DOI: http://dx.doi.org/10.23857/dc.v8i3

Ciencias Técnicas y Aplicadas
Artículo de Investigación

Análisis de métodos de clasificación y frecuencia de palabras para la autoría de textos en español

Analysis of classification methods and frequency of words for the authorship of texts in Spanish

Análise de métodos de classificação e frequência de palavras para autoria de textos em espanhol

César Espín-Riofrio [I]
cesar.espinr@ug.edu.ec
https://orcid.org/0000-0001-8864-756X

Alexis Proaño-Indacochea [II]
alexis.proanoin@ug.edu.ec
https://orcid.org/0000-0001-9787-0149

Christian Castro-Martínez [III]
christian.castroma@ug.edu.ec
https://orcid.org/0000-0002-5185-1820

Tania Peralta-Guaraca [IV]
tania.peraltag@ug.edu.ec
https://orcid.org/0000-0002-4879-6824

Correspondencia: cesar.espinr@ug.edu.ec

*Recibido: 29 de septiembre del 2022 *Aceptado: 28 de octubre del 2022 * Publicado: 8 de noviembre del 2022

I. Magíster en Sistemas de Información Gerencial, Universidad de Guayaquil, Guayaquil, Ecuador.
II. Universidad de Guayaquil, Guayaquil, Ecuador.
III. Universidad de Guayaquil, Guayaquil, Ecuador.
IV. Magíster en Ingeniería de Software y Sistemas Informáticos Universidad de Guayaquil, Guayaquil, Ecuador.

Resumen

El uso de machine learning de la mano con la estilometría es de mucha importancia para la determinación de autoría de textos en español. Mediante una investigación exhaustiva de artículos relevantes y el establecimiento del estado del arte de lo que es la estilometría y los métodos de clasificación para machine learning hasta la actualidad, se pretende establecer las técnicas y características que más nos beneficien para la atribución de autoría y por consiguiente entrenar y evaluar los métodos de clasificación utilizados. Se experimenta con librerías que contienen una biblioteca de estilometría, de la cual, se obtienen los métodos para extraer las características de tipo fraseológico. Se utiliza el dataset de las campañas PAN 2015 el cual otorga un corpus en español para varios autores. También se usa el corpus validado por la Real Academia Española de la Lengua llamado CREA con las palabras de uso más frecuente en el idioma español, con esto se alimenta los clasificadores para machine learning y mediante el uso de validación cruzada y las métricas de evaluación se obtiene qué método presenta mejores resultados en la fase de entrenamiento.

Palabras clave: Atribución de autoría; Estilometría; Procesamiento de Lenguaje Natural; Machine Learning.

Abstract

The use of machine learning hand in hand with stylometry is very important for determining the authorship of texts in Spanish. Through an exhaustive investigation of relevant articles and the establishment of the state of the art of what is stylometry and classification methods for machine learning to date, it is intended to establish the techniques and characteristics that most benefit us for the attribution of authorship and for consequently train and evaluate the classification methods used. Experiments are carried out with libraries that contain a stylometry library, from which the methods to extract the phraseological type characteristics are obtained. The dataset of the PAN 2015 campaigns is used, which provides a corpus in Spanish for several authors. The corpus validated by the Royal Spanish Academy of Language called CREA is also used with the words most frequently used in the Spanish language, with this the classifiers for machine learning are fed and through the use of cross validation and the evaluation metrics are used. obtains which method presents better results in the training phase.

Keywords: Attribution of authorship; Stylometrics; Natural Language Processing; machine learning.

Resumo

O uso do aprendizado de máquina de mãos dadas com a estilometria é muito importante para determinar a autoria de textos em espanhol. Através de uma exaustiva investigação de artigos relevantes e do estabelecimento do estado da arte do que é estilometria e métodos de classificação para aprendizagem de máquina até à data, pretende-se estabelecer as técnicas e características que mais nos beneficiam para a atribuição de autoria e consequentemente treinar e avaliar os métodos de classificação utilizados. Os experimentos são realizados com bibliotecas que contêm uma biblioteca de estilometria, da qual são obtidos os métodos para extrair as características do tipo fraseológico. Utiliza-se o conjunto de dados das campanhas PAN 2015, que fornece um corpus em espanhol para diversos autores. O corpus validado pela Real Academia Espanhola de Línguas chamado CREA também é usado com as palavras mais usadas na língua espanhola, com isso são alimentados os classificadores para aprendizado de máquina e através do uso de validação cruzada e as métricas de avaliação são usadas. qual método apresenta melhores resultados na fase de treinamento.

Palavras-chave: Atribuição de autoria; Estilometria; Processamento de linguagem natural; aprendizado de máquina.

Introducción

Machine Learning y estilometría han ido avanzando en sus métodos para encontrar los resultados más eficaces cuando se trata del análisis de datos, en base a esto, se precisa combinar estas dos disciplinas tan significativas en la actualidad y adaptar estas herramientas para obtener información a partir de textos en español y realizar las predicciones más acertadas posibles. La importancia del estudio reside en la necesidad de usar los datos que se pueden obtener a partir de textos de diferentes autores y así poder usar esta información para varios casos investigativos referentes a la atribución de autoría como en casos de desconocimiento de un autor o plagios, debido a que no existen muchos estudios que se han desarrollado o trabajado con textos en países hispanohablantes, así todo esto beneficiará a futuras investigaciones en el área, como las personas que estudian la

lengua y la literatura, proporcionándoles una herramienta segura para sus estudios. Desde los años 80's hay varios autores que hicieron estudios y nos relatan sobre la estilometría como (Lutoslawski, 1898), quien acuñó el término y estableció las bases de la estilometría en su artículo aplicado a la cronología de las obras de Platón. Años después (FredeRick & David., 1964) llevaron a cabo un estudio estilométrico en el que ya se aplicaban modernas técnicas de análisis estadístico. Otro avance lo tuvieron (Tweedie et al., 1996), los cuales, nos relatan sobre las redes neuronales y su aplicación para la estilometría, diciendo que esencialmente la estilometría se trata de patrones. Por otro lado (Burrows, 2002), propuso una nueva forma de usar frecuencias relativas de palabras comunes para comparar textos y eficaz para la atribución. (Juola et al., 2006) establece el estado del arte y un marco para el desarrollo uniforme, que coordina y combina varios enfoques diferentes para la atribución de autoría. (Akcapinar Sezer et al., 2020), proponen una combinación de características de estilometría con redes neuronales profundas para mejorar la consistencia de las soluciones. (Adebayo & Yampolskiy, 2022), elaboraron un artículo muy reciente en el que con la estilometría y el uso de machine learning tratan de estimar el coeficiente intelectual de las personas mediante el uso de un conjunto de datos correctos.

Los clasificadores de machine learning también han ido avanzando durante el paso del tiempo comenzando por el clasificador Navies Bayes, el cual, está basado en el cálculo de probabilidades y algunas hipótesis simplificadoras adicionales (Bayes, 1763). (Rosenblatt, 1958), desarrollaron un modelo llamado "Perceptron", el cual, efectúa cálculos para detectar características o tendencias en los datos de entrada. Después (Cover & Hart, 1967), elaboraron algunas de las propiedades formales de la regla del k-vecino más cercano. La Recurrent Neural Network fue basado en el trabajo de (Rumelhart et al., 1986), los cuales, proponen un nuevo procedimiento de aprendizaje para redes de unidades similares a neuronas con funciones de transferencia no lineales conocida como Multilayer Perceptron a través de la llamada "Regla Delta Generalizada". Las máquinas de vectores de soporte o SVM son un conjunto de algoritmos de machine learning desarrollados por (Cortes & Vapnik, 1995). Estos métodos mapean la entrada de vectores en algún espacio de características de alta dimensión a través de algún mapeo no lineal elegido y se usan en varios problemas de clasificación y regresión. (Stoean et al., 2008), nos relata sobre un nuevo enfoque que adopta la estrategia de SVM llamada optimización evolutiva. (Chen & Ishwaran, 2012), en su artículo nos definen el clasificador random forest como una herramienta de conjunto basada en árboles que es altamente adaptable a los datos. (Charbuty & Abdulazeez, 2021), proporciona un

enfoque detallado de los árboles de decisión y nos dice que el clasificador Decision Tree es una técnica basada en árboles, en la cual, todos los caminos a partir de la raíz se describen mediante una separación de secuencia de datos hasta que se logra un resultado booleano en el nodo hoja.

Método

Se realizó una revisión bibliográfica de los aportes relevantes de diferentes autores, de donde, se obtuvo la información necesaria para el proceso de estudio y así poder realizar los objetivos propuestos. El proyecto tiene un enfoque experimental, puesto que se meden variables y se establecen las características estilométricas que dan los mejores resultados para textos en el idioma de estudio. Se adaptan y realizan cambios al modelo en algunos algoritmos que hacen uso de la estilometría en la extracción de las características orientados al idioma inglés, para que analicen los textos en el idioma a estudiar, en este caso, el idioma español. Se establecen los datasets correspondientes para la obtención de los datos, luego se realiza un procesamiento de los datos a usar para una mejor obtención en los resultados. A continuación, se procede con extracción de las características estilométricas seleccionadas para el estudio y se entrenan los métodos de clasificación con estas características extraídas, para verificar que nuestros clasificadores nos arrojen las predicciones más acertadas se evaluaron estos métodos con las métricas de evaluación y el uso de validación cruzada. Posterior a esto, se obtuvieron y analizaron los resultados del estudio, en la Figura 1 se detallan los pasos a seguir.

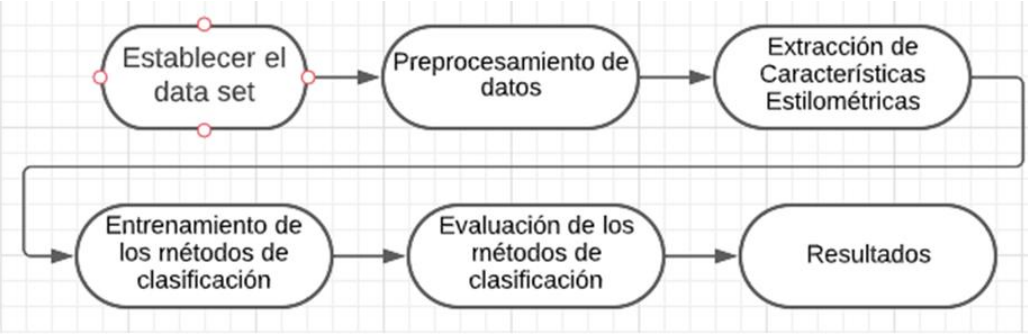

Figura 1: Procedimiento a seguir.

Dataset utilizado

Se solicitó el dataset a las campañas PAN 2015, dentro encontramos el corpus donde su estructura se encuentra compuesta por textos de 100 autores de varios idiomas, entre ellas el idioma español,

el cual, cuenta con 100 carpetas y dentro tendrán 4 archivos por autor y un documento cuestionado como se observa en las Figuras 2 y 3.

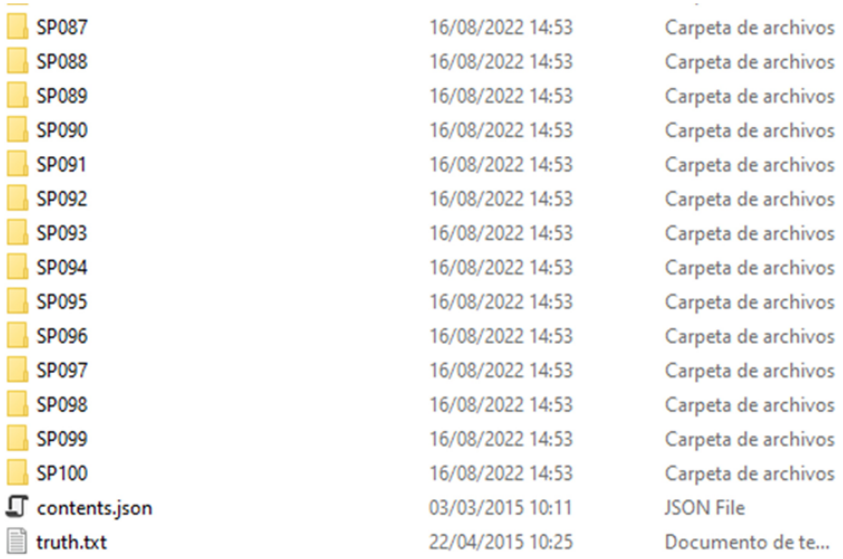

Figura 2: Directorios de los textos en español.

Figura 3: Textos por autor.

Preprocesamiento de datos

En el preprocesamiento de datos, se usó la librería re de Python para poder aplicar expresiones regulares, separar signos y convertir las palabras que tengan cualquier carácter en mayúsculas a minúsculas como se muestra en las Figuras 4 y 5.

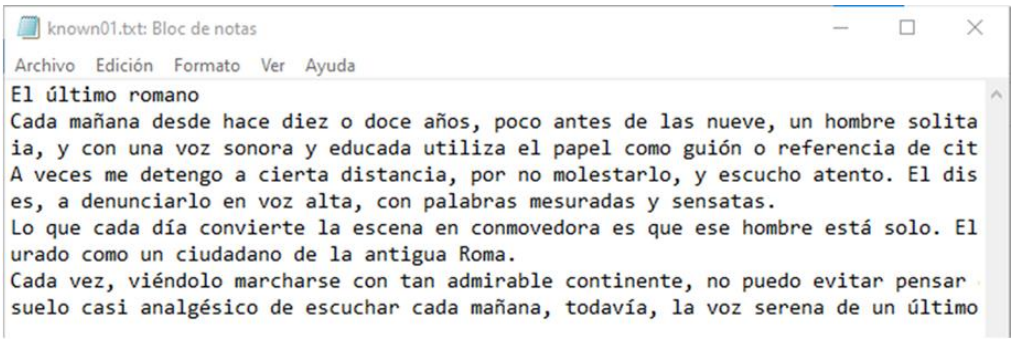

Figura 4: Archivo de texto antes del preprocesamiento.

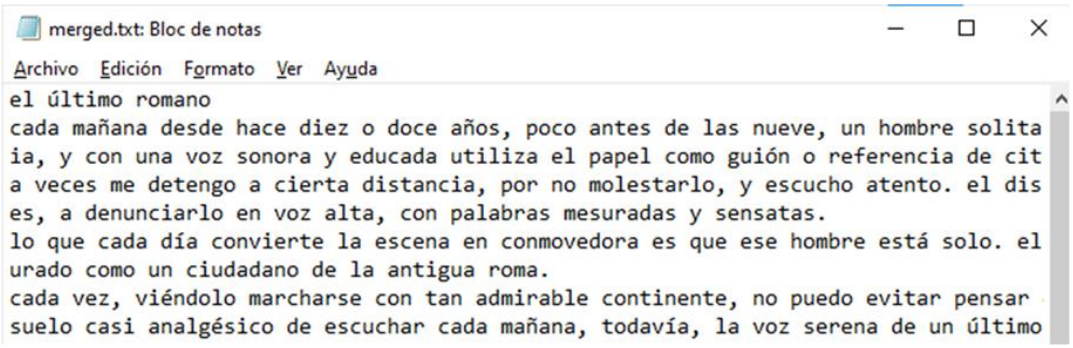

Figura 5: Archivo de texto después del preprocesamiento.

Extracción de características estilométricas

Para realizar la extracción de las características estilométricas de tipo fraseológico se usó una biblioteca de estilometría para Python desarrollada por Jeff Potter1 (GitHub - Jpotts18/Stylometry: A Stylometry Library for Python, n.d.), la cual, fue adaptada para reconocer los signos de puntuación y palabras del idioma español, las características estilométricas de tipo fraseológico usadas fueron: Lexical Diversity (Diversidad Léxica), Mean Word Length (Longitud media de las palabras), Mean Sentence Length (Longitud media de la frase), Sentence Length (Longitud de la frase), Mean Paragraph Length (Longitud media del párrafo), Document Length (Longitud del documento). Así mismo, para poder extraer las palabras de uso frecuente se usó un corpus llamado "El Corpus de Referencia del Español Actual" (CREA), de este listado se extraerán las 1000 primeras palabras, cabe indicar que este corpus es otorgado y validado por la Real Academia Española de la Lengua2. Una vez formado el vector que contiene el cálculo de las características estilométricas para cada uno de los textos, se procede a calcular las distancias que existen entre vectores. Las distancias calculadas fueron: cosine, euclidean, manhattan, chi2, cityblock, l1, l2,

como se puede observar en las Figuras 6 y 7, para calcular estas distancias se utilizan métodos importados de la librería scikit-learn3 de Python.

Figura 6: Cálculo de las características estilométricas de tipo fraseológico.

Author	LexicalDiversity	MeanWordLen	MeanSentenceLen	StdevSentenceLen	MeanParagraphLen	DocumentLen
SP001	31.35820096	7.105470953	23.31884058	18.30088783	1206.75	28660
SP002	33.20703654	6.723214286	20.55194805	12.63115207	791.25	18614
SP003	32.3913623	6.703673469	19.6196319	12.91938601	799.5	18650
SP004	32.74619696	6.788398693	18.29378531	11.75520636	809.5	18687
SP005	33.04289544	6.823174816	26.80555556	19.31870056	965	22463
SP006	32.4970964	7.520996441	29.6875	19.35876336	950	23941
SP007	32.42085218	7.541391941	29.72440945	18.11758495	943.75	23635
SP008	31.51711566	7.422151899	18.3487395	11.67883502	1091.75	26506
SP009	32.15381838	7.12	24.88020833	21.54504061	1194.25	28532
SP010	31.65662025	7.448742747	17.59917355	10.88674223	1064.75	25690
SP011	30.18039216	7.110309278	24.66517857	19.08321631	1381.25	32595
SP012	32.37472767	7.505312085	17.92982456	10.95182377	1022	25065
SP013	31.92699692	7.477810651	28.6870229	13.81728311	939.5	23384
SP014	32.3913623	6.703673469	19.6196319	12.91938601	799.5	18650
SP015	31.92895498	7.522264631	16.95256917	10.3735432	1072.25	26136

Figura 7: Cálculo de las distancias entre palabras.

Author	cosine	euclidean	manhattan	chi2	cityblock	l1	l2	braycurtis	canberra	chebyshev	correlation	minkowski	sqeuclidean	final_array
SP001	0.00820027	0.12979462	0.2231136	0.86373231	0.2231136	0.2231136	0.12979462	0.09516524	49.3209008	0.12766961	0.00804203	0.12979462	0.01684664	29794621492
SP002	0.00719512	0.12145931	0.16710938	0.90912516	0.16710938	0.16710938	0.12145931	0.07250327	36.7940833	0.12110259	0.00716679	0.12145931	0.01475236	21459311421
SP003	0.0080627	0.12870913	0.2102519	0.87924449	0.2102519	0.2102519	0.12870913	0.08944332	41.3498911	0.12719506	0.00794469	0.12870913	0.01656604	87091294696
SP004	0.00800632	0.12826261	0.20931137	0.88060215	0.20931137	0.20931137	0.12826261	0.08901413	40.6409321	0.12674481	0.00788945	0.12826261	0.0164513	82626109292
SP005	0.00812242	0.12919204	0.22678916	0.86347474	0.22678916	0.22678916	0.12919204	0.09617547	48.2401499	0.12677327	0.00795905	0.12919204	0.01669058	29192043000
SP006	0.00819378	0.12967756	0.21154019	0.87195346	0.21154019	0.21154019	0.12967756	0.09074098	42.3708209	0.1284269	0.00806846	0.12967756	0.01681627	77559042835
SP007	0.00924912	0.13793611	0.23240054	0.85493026	0.23240054	0.23240054	0.13793611	0.0988173	48.2989493	0.1358261	0.00909322	0.13793611	0.01902637	37936110162
SP008	0.00791149	0.12742467	0.1966808	0.88374149	0.1966808	0.1966808	0.12742467	0.08493474	43.8200647	0.12645223	0.00781979	0.12742467	0.01623705	74246711004
SP009	0.00709801	0.12060385	0.19319936	0.88494751	0.19319936	0.19319936	0.12060385	0.08350154	48.1585669	0.11959816	0.0069977	0.12060385	0.01454529	60385497063
SP010	0.00869047	0.13367033	0.22728058	0.86024276	0.22728058	0.22728058	0.13367033	0.09676081	43.9604147	0.13203828	0.00853201	0.13367033	0.01786776	67033162680
SP011	0.00818311	0.12965993	0.22543149	0.85965176	0.22543149	0.22543149	0.12965993	0.09625631	50.306028	0.12740841	0.00801606	0.12965993	0.0168117	96599319504
SP012	0.00775347	0.12610574	0.25768037	0.83476251	0.25768037	0.25768037	0.12610574	0.10842963	54.3126604	0.12250371	0.00747719	0.12610574	0.01590266	61057356580
SP013	0.00684254	0.11834106	0.18577985	0.89220697	0.18577985	0.18577985	0.11834106	0.08053686	40.4258305	0.11731917	0.00675478	0.11834106	0.01400461	18341059240
SP014	0.00791965	0.12753983	0.17515028	0.90277359	0.17515028	0.17515028	0.12753983	0.07570345	36.6839161	0.12715674	0.0078923	0.12753983	0.01626641	75398255881
SP015	0.00648619	0.11519081	0.19135777	0.88630816	0.19135777	0.19135777	0.11519081	0.08285073	47.1614775	0.11406907	0.00637898	0.11519081	0.01326892	51908109396

Entrenamiento de los métodos de clasificación

Para realizar este paso se debe definir el porcentaje de datos que se usarán para el entrenamiento y el porcentaje correspondiente para pruebas, para este estudio se definió el 80% y el 20% respectivamente como se puede ver en la Figura 8, luego se seleccionan las características estilométricas extraídas y se procede con el entrenamiento de los métodos: Decision Trees (DT), Random Forest (RF), Multilayer Perceptron (MLP) y Gradient Boost (GBT). Como se puede observar en la Figura 9, en la definición de estos métodos se procede a establecer el parámetro random_state con un valor de 45 para los 3 primeros métodos y un valor de 0 para el método Gradient Boost, el cual, asegura que no importa las veces que se ejecute el problema siempre se reproduce de la misma forma para no obtener resultados diferentes, así mismo en el método

Random Forest se establece el parámetro n_estimators que se traduce al número de árboles que se va a utilizar en el bosque en este caso se definen 20 árboles. En el método MLP se establece el parámetro max_iter con un valor de 1000, con el cual, podemos indicar el número de iteraciones a realizarse sobre los datos, finalmente en el método Gradient Boost se establecen 2 parámetros más, los cuales son learning_rate y max_depth, donde se define la tasa de aprendizaje para reducir la contribución de cada árbol y se obtiene el mejor rendimiento respectivamente, estos métodos fueron potenciados con la técnica validación cruzada realizándose esta última 10 veces para un mejor entrenamiento. Cabe mencionar que estos métodos y técnicas ya están definidos en la librería scikit-learn4 de Python.

```
#función cargar archivo y asigno X y y, y hago 80-20

def mifun(n_File):

  ruta='/content/output_' + str(n_File) + '.csv'
  df = pd.read_csv(ruta, encoding='latin1')


  # Seleccionamos columnas de características
  feature_cols = ['cosine', 'euclidean', 'manhattan', 'chi2', 'cityblock', 'l1',

  X = df[feature_cols]
  #X

  # seleccionamos columna objetivo
  y = df['truth_binary']


  # Usaremos validación cruzada para evaluar
  from sklearn.model_selection import cross_validate
  from sklearn.model_selection import train_test_split

  # Dejamos 20% para validación final
  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

  return X_train, y_train
```

Figura 8: Selección de características estilométricas y porcentajes de prueba y entrenamiento

```
def eval_classifiers(X_train, y_train):
  clfs = [
          ('Decision tree', DecisionTreeClassifier(random_state=45)),
          ('RandomForest', RandomForestClassifier(n_estimators=20, random_state=45)),
          ('MLP', MLPClassifier(max_iter=1000, random_state=45)),
          ('GradientBoost',GradientBoostingClassifier(n_estimators=100, learning_rate=1.0, max_depth=1, random_state=0)),
          ]


  # Vamos devolver los resultados como una tabla
  # Cada fila un algoritmo, cada columna un resultado
  metrics = ['accuracy', 'precision', 'recall', 'f1']
  results = pd.DataFrame(columns=metrics)
  for alg, clf in clfs:
    scores = cross_validate(clf, X_train, y_train, cv=10, scoring=metrics) # por defecto, es estratificado
    results.loc[alg,:] = [scores['test_'+m].mean() for m in metrics]
  return results
```

Figura 9: Definición de métodos clasificadores y validación cruzada.

Resultados

Para evaluar los métodos clasificadores se usaron las siguientes métricas: Accuracy, Precision, Recall y F1. Como podemos observar a través de la ejecución de los apartados anteriores, se obtuvo los siguientes resultados de las métricas de evaluación para cada método como se muestra en la Tabla 1 y se gráfica en la Figura 10.

Figura 10: Métricas de Evaluación para cada uno de los métodos clasificadores.

	Accuracy	Precision	Recall	F1
Decision Tree	0.8375	0.8617	0.8250	0.8166
Random Forest	0.7750	0.7500	0.8250	0.7776
MLP	**0.8750**	**0.8650**	**0.9250**	**0.8840**
Gradient Boost	0.8500	0.8200	**0.9250**	0.8622

Nos centraremos en los resultados de la métrica F1, puesto que es una combinación de las métricas Precision y Recall. Siendo así, se puede evidenciar en la tabla 1 y la figura 9 que el clasificador MLP tiene un F1 de 0.8840, proclamándose como el de mejor resultado en la etapa de entrenamiento.

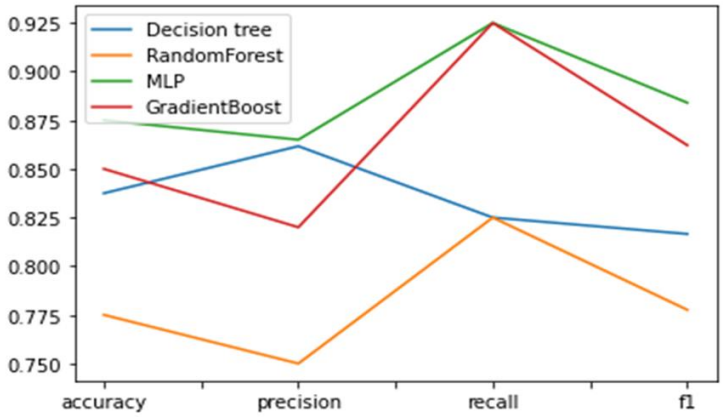

Figura 11: Métodos Clasificadores vs las métricas de evaluación de la tabla 3.

Para determinar con qué cantidades de palabras se obtienen mejores resultados podemos observar en la Tabla 2, los resultados de los experimentos que se realizaron con los métodos clasificadores en un intervalo de 50 en 50 palabras de las 1000 tomadas del CREA para cada uno de los textos en español. Gracias a estos datos podemos concluir que, si se utilizan más de 400 palabras con el clasificador MLP, los resultados empiezan a mostrar un decaimiento, evidenciando así que no es relevante usar una cantidad superior de palabras.

Figura 12: Métrica F1 para los métodos clasificadores respecto a la cantidad de palabras CREA

# palabrasCREA	Decision tree	Random Forest	MLP	Gradient Boost
50	0.7669	0.7938	0.8449	0.8344
100	0.6900	0.7804	0.8277	0.723
150	0.8204	0.8884	0.8974	0.8439
200	0.7865	0.7809	0.8889	0.7914
250	0.8250	0.8159	0.8935	0.8113
300	0.7785	0.8477	0.919	0.8112
350	0.7590	0.9083	0.9176	0.857
400	0.8056	0.9044	**0.9435**	0.8583
450	0.7774	0.8909	0.8824	0.8006
500	0.8490	0.8803	0.8935	0.8253
550	0.8713	0.874	0.8635	**0.877**
600	0.8151	**0.913**	0.9019	0.8464

650	0.8713	0.9102	0.8856	0.8674
700	**0.8909**	0.8763	0.8425	0.8512
750	0.8073	0.8348	0.899	0.8463
800	0.7747	0.8096	0.8824	0.8138
850	0.7994	0.8502	0.8588	0.8737
900	0.8550	0.9127	0.9102	0.8413
950	0.8360	0.8598	0.8209	0.8243
1000	0.8166	0.7776	0.884	0.8622

Discusión

Podemos observar que para este tipo de trabajos se debe establecer la cantidad de parámetros para cada uno de los métodos, la cantidad de características estilométricas de tipo fraseológico, e incluso la cantidad de palabras a utilizar antes de realizar el análisis de los textos, debido a que cualquier variación de estas cantidades pueden influir en los resultados finales. Así mismo, dichos resultados pueden varían o indicar que otro método puede ser el que tenga el mejor resultado si se llegase a utilizar métodos distintos a los usados en este estudio, como, por ejemplo, en el trabajo de (Maurya et al., 2016), donde se seleccionan 117 técnicas, se extrajeron 446 características estilométricas y realizaron una experimentación de textos de diferentes tamaños y diferente número de características con varias técnicas de aprendizaje automático concluyendo en dicho trabajo que el Support Vector Machine (SVM) es el método más adecuado para determinar la atribución de autoría de un autor.

Otro tema que se puede discutir es que la cantidad de los trabajos relacionados a este tópico se encuentran en su mayoría orientados para el idioma inglés, pero en este artículo se examinaron textos en idioma español para poder expandir este tipo de estudios en el área para la comunidad de hispanohablantes.

Conclusiones

Gracias al análisis realizado de contribuciones científicas de impacto sobre estilometría y machine learning realizadas para los diferentes idiomas, se identificó que los métodos clasificadores más utilizados en investigaciones para la atribución de textos de autores desconocidos. Se entrenaron Decision Trees (DT), Random Forest (RF), Multilayer Perceptron (MLP) y Gradient Boost (GBT)

potenciándolos con validación cruzada para poder obtener el método con los mejores resultados y con la ayuda de la métrica de evaluación F1 se evidenció que el método clasificador MLP con 0.9435 es el mejor para determinar la atribución de autoría de un texto en español en la presente investigación.

También que las palabras de uso frecuente como característica estilométrica es importante hasta cierto número de palabras que podrían considerarse influyentes, sobre ese umbral no hay mayor variación y más bien disminuye el rendimiento de los clasificadores.

Con la constante evolución de importantes tecnologías como Lenguaje de Procesamiento Natural, Inteligencia Artificial y Aprendizaje Automático, es totalmente factible entrenar métodos clasificadores para el idioma español y determinar la atribución de autoría mediante el análisis estilométrico y el uso de validación cruzada.

Referencias

1. Adebayo, G. O., & Yampolskiy, R. v. (2022). Estimating Intelligence Quotient Using Stylometry and Machine Learning Techniques: A Review. Big Data Mining and Analytics, 5(3), 163– 191. https://doi.org/10.26599/bdma.2022.9020002
2. Akcapinar Sezer, E., Sever, H., & Canbay, P. (2020). Deep Combination of Stylometry Features in Forensic Authorship Analysis. International Journal of Information Security Science, 9(3), 154–163. https://www.researchgate.net/publication/344408746
3. Bayes, T. (1763). Thomas bayes, an essay towards solving a problem in the doctrine of chances (1764). 199–207. https://doi.org/10.1016/B978-044450871-3/50096-6
4. Burrows, J. (2002). ‘Delta’: A Measure of Stylistic Difference and a Guide to Likely Authorship.
5. Literary and Linguistic Computing, 17(3), 267–287. https://doi.org/10.1093/llc/17.3.267
6. Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2(01), 20–28.
7. https://doi.org/10.38094/jastt20165
8. Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. Genomics, 99(6), 323–329. https://doi.org/10.1016/j.ygeno.2012.04.003
9. Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. 20, 273–297.
10. Cover, T. M., & Hart, P. E. (1967). Nearest Neighbor Pattern Classfication. I, 1–28.

11. FredeRick, M., & David., W. (1964). Inference in an Authorship Problem. Journal of the American Statistical Association, 274(6), 509. https://doi.org/10.1001/jama.1995.03530060085046
12. GitHub - jpotts18/stylometry: A Stylometry Library for Python. (n.d.). Retrieved August 23, 2022, from https://github.com/jpotts18/stylometry
13. Juola, P., Sofko, J., & Brennan, P. (2006). A prototype for authorship attribution studies. Literary and Linguistic Computing, 21(2), 169–178. https://doi.org/10.1093/llc/fql019
14. Lutoslawski, W. (1898). Principes de stylométrie appliqués à la chronologie des œuvres de Platon.
15. Revue Des Études Grecques, 11(41), 61–81. https://doi.org/10.3406/reg.1898.5847
16. Maurya, R. K., Saxena, M. R., & Akhil, N. (2016). Intelligent Systems Technologies and Applications. Advances in Intelligent Systems and Computing, 384, 247–257. https://doi.org/10.1007/978-3-319-23036-8
17. Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386–408. https://doi.org/10.1037/h0042519
18. Rumelhart, D. E., Hintont, G. E., & Williams, R. J. (1986). Learning Representations by Back- Propagating Errors. Cognitive Modeling, 2, 3–6. https://doi.org/10.7551/mitpress/1888.003.0013
19. Stoean, R., Preuss, M., Stoean, C., El-Darzi, E., & Dumitrescu, D. (2008). Support vector machine learning with an evolutionary engine. Journal of the Operational Research Society, 60(8), 1116–1122. https://doi.org/10.1057/jors.2008.124
20. Tweedie, F. J., Singh, S., & Holmes, D. I. (1996). Neural network applications in stylometry: The federalist papers. Language Resources and Evaluation, 30(1), 1–10. https://doi.org/10.1007/BF00054024