribución de autoría son los tweets, algunas investigaciones donde utilizan este tipo de corpus son presentadas en [12] donde el corpus estuvo compuesto por un conjunto de 600 tweets en español de 6 autores, realizaron dos pruebas, con 300 y 600 implementaron un modelo con representación de n-gramas y algoritmos de aprendizaje como la Máquina de Soporte Vectorial (SVM), Naive Bayes y un árbol J48, obteniendo una exactitud de 70% con una representación de 2-gramas y SVM. Por su parte [13] utiliza un corpus compuesto por 9000 tweets, implementando una red neuronal y una representación de n-gramas de caracteres, donde incluyen 3 módulos: incrustación de caracteres, la capa convolucional y la función de activación SoftMax, la red neuronal obtiene mejores resultados al incrementar el número de tweets por autor donde con 500, la red obtiene una precisión de 72.4% seguido de 66.5% para 200 tweets por autor.

Finalmente, en [14] crean un corpus utilizando textos de blog bengalíes, entre los blogs esta SaltedBadam, DeshKaal y LaljiperDiary, preprocesaron los textos y como resultado obtuvieron 3000 muestras aleatorias de texto de los 3 blogs, creando así 1000 para cada etiqueta, de las cuales 750 formaron parte del conjunto de entrenamiento y 250 muestras para el conjunto de validación, equilibraron el corpus a través de la selección de 25 párrafos al azar de entre los blogs. Analizaron características de tipo léxico bajo 3 representaciones: Frecuencia del término (*TF*), Frecuencia del término Inversa a la Frecuencia del Documento (*TF-IDF*) y una representación booleana. Como método de aprendizaje supervisado emplearon 3 clasificadores: Naive Bayes (*NB*), Máquina de Soporte Vectorial (*SVM*), árboles de decisión, un Perceptrón Multicapa, los resultados que obtuvieron muestran que *SVM* y el Perceptrón multicapa obtiene una precisión superior al 90%, el bigrama de 500 letras y números obtiene el mayor valor de precisión obteniendo una precisión del 99.47%.

En la Tabla 1 se pueden observar los diferentes corpus que han abordado el problema de atribución de autoría en diferentes contextos, como se puede observar el número mínimo de autores que consideran son 6 autores y el máximo de autores son 20, a través de la revisión de la literatura se observa que el número de autores entre los que se disputa la autoría de un documento es entre 6 y 20. Tomando como referente la literatura, se determina el número de autores y textos utilizados para la construcción del corpus utilizado en la presente investigación.

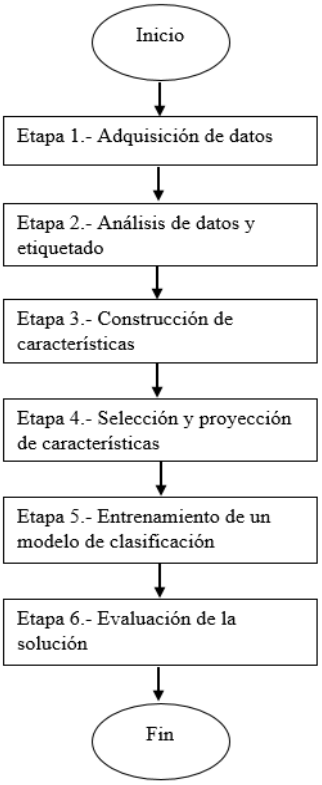
Tabla 1 Descripción de corpus en diferentes contextos.

Corpus	Número de Autores	Número de Documentos por autor	Contexto
C10 [15]	10	50	Noticias corporativas e industriales
PAN12 [15]	14	3	Novelas
Guardia 10 [16]	13	100	Artículos de opinión y reseñas de libros
Corpus de tweets [17]	6	100	Tweets
Greek blog [18]	20	50	Blogs

3. METODOLOGÍA

Para el desarrollo de la investigación donde se analizará el corpus multicontexto, se utilizó la metodología de clasificación de textos propuesta por [15],[19], las etapas de esta se pueden observar en la Figura 1.

Figura 1. Etapas de la metodología de clasificación de textos



3.1 ADQUISICIÓN DE DATOS.

Para la creación del corpus multicontexto se estableció que los textos de los autores fueran de diferentes contextos, por lo que se consideraron autores que realizaran notas periodísticas, que tuvieran una cuenta en Twitter y un blog, en un principio se consideraron 83 autores, sin embargo, no todos contaban con los diferentes contextos que se plantearon, por lo que solo se consideraron los autores que contaran con documentos en los tres contextos y que tuvieran un mayor número de notas periodísticas, notas en su blog, y tweets publicados. Quedando 15 autores con tres contextos diferentes, donde en la Tabla 2 se pueden apreciar.

Tabla 2 Número de documentos por contexto del corpus multicontexto.

Autores	Número de documentos		
	Notas periodísticas	Blog	Twitter
1	25	17	50
2	25	17	50
3	25	17	50
4	19	17	50
5	25	17	50
6	25	17	50
7	25	17	50
8	13	17	50
9	25	17	50
10	25	17	50
11	25	17	50
12	25	17	50
13	8	17	50
14	15	17	50
15	9	17	50

3.2 ANÁLISIS Y ETIQUETADO DE DATOS.

Los modelos de representación utilizados fueron n-gramas de tamaño 1,2, y 3 de palabras, y n-gramas de tamaño 2,3, y 4 de carácter, no se realizó ningún tipo de pre procesamiento.

3.3 CONSTRUCCIÓN Y PONDERACIÓN DE CARACTERÍSTICAS.

En esta etapa se utilizó la ponderación booleana o binaria la cual consiste en asignar un valor a una característica dentro de un documento. Si la característica se encuentra presente en el documento se asigna el valor de 1 por el contrario se asigna 0.

3.4 ENTRENAMIENTO DEL MODELO.

El modelo se construyó a partir de métodos de aprendizaje supervisado, para esta investigación se utilizaron métodos como Máquina de Soporte Vectorial (*SVM*), Regresión logística (*RL)*, Multinomial Naive Bayes (*NB*). Los parámetros de Máquina de Soporte Vectorial fueron un kernel lineal, el parámetro C igual a uno, utilizando una clasificación de uno contra todos.

3.5 EVALUACIÓN DE LA SOLUCIÓN

La métrica de exactitud es la que se utilizó para evaluar el método propuesto. Esta consiste en el porcentaje de instancias que se clasifican correctamente. Se define en términos de verdaderos positivos (VP), Falsos positivos (FP), Verdaderos negativos (VN) y falsos negativos (FN) como se muestra en la ecuación 1[20]:

$$Exactitud = \frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{VP} + \mathrm{VN} + \mathrm{FP} + \mathrm{FN}} \quad (1)$$

4. EXPERIMENTACIÓN Y RESULTADOS

Los experimentos planteados fueron 3 donde se utilizaron diferentes contextos para el entrenamiento y validación con la finalidad de determinar la autoría a partir de corpus conformado por diferentes contextos. En los experimentos se analizaron las representaciones de palabras y caracteres mediante el modelo de n-gramas. En el primer experimento se consideró al contexto de blog como la muestra de entrenamiento y al contexto de periódico y Twitter como la muestra de validación de forma separada, analizando diferentes representaciones de n-gramas de palabras y caracteres, los resultados que se obtuvieron se muestran en la Tabla 3 y 4.

Tabla 3 Muestra de entrenamiento blog y muestra de validación periódico

Método	Palabras			Caracteres		
	1 grama	2 grama	3 grama	2 grama	3 grama	4 grama
Máquina de soporte vectorial	**80.19%**	72.2%	45.36%	55.91%	68.37%	**69.96%**
Multinomial Naive Bayes	67.7%	61.98%	58.78%	46%	38%	16.61%
Regresión logística	76.35%	66.13%	29.07%	52.39%	61.66	64.53%

Tabla 4 Muestra de entrenamiento blog y muestra de validación Twitter

Métodos	Palabras			Caracteres		
	1 grama	2 grama	3 grama	2 grama	3 grama	4 grama
Máquina de soporte vectorial	10.01%	10.01%	7.20%	6.80%	8.01%	10.54%
Multinomal Naive Bayes	**34.57%**	**34.57%**	22.83%	12.41%	32.04%	**35.64%**
Regresión logística	10.28%	10.28%	6.94%	6.80%	7.87%	8.41%

Como se puede observar en la Tabla 3, el modelo de 1-grama de palabras obtiene una exactitud de 80.19% y 69.96% con el

modelo de 4-gramas de carácter, esto nos indica si es posible determinar la autoría a partir de diferentes contextos. Sin embargo, cuando la muestra de validación son textos cortos el rendimiento del sistema no supera más del 35.64% de exactitud.

En el segundo experimento se consideró al contexto de periódico como la muestra de entrenamiento y al contexto de blog y Twitter como la muestra de validación de forma separada cada contexto. Como se puede observar en la Tabla 5 se alcanzó una exactitud de 67.32%, lo que nos sigue indicando que si es posible determinar la autoría a partir de diferentes contextos. Sin embargo, con textos cortos el rendimiento no supera el 25.56% de exactitud.

Tabla 5 Muestra de entrenamiento periódico y muestra de validación blog

Métodos	Palabras			Caracteres		
	1 grama	2 grama	3 grama	2 grama	3 grama	4 grama
Máquina de soporte vectorial	**67.32 %**	49.21 %	22.83 %	52.75 %	62.20 %	**63.77 %**
MultinomialNaive Bayes	41.73 %	41.33 %	48.4%	44.09 %	34.64 %	24.80 %
Regresión logística	59.84 %	43.30 %	11.41 %	55.90 %	57.48 %	53.93 %

Tabla 6 Muestra de entrenamiento periódico y muestra de validación Twitter

Métodos	Palabras			Caracteres		
	1 grama	2 grama	3 grama	2 grama	3 grama	4 grama
Máquina de soporte vectorial	6.67%	6.67%	6.67%	6.67%	6.67%	6.67%
Multinomial Naive Bayes	25.10%	**25.5 %**	21.76%	11.08%	21.36%	25.36%
Regresión logística	6.67%	6.67%	6.67%	6.67%	6.67%	6.67%

En el tercer experimento se consideró al contexto de Twitter como la muestra de entrenamiento y al contexto de blog y periódico como la muestra de validación de forma separada cada contexto. Como se puede observar en las tablas 7 y 8, el rendimiento del método no supera el 47.63%, esto se debe posiblemente a que los textos de Twitter no tienen el tamaño suficiente para poder determinar la autoría.

Tabla 7 Muestra de entrenamiento Twitter y muestra de validación blog

Métodos	Palabras			Caracteres		
	1 grama	2 grama	3 grama	2 grama	3 grama	4 grama
Máquina de soporte vectorial	38.18%	36.61%	38.18%	9.05%	26.37%	35.03%
Multinomial Naive Bayes	**47.63%**	47.24%	38.58%	9.84%	27.95%	21.65%
Regresión logística	28.74%	32.67%	36.61%	8.66%	24.80%	29.13%

Tabla 8 Muestra de entrenamiento Twitter y muestra de validación periódico

Métodos	Palabras			Caracteres		
	1 grama	2 grama	3 grama	2 grama	3 grama	4 grama
Máquina de soporte vectorial	36.73%	30.61%	36.73%	9.18%	23.46%	25.85%
Multinomial Naive Bayes	**46.25%**	43.87%	36.39%	13.26%	23.80%	25.17%
Regresión logística	25.85%	24.82%	37.41%	8.86%	20.06%	24.48%

4. CONCLUSIONES

Como se puede observar en la experimentación cuando el contexto es de blog o notas periodísticas se obtienen resultados de 80.19% y 67.32% de exactitud, esto se debe a que los textos tienen tamaños similares a un que los contextos son diferentes. Sin embargo, con textos cortos no se obtuvieron resultados superiores a 35.64%, esto se debe a que no importa el contexto, pero si influye el tamaño de los textos.

8. REFERENCIAS

[1] F. López, "Donde se muestran algunos resultados de atribución de autor en torno a la obra cervantina". *Revista Colombiana de Estadística*, Vol. 34(1), 15-37, 2011.

[2] V. Mercado, A. Villagra, M.G Leguizamón, & M. Errecalde. "Atribución de autoría y determinación de la orientación política en documentos periodísticos". *In XVII Workshop de Investigadores en Ciencias de la Computación*, 2015.

[3] M. Arellano. "Poetic language and the dissolution of the subject in" La gintanilla" and" El licenciado Vidriera". Calíope: *Journal of the Society for Renaissance and Baroque Hispanic Society*, Vol. 2(2), 1, 1996.

[4] D. Castro, Adame, M. Peláez, & R. Muñoz. "Authorship verification, average similarity analysis". *International Conference Recent Advances in Natural Language Processing*, 84-90, 2015.

[5] R. Álvarez. "La huella del Príncipe de Maquiavelo en la literatura inglesa". *RSEI: Revista de la sociedad Española de Italianistas*, Vol. 9, 19-41, 2013.

[6] J. Blasco, U. Ruiz. "Evaluación y cuantificación de algunas técnicas de atribución de autoría en textos españoles". *Castilla: Estudios de Literatura*, 27-47, 2009.

[7] F. Castillo, J. Martínez, M. Torres, P. Zavala, A. Becerra, & J. Rizzo. "Atribución de autoría de mensajes de Twitter a través del análisis sintáctico automático". *Research in Computing Science*, Vol. 149(11), 91-101, 2020.

[8] H. J. Escalante, T. Solorio, & M. Montes-y Gómez. “Local histograms of character n-grams for authorship attribution”. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 288-298, 2011.

[9] E. Abdallah, A. Abdallah, M. Bsoul, & Otoom. “A Simplified features for email authorship identification.” *International Journal of Security and Networks*, 8(2), 72-81.

[10] W. Anwar, I. Bajwa, & S. Ramzan. (2019). Design and Implementation of a Machine Learning Based Authorship Identification Model. In: Scientific Programming, 1–14, 2013.

[11] O. Abdulfattah, & W. Ibrahim. “The Effectiveness of Stemming in the Stylometric Authorship Attribution in Arabic”. IJACSA: *International Journal of Advanced Computer Science and Applications*, Vol. 11(1), 116-121, 2020

[12] F. Castillo, J. Zavala, M, Sánchez, A. Márquez, I. Morales, & J. Flores. “Aplicación del análisis sintáctico automático en la atribución de autoría de mensajes en redes sociales”. *Research Computing Science*, 109-119, 2017.

[13] P. Sherestha, S. Sierra, & F. González. Convolutional “Neural Networks for Authorship Attribution of Short Texts”. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* Vol. 2, 669-674, 2017.

[14] S. Phani, S. Lahiri, & A. Biswas. “A machine learning approach for authorship attribution for Bengali blogs”. *In 2016* International Conference on Asian Language Processing *(IALP), (IEEE)*, 271-274, 2016.

[15] O. González Brito, J. L. Tapia Fabela, & S. Salas Hernández. “New approach to feature extraction in authorship attribution”. *International Journal of Combinatorial Optimization Problems & Informatics*, Vol. 12(3), 2021.

[16] P. Stamatatos, “On the robustness of authorship attribution based on character n-gram features”. *Journal of Law and Policy*, 21(2), 2013.

[17] F. Velásquez, J. Godoy, M. Falcón, J. De Paz, Chávez, & J. Sierra. Atribución de autoría de mensajes de Twitter a través del análisis sintáctico automático. *Research and Computing Science*, 149(11), 91-101, 2020.

[18] G. Mikros. “Authorship attribution and gender identification in Greek blogs”. *Methods and Applications of Quantitative Linguistics*, 21, 21-32, 2012.

[19] M. Mirończuk, & J. Protasiewicz. “A recent overview of the state-of-the-art elements of text classification.” *In Expert Systems with Applications* Vol. 106, 36-54, 2018.

[20] J. Posadas-Durán, H. Gómez-Adorno, G. Sidorov, I. Batyrshin, D. Pinto, & L. Chanona-Hernández, “Application of the distributed document representation in the authorship attribution task for small corpora”. *In Soft Computing*, Vol. 21, 627-639, 2017.