Selección de características mediante intersección de conjuntos en la tarea de Atribución de Autoría

Omar González Brito [A]., María de los Ángeles Mota Segura [B]., Silvia Salas Hernández [C]., Laura Cleofas-Sánchez [D].

[A] Tecnológico de Estudios Superiores Tianguistenco, Carretera Tenango, Santiago-La Marquesa 22, 52650, Santiago Tilapa, omar_g@test.edu.mx
[B] Tecnológico de Estudios Superiores Tianguistenco, Carretera Tenango, Santiago-La Marquesa 22, 52650, Santiago Tilapa, maria_201823068@test.edu.mx
[C] Tecnológico de Estudios Superiores Tianguistenco, Carretera Tenango, Santiago-La Marquesa 22, 52650, Santiago Tilapa, silvia_sh@test.edu.mx
[D] Tecnológico de Estudios Superiores Tianguistenco, Carretera Tenango, Santiago-La Marquesa 22, 52650, Santiago Tilapa, laura_cs@test.edu.mx

Resumen

Establecer la autoría requiere del análisis de rasgos lingüísticos o características que permitan identificar el estilo de escritura de cada autor, dentro del conjunto de características que describen el estilo de escritura de un autor hay características relevantes, irrelevantes y redundantes, por lo que se requieren métodos de selección de características que mejoren el desempeño del clasificador, en la presente investigación se implementó el método de clasificación de textos que consta de las siguientes etapas; adquisición de datos, análisis de datos y etiquetado, construcción de características y ponderación, selección y proyección de características, entrenamiento de un modelo de clasificación, y evaluación de la solución. En la etapa de Selección y proyección de características se propone la extracción de subconjuntos de características a partir de la intersección de conjuntos, y como se puede observar en la experimentación realizada se obtuvieron mejores resultados en comparación a los métodos de análisis de componentes principales, ganancia de información, puntuación de información mutua, máxima relevancia mínima de redundancia, y *random Forest*.

Palabras clave—Atribución de Autoría, Selección de características, Máquina de Soporte Vectorial, Regresión logística, Naive Bayes.

Abstract

Establishing authorship requires the analysis of linguistic features or features that allow identifying the writing style of each author. Within the set of features that describe the writing style of an author, there are relevant, irrelevant, and redundant features, which is why methods are required. to select features that improve the performance of the classifier, in this research, the text classification method was implemented, which consists of the following stages; data acquisition, data analysis and labeling, feature construction and weighting, feature selection and projection, training a classification model, and solution evaluation. In the feature selection and projection stage, the extraction of subsets of features from the intersection of sets is proposed, and as can be seen in the experimentation carried out, better results were obtained compared to the principal component analysis methods, gain of information, mutual information score, maximum relevance minimum redundancy, and random forest.

Keywords— *Author Attribution, Feature Selection, Support Vector Machine, Logistic regression, Naive Bayes.*

1. INTRODUCCIÓN

El análisis de autoría se ha convertido en un problema importante en áreas como la recuperación de información, la lingüística computacional y la lingüística forense. La atribución de autoría requiere del análisis de características lingüísticas, rasgos lingüísticos o características que permitan identificar el estilo de escritura de cada autor. Esto permite determinar si un documento pertenece o no a un autor entre un conjunto de candidatos [1], [2]. Un claro ejemplo de atribución de autoría se da a través de la literatura, con la determinación del autor del testamento de la roma imperial, es así como en el año de 1440 Lorenzo Valla determino la autoría de dicho documento empleando técnicas como los anacronismos lingüísticos, estilísticos y de contenido [3].

La inteligencia artificial ha abordado la tarea de atribución de autoría, identificando características textuales que los autores plasman en la redacción de sus textos, esta identificación se realiza desde el enfoque de la clasificación de textos en donde una de las principales etapas, es la selección de las características [4]. En [5] y [6] mencionan que en la tarea de clasificación se encuentran características que pueden ser relevantes, irrelevantes y redundantes, estos dos últimas perjudican el rendimiento del clasificador.

En la presente investigación se propone la extracción de características por medio de la intercesión de características, a partir de los documentos de los autores, de esta forma se identificaron las características en común, así el rendimiento del clasificador , los resultados obtenidos son comparados con los siguientes métodos de selección de características; análisis de componentes principales, ganancia de información, puntuación de información mutua, máxima relevancia mínima de redundancia, y bosques aleatorios.

2. MARCO TEORICO

La selección de características es un paso crucial para reducir la dimensionalidad de conjuntos de datos, y al mismo tiempo permite elegir las características más importantes optimizar el rendimiento del clasificador, existen diversos métodos de clasificación, que han sido utilizados en diferentes áreas como el reconocimiento de patrones, aprendizaje automático, minería de datos, análisis estadístico y clasificación de textos, entre otras [7], [8].

Las características que han sido implementadas para autoría son de tipo léxicas, es decir, palabras y caracteres, debido a que en ellos se encuentra contenido el estilo de escritura de un autor [9], [10]. Los n-gramas son una de las técnicas del procesamiento de lenguaje natural que ayuda al tratamiento del texto, estos son definidos como una sub-secuencia de elementos: palabras, fonemas, letras, lemas, caracteres,

etcétera, donde n es el número de elementos que componen la secuencia [11]. La mayoría de las investigaciones analizan un conjunto de características de manera general de todos los autores; sin embargo, no se han extraído subconjuntos representativos de características por autor que permitan describir la autoría del texto. Debido a lo observado en la revisión de la literatura, en esta investigación se propone extraer las características a partir de la intercesión de los documentos de los autores, teniendo un conjunto A y B de documentos, es posible obtener conjuntos que contengan elementos en común, esto se representa matemáticamente de la siguiente manera [12], [13]:

$$A \cap B \tag{1}$$

Los métodos más utilizados para la selección de características son métodos de filtro. Estos seleccionan las características por medio de un criterio de relevancia, que puede ser una medida de distancia o dependencia algunos de los más utilizados son [14]:

2.1 Análisis de Componentes Principales.

En [15] describe el Análisis de Componentes Principales (ACP) como un método algebraico/estadístico que busca sintetizar y estructurar la información de una matriz de datos. Se homologa la matriz a un espacio vectorial para identificar dimensiones que, como combinación lineal de las variables, conserven la varianza total, no tengan correlación entre sí y proporcionen una explicación diferencial y conocida de la varianza total de las variables originales.

2.2 Ganancia de Información.

En [16] define la Ganancia de Información como una medida utilizada en teoría de la información y estadística para seleccionar características relevantes en un conjunto de datos. Se calcula como la diferencia entre la entropía inicial y la entropía condicional después de la selección de características, indicando la capacidad de la característica para reducir la incertidumbre en la clasificación de datos. Matemáticamente es expresada en la formula:

$$IG(x) = \log(n) + \frac{n_x^3}{n^2}\log\frac{n_x}{n} + \frac{n_x^3}{n^2}\log\frac{n_x}{n} \tag{2}$$

Donde:

n es el número total de documentos en el conjunto de datos.

n_x es el número total de documentos que contiene la palabra x.

2.3 Puntuación de Información Mutua.

También en [16] menciona que la Puntuación de Información Mutua (PMI) es una medida que evalúa la asociación entre dos términos de texto. Se calcula a partir de la frecuencia conjunta de aparición de los términos y su frecuencia individual en un corpus. Una puntuación alta indica una fuerte asociación entre los términos cuando la probabilidad de encontrarlos juntos es mayor que la probabilidad de encontrarlos por separado. Esta es expresada matemáticamente de la siguiente forma:

$$PMI(x,y) = log_2\frac{m * n_{xy}}{n_x n_y} \tag{3}$$

Donde:

n es el número total de documentos en la colección.

n_x es el número total de documentos que contiene la palabra x.

n_y es el número total de documentos que contiene la palabra y

n_{xy} es el número total de documentos que contienen ambas palabras x e y.

m es el número total de palabras diferentes en la colección.

2.4. Mínima de Redundancia Máxima Relevancia

En [17] menciona que el método RMR (Mínima Redundancia Máxima Relevancia) se utiliza para seleccionar variables relevantes en la predicción de una variable objetivo, evitando aquellas ya presentes en el modelo. Primero, se identifican las variables más relevantes, y luego se eliminan las redundantes, optimizando así la eficiencia del modelo. La relevancia y la redundancia se representa matemáticamente de la siguiente manera:

$$Relevancia(V_C) = \frac{1}{|S|}\sum_{x_i x_j \in S} I(X_i, Y) \tag{4}$$

$$Redundancia(W_i, X_j) = \frac{1}{|S|^2}\sum_{x_i x_j \in S} \left|I(X_i, Y_j)\right| \tag{5}$$

Donde:

(S) Es el conjunto de variables seleccionadas.

$|S|$ La cardinalidad.

I Una función que mide el grado de relación entre dos variables.

2.5 Random Forest

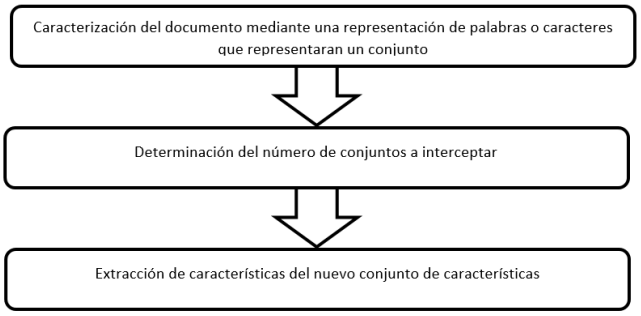
Random Forest es un algoritmo de aprendizaje automático que utiliza múltiples árboles de decisión para mejorar la precisión y reducir la varianza en las predicciones, mediante un enfoque de ensamble [18].

3. METODOLOGÍA

El esquema metodológico para el desarrollo de la investigación estuvo basado en dos partes una para la

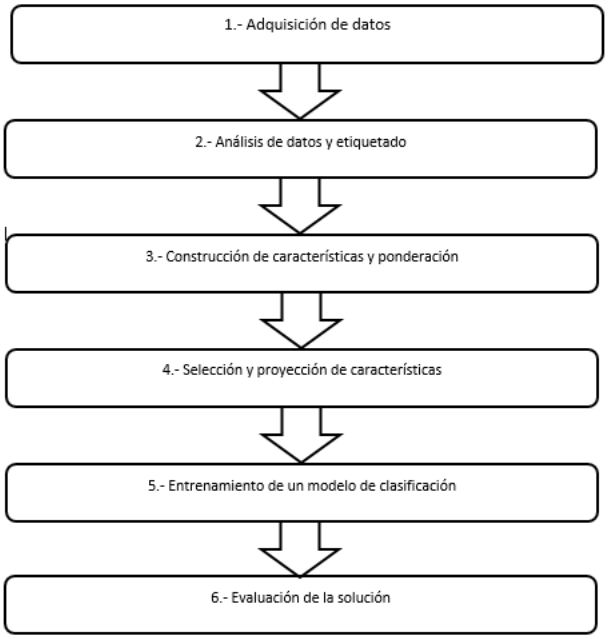
intersección de los conjuntos, la cual se puede observar en la Figura 1.

Figura 1. Esquema metodológico para selección y proyección de características



El esquema metodológico de la presente investigación se observa en la Figura 2.0, este se compone de 6 fases, las cuales se describen a continuación.

Figura 2.0 Esquema metodológico general



3.1 Adquisición de datos

El corpus utilizado para la experimentación fue el corpus C10, este se puede obtener de la página web de PAN (https://pan.webis.de/), el cual se encuentra integrado por documentos del corpus REUTERS volumen 1, este está conformado por 10 autores y por cada autor contiene 50 documentos de entrenamiento y 50 de validación, los autores son; Alan Crosby, Alexander Smith, Benjamín KangLim, David Lawder, Jane Macartney, Jim Gilchrist, Marcel Michelson, Mure Dickie, Robin Sidel, Todd Nissen.

3.2.- Análisis de datos y etiquetados

La representación de características utilizadas son las palabras y caracteres utilizando el modelo de n-gramas, este se define como una subsecuencia de elementos, donde n es el número de elementos que componen la secuencia [11].

3.3.- Construcción de características y ponderación

Se utilizó una representación vectorial, con la ponderación booleana donde 1 indica la presencia de la característica en el documento y un 0 indica su ausencia de esta.

3.4.- Selección y proyección de características

El método propuesto consiste:
1.- Caracterizar el documento mediante una representación de palabras o caracteres que representaran un conjunto
2.- Determinar el número de conjuntos a interceptar
3.- Extraer las características que se encuentren en la intercepción de los conjuntos. Para poder comparar el método propuesto se desarrollaron los métodos de análisis de componentes principales, ganancia de información, puntuación de información mutua, máxima relevancia mínima de redundancia, *Random forest.*

3.5.- Entrenamiento de un modelo de clasificación

Los modelos de aprendizaje supervisado que se utilizaron son:

- **Máquina de Soporte Vectorial**: En [20] mencionan que el enfoque de las máquinas de vectores de soporte (SVM) constituye un algoritmo dentro del ámbito de Machine Learning, empleado en tareas de clasificación y regresión. Su principio recae en la identificación del hiperplano de máxima separación entre dos clases de datos, con el propósito de lograr la mejor distinción posible. Este método ha demostrado su eficacia en diversas aplicaciones y escenarios, destacando por su capacidad para abordar problemas de clasificación y regresión mediante la identificación de un hiperplano óptimo.

- **Naive Bayes:** En [21] mencionan que el método Naive Bayes se configura como un algoritmo de clasificación ampliamente empleado en minería de datos. Su fundamentación radica en la teoría de la probabilidad, utilizada para calcular probabilidades condicionales y realizar predicciones en nuevos casos. Se caracteriza por ser tanto predictivo como descriptivo, presentando una simplicidad que no compromete su éxito, evidenciado por los resultados favorables obtenidos en diversas aplicaciones. La designación "naive" alude a la suposición ingenua, aunque no siempre acertada, de independencia entre las variables en el conjunto de datos.

- **Regresión Logística:** En [22] menciona que la Regresión Logística es un modelo de regresión empleado específicamente cuando la variable de interés es dicotómica. Este modelo facilita la estimación de la probabilidad de que la variable dependiente adquiera el valor del evento previamente definido. A diferencia de construir un modelo de regresión para estimar valores

reales, la Regresión Logística se basa en el cálculo de probabilidades. La variable dependiente resultante puede asumir cualquier valor, y se recurre a métodos de estimación propios de los modelos de regresión tradicionales para desarrollar el modelo de Regresión Logística.

3.6.- Evaluación de la solución

La medida de evaluación utilizada fue la exactitud, la cual se define como la proporción de predicciones correctas hechas por el modelo en comparación con el total de predicciones realizadas. Se expresa en términos de verdaderos positivos (VP), falsos positivos (FP), verdaderos negativos (VN) y falsos negativos (FN). Este valor se calcula dividiendo el número de predicciones correctas entre el total de predicciones [23].

$$Exactitud = \frac{VP + VN}{VP + VN + FN} \qquad (2)$$

4.- EXPERIMENTACIÓN Y RESULTADOS

Para la experimentación se consideraron la representación de palabras y caracteres utilizando diferentes tamaños de n-gramas, cabe mencionar que la selección de características por conjuntos, no se puede determinar un cierto número de características a diferencia de los métodos propuestos, por lo que a partir de la intersección de conjuntos se determinó el número de características a utilizar para los demás métodos.

4.1 Experimentación utilizando una representación de características de palabras

Para la experimentación de la investigación se tomaron en cuenta los siguientes métodos de selección de características: análisis de componentes principales **(M1)**, ganancia de información **(M2)**, puntuación de información mutua **(M3)**, máxima relevancia mínima de redundancia **(M4)**, Random Forest **(M5)**, y el Método propuesto **(MP)**, y como métodos de aprendizaje Máquina de Soporte Vectorial, Naive Bayes y Regresión logística.

Los resultados se pueden observar en la Tabla 1 y 2, donde el mejor resultado lo obtuvo el método de aprendizaje de Máquina de soporte vectorial y el método propuesto **(MP)** con 81.2% de exactitud, y 77.8 % de exactitud en la Tabla 2.

Tabla 1 Características 3178, 1-grama

Métodos de aprendizaje supervisado	**Métodos de Selección de características**					
	M1	**M2**	**M3**	**M4**	**M5**	**MP**
Máquina de soporte vectorial	63.2	64.6	76	76.6	74.4	**81.2**
Naive Bayes	62.8	65.6	75.2	74.4	71	77
Regresión logística	65.6	68.4	77.2	77.4	76.2	79.4

Tabla 2 Características 2659, 2-gramas

Métodos de aprendizaje supervisado	**Métodos de Selección de características**					
	M1	**M2**	**M3**	**M4**	**M5**	**MP**
Máquina de soporte vectorial	33.6	42	46.4	39.2	40.2	**77.8**
Naive Bayes	35.6	42	52.2	50.8	46.4	75.2
Regresión logística	34.4	46.8	45.4	42.8	40.4	77.2

Mientras que en la Tabla 3 se obtienen mejores resultados con Naive Bayes y el método propuesto MP con 63.2% de exactitud. Lo que nos indica que el método propuesto obtiene un mejor desempeño en referencia a los demás métodos.

Tabla 3 Características 715, 3-gramas

Métodos de aprendizaje supervisado	**Métodos de Selección de características**					
	M1	**M2**	**M3**	**M4**	**M5**	**MP**
Máquina de soporte vectorial	16	15.2	17.2	24	17.8	57.19
Naive Bayes	18.2	14.6	20.8	21.4	19.2	**63.2**
Regresión logística	15.6	13.6	17.8	23	20.8	61

4.2 Experimentación utilizando una representación de características de caracteres

Como se puede observar en la Tabla 4 el método que obtiene mejores resultados es el método **(M3)** con el método de aprendizaje de regresión logística alcanzando una exactitud de 76%, sin embargo, con el método propuesto **(MP)** apenas hay una diferencia del 1.4% siendo un resultado competitivo.

Tabla 4 Características 2094, 2-gramas

Métodos de aprendizaje supervisado	**Métodos de Selección de características**					
	M1	M2	M3	M4	M5	MP
Máquina de soporte vectorial	71.8	69	75.6	74.2	73.4	74
Naive Bayes	69.1	68.6	72	73	71.2	73.2
Regresión logística	73.6	71.6	**76**	75	73.8	**74.6**

En la Tabla 5 se puede observar que el mejor resultado fue obtenido con el método de aprendizaje de regresión logística y el método propuesto con 79.2% de exactitud.

Tabla 5 Características 8938, 3-gramas

Métodos de aprendizaje supervisado	**Métodos de Selección de características**					
	M1	**M2**	**M3**	**M4**	**M5**	**MP**
Máquina de Soporte Vectorial	77.6	71.8	78.8	78.6	77.6	79
Naive Bayes	72.8	70.1	74.6	73.8	72	74.4
Regresión logística	76.4	73.8	78	77.2	76.6	**79.2**

En la Tabla 6 se puede observar que el mejor resultado fue obtenido con el método de aprendizaje de Máquina de Soporte Vectorial y el método propuesto **(MP)** con 79.2% de exactitud.

Tabla 6 Características 1997, 4-gramas

Métodos de aprendizaje supervisado	**Métodos de Selección de características**					
	M1	**M2**	**M3**	**M4**	**M5**	**MP**
Máquina de Soporte Vectorial	75.2	71.8	78	77.2	79.2	**81**
Naive Bayes	69.3	69.6	73.2	73.8	74.2	74
Regresión logística	75	73.4	77.4	76.4	78.2	80.4

4.3 Resultados con respecto al estado del arte

Tabla 7. Comparación del método propuesto con el estado del arte.

Métodos del Estado del Arte	**Exactitud**
Método Propuesto	**81.2 %**
Extracción de características con Máquina de Soporte Vectorial [24]	79.68 %
Trigrama con Frecuencia del Término y Máquina de Soporte Vectorial [25] 19	80.8 %
LenFreqLexMorVocSuf [26]14	68 %
Doc2vec con palabras [27] 17	80.2 %
Medida R [28] 18	77.2 %

Como se puede observar en la tabla 7, los resultados obtenidos con el método propuesto son supriores a los metidos del estado del arte. Una de las ventajas de este método es el costo computacional a diferencia de los revisados en el estado del arte.

5. CONCLUSIONES Y TRABAJO FUTURO

El método propuesto de extracción de características a partir de la intercesión de conjuntos es una nueva propuesta de extracción de características que no requiere de un alto costo computacional, fácil de implementar y con mejores resultados que algunas propuestas del estado del arte. El número de conjuntos se determina a partir de número de clases a clasificar, en la presente investigación se utilizaron 10 conjuntos, posteriormente se extraen las características que se encuentran en la intercesión de estos, este método obtiene buenos resultados debido a que en la intercesión se encuentran las características que tienen en común los documentos de cada autor de esta observación se puede determinar por qué el método obtiene buenos resultados.

El desarrollo de este método permite la reducción de la dimensionalidad de las características, conservando el desempeño del clasificador de aprendizaje automático, obteniendo mejores resultados que en el estado del arte como se puede observar en la Tabla 7.

Como trabajo futuro se pretende implementar el método propuesto en otras tareas del procesamiento del lenguaje natural como perfil de autor, minería de opiniones, análisis de sentimientos, entre otras, para consolidar el método propuesto como un método de selección de características.

5. REFERENCIAS

[1] Castro, D., Adame, y., Peláez, M., & Muñoz, R. (2015). Authorship verification, average similarity analysis. International Conference Recent Advances in Natural Language Processing.

[2] Puig, X., Font, M., y Ginebra, J. (2016). A unified approach to authorship attribution and verification. The American Statistician, 70(3), 232-242.

[3] Frontini, F., Lynch, G. & Vogel, C. (2008). Revisiting the Donation of Constantine. Proceccedings of the AISB 2008 Simposium on Style in Text: Creative Generation and Identification of Authorship, (7), pp. 1-9.

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

[5] Xue, B., Zhang, M., y Browne, W. (2015). A comprehensive comparison on evolutionary feature selection approaches to classification. International Journal of Computational Intelligence and Applications, 14(02), 1550008.

[6] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 16-28.

[7] Virginia Ahedo, & Galán., J. M. (2019). Métodos de selección de características para la reducción de dimensionalidad. Revista de Ciencia de Datos., 5(2), 123-145.

[8] Solorio-Fernández, S., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2020). A review of unsupervised feature selection methods. Artificial Intelligence Review, 53(2), 907-948.

[9] Ramírez, J. M., Ruíz, M. C., & Somodevilla, M. J. (2014). Atribución de autoría combinando información léxico-sintáctica mediante representaciones holográficas reducidas. Res. Comput. Sci., 88, 103-113.

[10] Queralt, S. (2014). Acerca de la prueba lingüística en atribución de autoría hoy. Revista de Llengua i Dret, (62).

[11] Sidorov, G. (2013). N-gramas sintácticos no-continuos. Polibits, (48), pp. 69-78.

[12] Lipschutz, S. (1991). Schaum's Outline of Linear Algebra. New York: McGraw-Hill.

[13] Alonso del Corral, Aurora; (2004). La intersección edu-comunicativa. España: Comunicar.

[14] Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., ... & Nowe, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM transactions on computational biology and bioinformatics, 9(4), 1106-1119.

[15] Colina, C. L., & Roldán, P. L. (1991). El análisis de componentes principales: aplicación al análisis de datos secundarios. Papers: revista de sociologia, 31-63.

[16] Perea-Ortega, J., Díaz-Galiano, M., Montejo-Ráez, A., & García-Cumbreras, M. (2011). Análisis de la expansión de consulta para colecciones médicas utilizando información mutua, ganancia de información y la ontología MeSH. Procesamiento del Lenguaje Natural, 1-8.

[17] Vicente González, F. (2022). Métodos de Selección de Variables en Modelos de Regresión. UNIVERSIDA DE SANTIAGO DE COMPOSTELA, 35.

[18] Alaminos Fernández, A. (2023). Árboles de decisión en R con Random Forest. Alemania: Universidad de Alicante. Obets Ciencia Abierta. Alicante: Limencop.

[19] Mirończuk, M., Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. In: Expert Systems with Applications, 106, 36-54.

[20] Villasana, S., Caralli, A., Seijas, C., Villanueva, C., Sáenz, L., & Arteaga, F. (2006). Un novedoso método para estimar la resistencia rotórica del motor de inducción usando máquinas de vectores de soporte. Revista INGENIERÍA UC, 7-14.

[21]Pacheco Leal, S., Díaz Ortiz, L., & García Flores, R. (2005). El clasificador Naive Bayes en la extracción de conocimiento de bases de datos. Ingenierías, 24-33.

[22]Moral Peláez, I. (2006). Modelos de regresión: lineal simple y regresión logística. Atención Primaria, 195-214.

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

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

[25] Plakias. S and Stamatatos, E. (2008). Tensor space models for authorship attribution. In Proc. of the 5th Hellenic Conference on Artificial Intelligence

[26] Gómez-Adorno, H., Sidorov, G., Pinto, D., Vilariño, D., & Gelbukh, A. (2016). Automatic authorship detection using textual patterns extracted from integrated syntactic graphs. Sensors, 16(9), 1374.

[27] Posadas-Durán, J. P., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Pinto, D., & Chanona-Hernández, L. (2017). Application of the distributed document representation in the authorship attribution task for small corpora. Soft Computing, 21(3), 627-639

[28] Potthast, M., Braun, S., Buz, T., Duffhauss. (2016). Who Wrote the Web? Revisiting Influential Author Identification Research Applicable to Information Retrieval. European Conference on Information Retrieval. Lecture Notes in Computer Science, 9626, 393-407, Springer.