ARTÍCULO DE INVESTIGACIÓN / RESEARCH ARTICLE
https://dx.doi.org/10.14482/inde.44.01.215.568

Authorship Classification in Academic and Scientific Documents: A Machine Learning-Based Approach

Clasificación de autoría en documentos académicos y científicos: un enfoque basado en aprendizaje automático

Pablo Pico-Valencia*
Sahory Maila-Herrera**

* Universidad de Granada (España). Software Engineering Department, Research Centre for Information and Communication Technologies (CITIC-UGR). Ph.D. in Information and Communication Technologies (ICT).
Orcid-ID: https://orcid.org/0000-0003-3518-3313. pablo.pico@ugr.es

** Pontificia Universidad Católica del Ecuador Sede Esmeraldas. Systems and Computing Engineering. Eng. in System and Computing.
Orcid-ID: https://orcid.org/0009-0007-1702-9749. sahory.maila@pucese.edu.ec
Corresponding: Pablo Pico-Valencia, "Periodista Daniel Saucedo Aranda" Street, 18071, Granada (España).

Abstract

This paper presents a machine learning-based system that incorporates text mining to analyze and classify writing styles in scientific reports authored by faculty members at the Pontificia Universidad Católica del Ecuador, Esmeraldas. The system aims to enhance academic integrity by identifying potential cases of false authorship. A dataset of research papers written in Spanish by faculty professors was processed using TF-IDF (Term Frequency-Inverse Document Frequency) and Word Embeddings for feature extraction. To assess classification performance, seven supervised learning models were tested: Linear Support Vector Classifier (SVC), SVC with RBF kernel, Random Forest, Decision Tree, Logistic Regression, k-Nearest Neighbors (k-NN), and Naïve Bayes. The Logistic Regression model yielded the highest accuracy (89.62%), closely followed by Linear SVC (87.36%) and RBF SVC (86.59%), outperforming tree-based and probabilistic methods with statistical significance ($p<0.05$). The Wilcoxon test showed no significant performance differences among the best classifiers, confirming their reliability in authorship attribution. The findings highlight the promise of incorporating writing style analysis into institutional systems to enhance conventional methods for detecting plagiarism.

Keywords: data mining, machine learning, natural language processing, prediction, writing style.

Resumen

Este artículo presenta un sistema basado en aprendizaje automático que implementa minería de texto para analizar y clasificar estilos de escritura en informes científicos elaborados por docentes de la Pontificia Universidad Católica del Ecuador, sede Esmeraldas. El objetivo del sistema es fortalecer la integridad académica mediante la identificación de posibles casos de autoría falsa. Se procesó un conjunto de datos compuesto por artículos de investigación redactados en español por profesores universitarios, aplicando TF-IDF (Frecuencia de Término - Frecuencia Inversa de Documento) y Word Embeddings para la extracción de características. Para evaluar el rendimiento en la clasificación, se probaron siete modelos de aprendizaje supervisado: Clasificador Lineal de Vectores de Soporte (SVC), SVC con kernel RBF, Random Forest, Árbol de Decisión, Regresión Logística, k-Vecinos más Cercanos (k-NN) y Naïve Bayes. El modelo de Regresión Logística obtuvo la mayor precisión (89.62 %), seguido de cerca por el SVC Lineal (87.36 %) y el SVC RBF (86.59 %), superando con significancia estadística a los métodos basados en árboles y probabilísticos ($p<0.05$). La prueba de Wilcoxon no mostró diferencias significativas en el rendimiento entre los mejores clasificadores, lo que confirma su fiabilidad en la atribución de autoría. Los hallazgos subrayan el potencial de incorporar el análisis del estilo de escritura en los sistemas institucionales para mejorar los métodos convencionales de detección de plagio.

Palabras clave: aprendizaje automático, estilo de redacción, minería de datos, predicción, procesamiento del lenguaje natural.

Fecha de recepción: 11 de octubre de 2024
Fecha de aceptación: 3 de mayo de 2025

INTRODUCTION

Text mining enables machines to efficiently search and extract valuable information from text documents [1]. This is achieved by identifying characteristic patterns in the natural language used in these documents. Machines can now acquire explicit and structured information through text mining [2]. However, interpreting deeper meaning—as human do—remains a significant challenge [1].

Additionally, text mining applications are diverse and span multiple fields. In security, for example, text mining can anticipate and counteract terrorist activities. It does this by identifying connections among individuals and entities, and by analyzing patterns in social and economic behavior [3], [4]. In politics, it has been used to analyze public opinion on social networks such as X (formerly Twitter) [5]. In business and marketing, text mining helps evaluate customer feedback and comments. It goes beyond simple sentiment classification (positive, neutral, or negative) to identify customer needs, analyze product reviews, and track emerging trends [6]. For instance, it has been applied to improve and/or realign services in the tourism sector based on customer reviews [7]. It is also used to identify key attributes in banking services and perform customer sentiment analysis from online user reviews [8], and to enhance healthcare service design and delivery in hospitals by leveraging customer satisfaction metrics, which vary significantly depending on the service context [9].

Moreover, text mining has important applications in education and healthcare. In education, it is increasingly combined with machine learning to automate the extraction and classification of bibliographic materials in online learning environments. This enables personalized learning experiences and the early identification of students who may require additional support [10], [11]. Additionally, text mining is employed to detect plagiarism in student essays and assess writing styles [12]. In healthcare, it is utilized in biomedical and clinical settings to analyze clinical data, identify potential drug interactions, and track disease outbreaks [13]. Overall, text mining serves as a valuable tool for extracting and interpreting text from various sources, including books, articles, reports, theses, websites, and social networks.

The global pandemic, technological advancements, and the increasing demand for online education have driven a shift toward telematics-based work and study models. This transformation has led education to move from traditional classrooms to virtual learning environments, which facilitate digital content delivery and enable interaction with students worldwide [14]. However, the ease of accessing information online has contributed to a rise in plagiarism among students. Common practices include copying and pasting verbatim text, paraphrasing without proper citation, and using online tools to bypass originality checks. These behaviors hinder

the development of critical thinking, analytical reasoning, and writing skills—key competencies for academic success. This issue is particularly concerning in assignments such as undergraduate and master's theses, where originality and independent research are paramount.

Two primary ethical concerns affect academic writing, particularly in undergraduate and postgraduate reports: plagiarism and false authorship. Plagiarism occurs when content is copied from existing sources without proper citation, often assessed by measuring textual similarity to existing databases. Universities typically establish thresholds for acceptable similarity levels (e.g., a maximum of 15% similarity without citation). False authorship, on the other hand, involves presenting someone else's work as one's own, either by commissioning it or by using another person's ideas without proper credit. Both issues undermine the integrity of academic research.

To address plagiarism, universities employ similarity detection tools such as Turnitin, which compare submitted work against existing sources. However, detecting false authorship—such as contract cheating—is more complex and difficult to prove, even when suspected. This raises a critical research question: Can an artificial intelligence (AI) tool based on text mining and machine learning classify writing styles in academic reports to identify potential cases of false authorship?

This study aims to develop a predictive system that analyzes content to classify scientific writing styles in digital academic reports. Although this research is motivated by concerns regarding student theses, it utilizes reports and papers authored by professors at the Pontificia Universidad Católica del Ecuador Sede Esmeraldas (PUCESE) to train the AI model. The model is designed for future application to student reports, with the potential for expansion to undergraduate and graduate theses. Due to the lack of historical reports tracking students' academic progress, this study does not involve actual student cases. However, the classifier's logic can be applied similarly to student papers. By analyzing professors' manuscripts, this research establishes a foundation for addressing academic integrity issues in student work.

PUCESE stands to benefit from validating not only textual similarity but also the authenticity of writing styles in academic reports. This novel system holds promise for universities worldwide, with PUCESE's Academic and Research Department, administrative bodies, and faculty advisors as the primary beneficiaries.

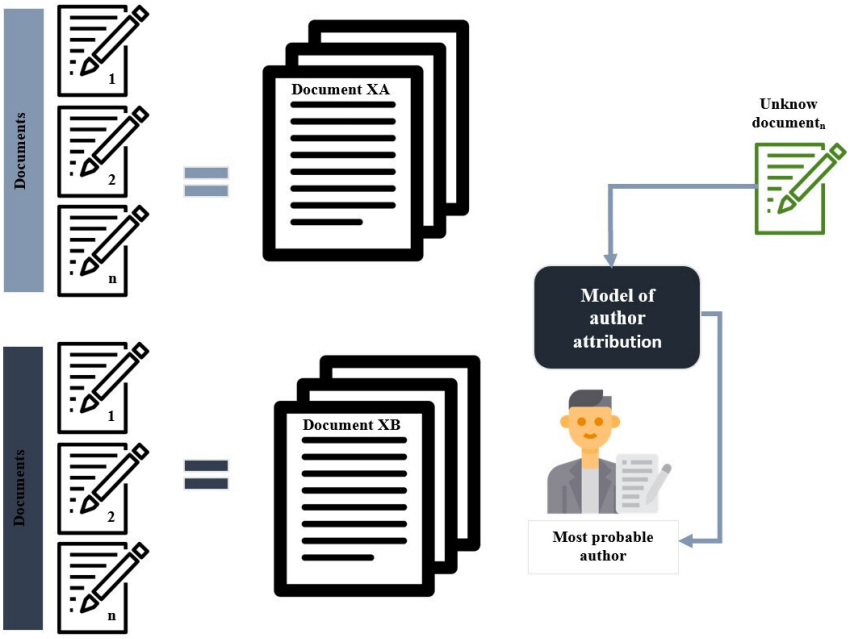
This paper is organized as follows: Section 2 presents the theoretical framework of the study. Section 3 introduces authorship attribution and machine learning-based classification systems. Section 4 describes the system's design. Section 5 details the training results of the classifier in terms of the accuracy metric. Finally, Section 6 presents the conclusions and future work.

THEORETICAL BACKGROUND

Attribution of Authorship

Authorship analysis, a field of growing interest, has made significant contributions to areas such as homeland security, intelligence, and market analysis [15]. It focuses on the automatic classification of texts based on authors' writing styles, encompassing tasks such as authorship attribution (identifying the author) and plagiarism detection. This study leverages the concept of authorship analysis, specifically authorship attribution, to identify potential cases of false authorship in academic reports.

Within authorship analysis, two primary approaches have been proposed for determining authorship based on writing style. The first approach, known as the profile-based method, aggregates all of an author's documents into a single dataset for training. This method creates a characteristic profile of the author's writing style, as illustrated in Figure 1 from Ramírez et al. [16]. However, it requires a substantial volume of documents from a single author, which may not always be available.



Source: [12].

Figure 1. Author profile approach

The second approach, instance-based, focuses on individual documents. Each document is transformed into a vector representation, from which features are extracted

to train the model. This method enables the model to predict the authorship of unknown texts. While Ramírez et al. [16] suggested using a single document, such as an abstract, to capture an author's writing style, our study proposes leveraging a larger set of documents to provide a more comprehensive representation of each author's writing style.

Furthermore, authorship analysis encompasses subfields such as author profiling (AP), which aims to identify patterns shared by groups of authors based on attributes such as age, gender, or political orientation [15]. This complements authorship attribution (AA) by providing additional insights into the potential author of a text.

Text Mining

Text mining involves extracting valuable information from natural language text by identifying patterns and insights to address specific questions or objectives [17]. The process typically consists of multiple steps: first, preprocessing the text by cleaning and transforming it into a structured format, followed by the application of analytical techniques to uncover meaningful information.

Several techniques are commonly employed in text mining to extract relevant insights from textual data. Three prominent methods include [18]:

- Boolean Method: This approach represents documents using keywords and evaluates their presence or absence. While effective for initial searches, it may overlook relevant documents due to its strict matching criteria.
- Latent Semantic Analysis (LSA): LSA examines the underlying relationships between words in a vector space, enabling the identification of hidden connections and concepts within the text. This method is useful for performing tasks such as topic modeling, but requires greater computational resources than the Boolean method.
- Semantic Vector Space Model: Building on the Boolean approach, this model decomposes words into distributional and compositional semantic components, allowing for a more nuanced understanding of word meanings and improving the accuracy of text analysis tasks.

Additionally, text mining employs advanced techniques that complement these foundational methods. Two notable approaches are Term Frequency-Inverse Document Frequency (TF-IDF) and Word Embedding:

- TF-IDF evaluates the relative importance of terms within a document by considering both their frequency within the specific document and their rarity across the entire collection [19].
- Word Embedding represents words as mathematical vectors in a continuous space, capturing semantic and syntactic relationships in specific contexts [20].

These advanced techniques facilitate various natural language processing (NLP) tasks, such as sentiment analysis and machine translation [21], [22], [23]. Due to their ability to capture nuanced linguistic relationships, TF-IDF and Word Embedding are utilized in this study to address the challenge of identifying potential cases of false authorship in academic reports.

METHODOLOGY

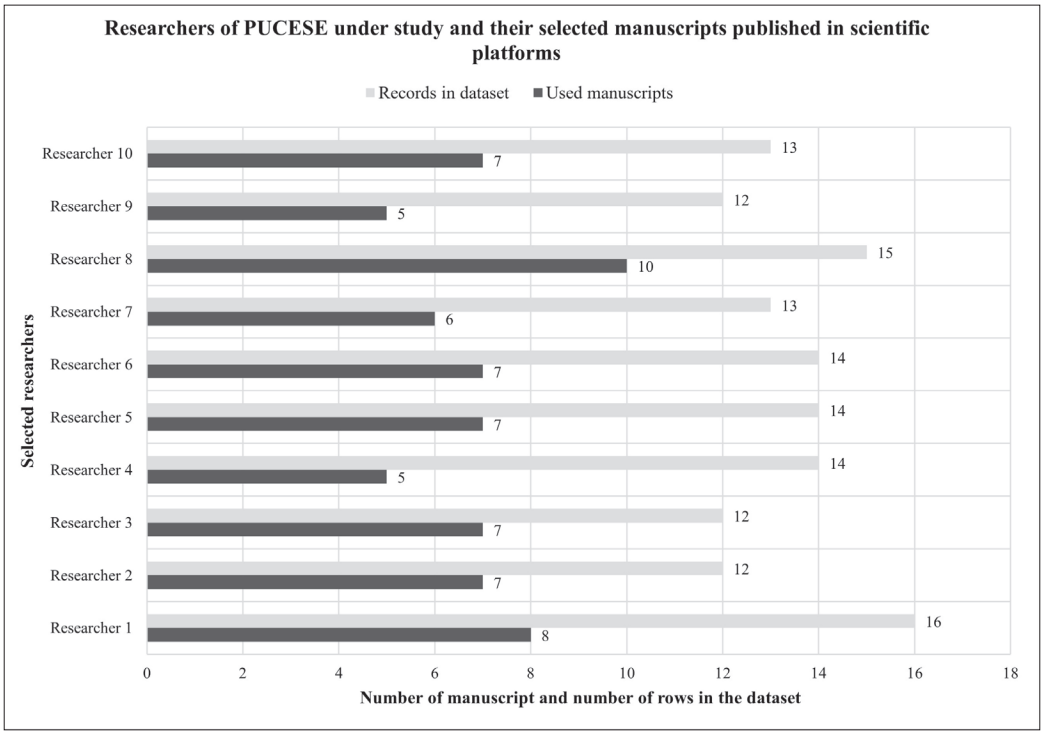
Research Design

To address the research question, our study employed a mixed-method approach, integrating qualitative and experimental components. Subsequently, a deductive approach was used during the development phase to apply the acquired knowledge in building a model for classifying writing styles. The proposed model was trained using a dataset of research papers and reports published by PUCESE faculty members.

Our study focused on analyzing the writing styles of faculty members at PUCESE as reflected in their published works. Convenience sampling was used to select participants. Faculty members were eligible for inclusion if they had published at least five articles in indexed scientific journals or conference proceedings, as documented on their ResearchGate profiles. Seventeen professors had at least one publication; however, only 10 met the criterion of having at least five indexed papers written in Spanish. These 10 professors were selected for analysis. They were adjunct professors in the following programs: Nursing, Information Technology, Environmental Engineering, Accounting and Auditing, Education, and Graphic Design. Additionally, six of the selected professors held PhD degrees, while the remaining four held master's degrees.

Professors who did not meet the inclusion criterion were excluded because the proposed model requires a sufficient amount of data for effective training. Including individuals with very few samples could lead to unreliable predictions, such as false positives or false negatives, as the model might struggle to generalize accurately from limited data. This aligns with machine learning principles, where a larger, more representative dataset enhances model performance and reduces the risk of

misclassification. The distribution of the selected authors, their publications, and the dataset items is presented in Figure 2.



Source: own elaboration.

Figure 2. Researchers under study and their published papers

To assess the effectiveness of the classifier, a confusion matrix was employed alongside standard machine learning metrics such as accuracy and F1-score. Accuracy measures the proportion of correct predictions out of the total predictions, making it useful for balanced datasets. The F1-score, the harmonic mean of precision and recall, provides a balanced evaluation, particularly in imbalanced datasets where accuracy alone can be misleading [24].

Programming Tools and Datasets

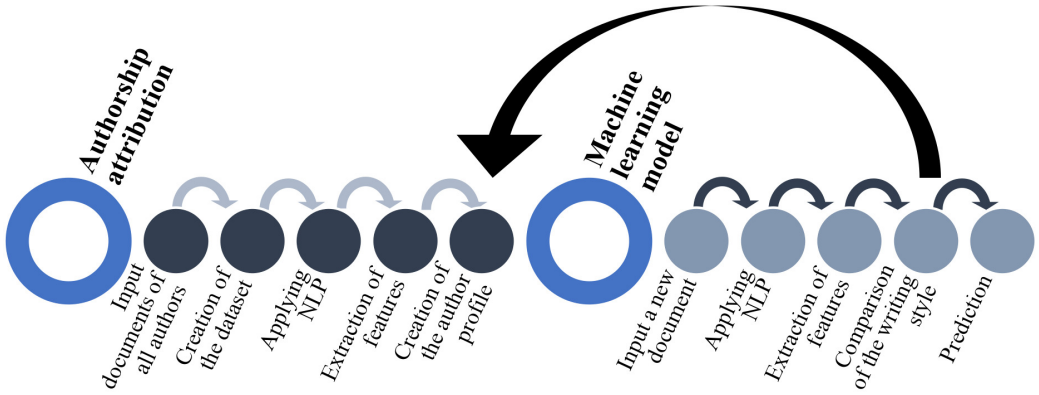
The development process leveraged Python, a cross-platform programming language with extensive libraries for data analysis. Specifically, the Natural Language Toolkit (NLTK) was used for text analysis tasks, while Scikit-learn (Sklearn), an open-source machine learning library, served as the foundation for building the writing style classification model.

For authorship analysis, the author profile approach was adopted (Figure 1). This process involved collecting manuscripts from each author under study. After applying preprocessing steps to ensure text consistency—such as removing formatting inconsistencies and special characters—all manuscripts by each author were concatenated into a single text file. These processed texts were then organized into a CSV file, forming the dataset used to train the classifier.

The dataset, named "authorship_dataset.csv", contained four columns: manuscript title, section of the manuscript (e.g., Introduction, Conclusions, Discussion, or Abstract), text from a specific section, and author identifier (labeled from 1 to 10). The dataset comprised 69 manuscripts, totaling 135 text samples from key sections of the manuscripts—those most representative of an author's distinctive writing style. This structured data allowed the model to learn and differentiate the unique characteristics of each author's writing style.

System Overview

The proposed system, illustrated in Figure 3, consists of two main stages. The first stage focuses on preparing the dataset using the author profile approach described earlier. The second stage involves the creation of a prediction model for authorship analysis.



Source: own elaboration.

Figure 3. Systematic process followed to develop the system

The first stage of the system is dedicated to preparing the dataset for the authorship analysis task. This involves preprocessing the data extracted from articles published by the selected authors. The preprocessing steps typically include cleaning the text by removing irrelevant characters and formatting inconsistencies. After cleaning,

the system extracts specific sections for analysis from each author's documents. In this study, the Introduction, Discussion, Conclusions, and in some cases, the Abstract sections were selected. These sections were chosen because they summarize the research topic and reflect the author's writing style, making them strong indicators of authorship. The extracted text is then used to create the dataset that serves as input for the classifier.

The prepared dataset is a text database structured into four columns. However, for model training, only two columns were utilized: "text of the specific section" (input feature) and "author" (target variable). The "author" column serves as the target variable for supervised learning, allowing the model to learn how to associate text samples with their corresponding authors.

Supervised learning models are particularly suitable for this task, as they map input text data (manuscripts) to the respective authors based on labeled examples. The dataset was divided into 80% for training and 20% for testing, a widely used practice in machine learning to ensure that the model learns effectively while retaining a portion of the data for evaluation on unseen samples.

The second stage, also illustrated in Figure 3, focuses on designing, training, and implementing the machine learning model. Before feeding the text data into the model, it undergoes several preprocessing steps to ensure compatibility with the algorithm's requirements. These steps allow the model to analyze the text effectively.

One essential preprocessing step is tokenization, which splits the text into individual units such as words and punctuation marks. Additionally, stop words—common words like "the" or "and" that provide little contextual meaning [25]—are removed to reduce noise in the dataset and improve model performance. Finally, lemmatization is applied, transforming words into their base form (e.g., "running" → "run") [26]. Lemmatization is preferred over stemming because it considers grammatical context, ensuring that words remain meaningful and correctly interpreted by the model.

Once the text was preprocessed, it was converted into a form suitable for the machine learning model. Two key approaches were employed: TF-IDF [19] and word embedding [20]. TF-IDF assigns a numerical value to each word based on its frequency within a document and its rarity across the entire dataset [19] . This helps the model understand the importance of each word in the context of a specific author's writing style. Word embedding, on the other hand, leverages deep learning algorithms to capture the semantic relationships between words [20]. This enables the model to go beyond word frequency and understand the nuanced meanings of words based

on their context. By combining these techniques, the model gains a comprehensive textual representation, improving accuracy in authorship classification.

The selection of a machine learning algorithm is a crucial factor that significantly impacts performance. In this study, multiple text classification algorithms were explored, including Linear Support Vector Machines (SVM), RBF SVC, Random Forest, Decision Tree, Logistic Regression, k-Nearest Neighbors (k-NN), and Naïve Bayes. These algorithms are depicted in [27]data science has positioned as an area of interest for decision makers in many organizations. Advances in Machine Learning (ML, and their training is detailed in the GitHub project available at: https://shorturl.at/zzVCQ.

The training process involved feeding the model with preprocessed training data, consisting of text features (from research manuscripts) and corresponding author labels. During training, a hyperparameter optimization procedure was applied to enhance performance. This involved testing different model configurations to optimize predictive capability. Ultimately, the best-performing model was selected for evaluation.

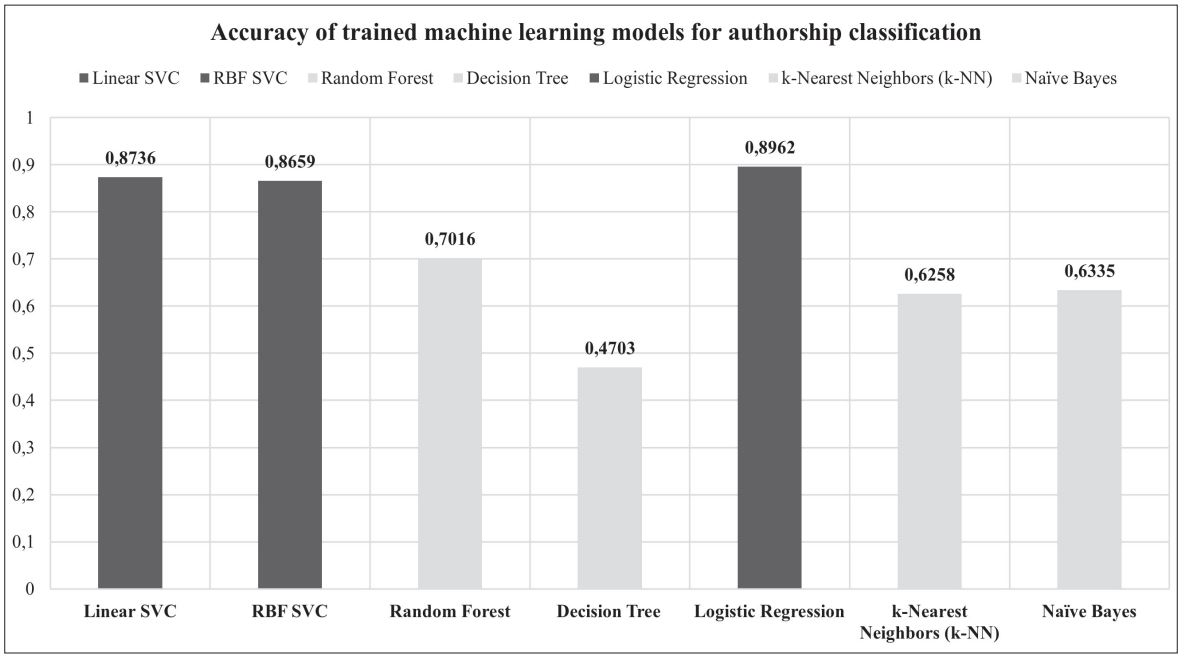
After training, the model's performance was evaluated using the test dataset. To assess its effectiveness, metrics such as accuracy and F1-score were computed. These metrics were obtained through cross-validation, ensuring robust model evaluation. The results are discussed in Section 5.

RESULTS AND DISCUSSION

Our study evaluated the performance of the classification models using accuracy as the primary metric. The results obtained for each of the seven machine learning algorithms indicate that the models trained with Linear SVC (87.36%), RBF SVC (86.59%), and Logistic Regression (89.62%) adapted best to the dataset used in this study. In contrast, the Decision Tree-based algorithm exhibited the poorest performance. Similarly, although achieving accuracy rates above 60%, k-NN and Naïve Bayes demonstrated only moderate performance. More detailed results for these and all remaining trained models are illustrated in Figure 4.

After training all models, we evaluated their performance using cross-validation technique—a fundamental approach in machine learning that ensures a robust and reliable assessment of model performance. Specifically, we applied 10-fold cross-validation, obtaining 10 evaluation results for each model. This method allowed us to estimate model performance more accurately, reducing the impact of variability that may arise from a single data split and ensuring that the trained models generali-

ze well. The dataset was divided into 10 equal subsets, with each model being trained and tested on different portions of the data.

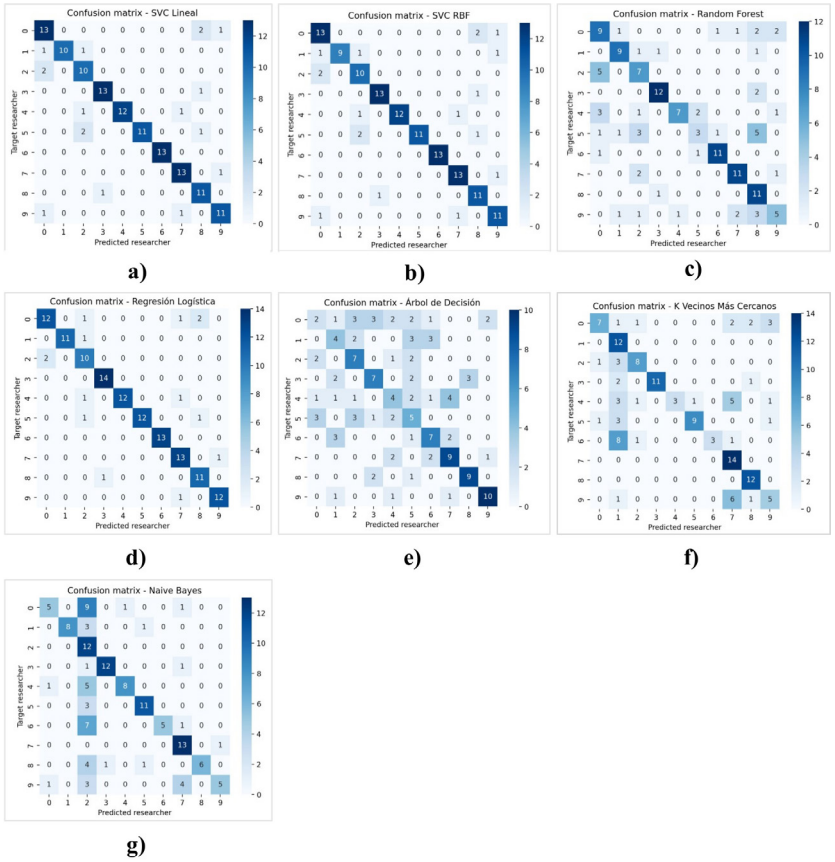


Source: own elaboration.

Figure 4. Accuracy obtained for trained models for authorship attribution

To further analyze the classification performance, a confusion matrix was used to visualize how each model assigned instances to their respective categories. This technique provides valuable insights into misclassification patterns. The confusion matrices for all models are shown in Figure 5.

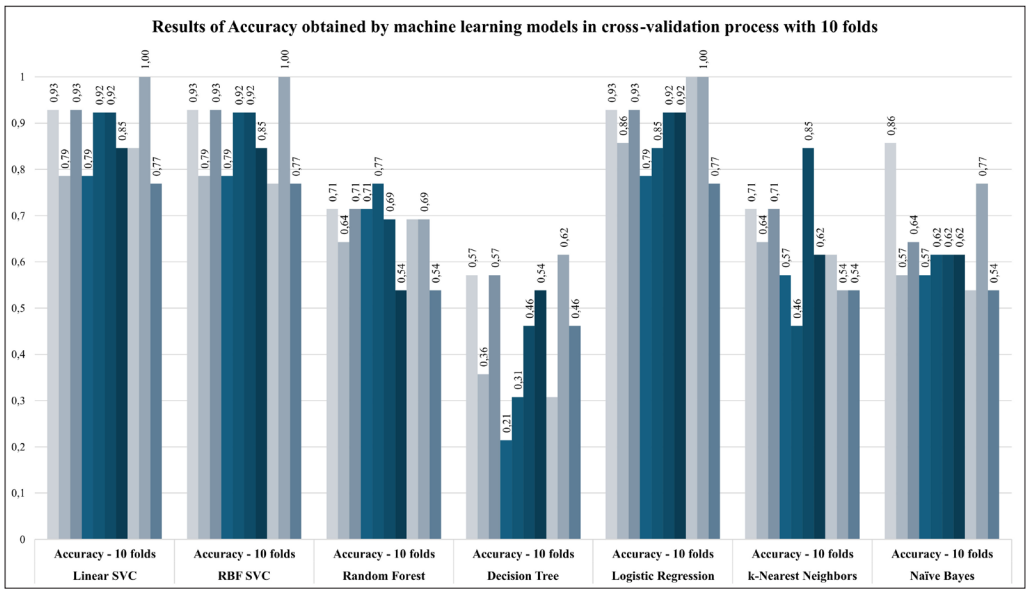
Additionally, Figure 6 and 7 present the results of accuracy and F1-score obtained in the evaluation process using cross-validation. Based on these results, the authorship attribution models were compared using the Friedman test to determine whether significant differences exist in the average performance between the models [28].



Source: own elaboration.

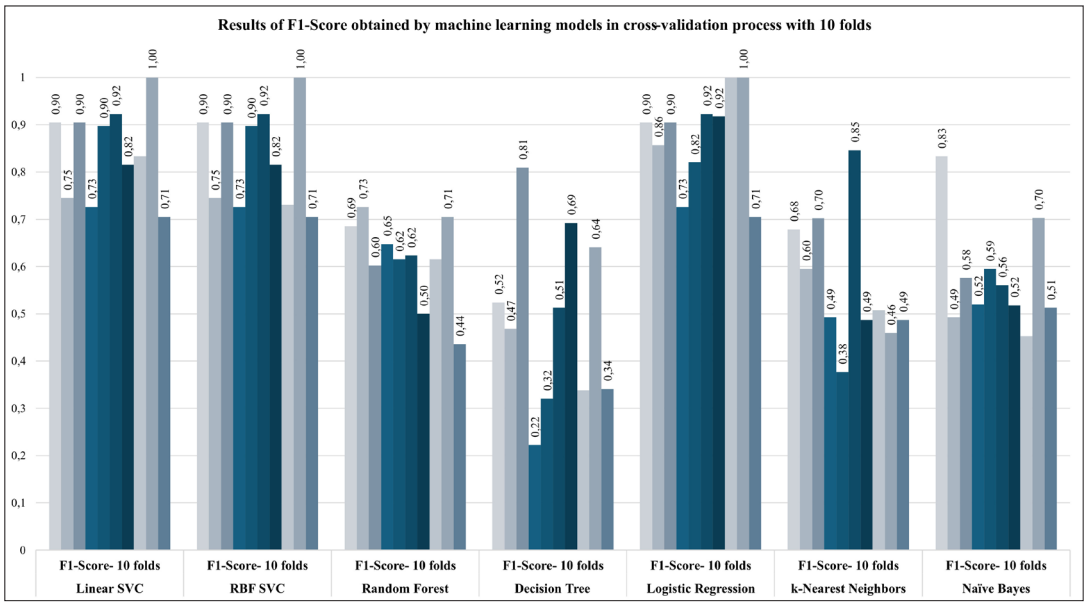
Figure 5. Confusion matrix of trained models using: a) Linear SVC, b) RBF kernel SVC, c) Random Forests, d) Logistic Regression, e) Decision Tree, f) k-NN, g) Naïve Bayes

The results of the Friedman test on accuracy data yielded a test statistic of 54.13 with a p-value of 6.96e-10, which is significantly lower than the conventional significance threshold ($\alpha = 0.05$). This strong statistical evidence indicates that at least one of the models performs significantly differently from the others. Similarly, for F1-score data, the Friedman test yielded a test statistic of 50.08 with a p-value of 4.52e-09, also demonstrating that the models exhibited statistically significant differences.



Source: own elaboration.

Figure 6. Results of accuracy obtained for trained models in cross-validation process



Source: own elaboration.

Figure 7. Results of F1-Score obtained for trained models in the cross-validation process

Because the Friedman test indicated an overall significant difference among the seven trained models, a post-hoc analysis using the Wilcoxon signed-rank test was conducted to pinpoint specific differences in predictive performance between pairs of these models [28].

The Wilcoxon signed-rank test results based on accuracy data (Table 1) indicate that Linear SVC, RBF SVC, and Logistic Regression perform significantly better than Random Forest, Decision Tree, k-NN, and Naïve Bayes ($p<0.05$). In contrast, Random Forest, k-NN, and Naïve Bayes do not show significant differences among themselves ($p > 0.05$), suggesting similar performance levels. Additionally, Linear SVC and RBF SVC do not exhibit significant differences between them, nor do they differ significantly from Logistic Regression. These findings reinforce the suitability of Linear SVC, RBF SVC, and Logistic Regression for authorship classification tasks, while highlighting the limitations of tree-based models and probabilistic classifiers in this context.

Table 1. P-values obtained from the Wilcoxon signed-rank test comparing model accuracy

	LSVC	RBF SVC	RF	DT	LR	K-NN	NB
Linear SVC (LSVC)	1.0000	0.3173	0.0020	0.0020	0.3573	0.0020	0.0020
RBF SVC (RBF-SVC)	0.3173	1.0000	0.0020	0.0020	0.3573	0.0020	0.0020
Random Forest (RF)	0.0020	0.0020	1.0000	0.0076	0.0020	0.4631	0.4413
Decision Tree (DT)	0.0020	0.0020	0.0076	1.0000	0.0020	0.0059	0.0020
Logistic Regression (LR)	0.3573	0.3573	0.0020	0.0020	1.0000	0.0020	0.0020
K-NN	0.0020	0.0020	0.4631	0.0059	0.0020	1.0000	0.7995
Naïve Bayes (NB)	0.0020	0.0020	0.4413	0.0020	0.0020	0.7995	1.0000

Source: own elaboration.

Based on the F1-Score analysis presented in Table 2, the Wilcoxon test indicates that Linear SVC, RBF SVC, and Logistic Regression significantly outperform ($p < 0.005$) Random Forest, Decision Tree, k-NN, and Naïve Bayes in achieving a balanced precision and recall. While Linear SVC and RBF SVC show comparable F1-Scores, the latter group generally does not exhibit significant differences among themselves ($p \geq 0.005$), suggesting a similar, though lower, level of balanced performance. These results underscore the superior effectiveness of Linear SVC, RBF SVC, and Logistic Regression in optimizing the F1-Score for this task.

Table 2. P-values obtained from the Wilcoxon signed-rank test comparing model F1-Score

	LSVC	RBF SVC	RF	DT	LR	K-NN	NB
Linear SVC (LSVC)	1.0000	0.3170	0.0019	0.0019	0.144	0.0019	0.0019
RBF SVC (RBF-SVC)	0.3170	1.000	0.0019	0.0019	0.144	0.0019	0.0019
Random Forest (RF)	0.0019	0.0019	1.000	0.105	0.0019	0.232	0.275
Decision Tree (DT)	0.0019	0.0019	0.105	1.000	0.0019	0.322	0.160
Logistic Regression (LR)	0.144	0.144	0.0019	0.0019	1.000	0.0019	0.0019
K-NN	0.0019	0.0019	0.232	0.322	0.0019	1.000	0.846
Naïve Bayes (NB)	0.0019	0.0019	0.275	0.160	0.0019	0.846	1.000

Source: own elaboration.

CONCLUSIONS AND FUTURE WORK

The findings of this study confirm that Linear SVC, RBF SVC, and Logistic Regression are the most effective machine learning models for authorship classification in academic reports, significantly outperforming Random Forest, Decision Tree, k-NN, and Naïve Bayes ($p < 0.05$). These results reinforce the suitability of support vector-based models and logistic regression for detecting writing style patterns, while highlighting the limitations of tree-based and probabilistic classifiers in this context. The system achieved an accuracy of 89.62% using Logistic Regression, demonstrating strong potential for real-world applications, though further refinement is needed to enhance reliability. Notably, Logistic Regression did not exhibit significant differences from Linear SVC and RBF SVC, which achieved 87.36% and 86.59% accuracy, respectively.

Our research serves as a foundational step for PUCESE in implementing a writing style-based authorship verification system to complement existing plagiarism detection tools such as Turnitin. However, to maximize the system's effectiveness, we strongly recommend establishing a student work repository from the beginning of their degree programs. This repository, built using assignments submitted through Moodle, would provide a historical dataset necessary for training more robust models capable of verifying whether a thesis was genuinely authored by the student.

Future research should focus on expanding the range of machine learning models, particularly deep learning approaches such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based models (e.g., BERT, GPT),

which could further improve classification accuracy. Additionally, developing larger and more balanced datasets in both English and Spanish would enhance the model's generalizability, allowing for broader application across different academic institutions and disciplines.

Lastly, integrating writing style analysis with other authorship verification techniques, such as keystroke dynamics, revision history tracking, and semantic similarity analysis, could strengthen the system's ability to detect false authorship more accurately. Further studies should also explore the ethical and institutional implications of AI-driven authorship verification to ensure fair, unbiased, and privacy-compliant implementation within academic integrity frameworks.

ACKNOWLEDGMENTS

We sincerely thank the Concurrent Systems Group at the University of Granada for their invaluable support.

REFERENCES

[1] A. Korkmaz, C. Aktürk, and T. Talan, "Analyzing the User's Sentiments of ChatGPT Using Twitter Data -," *Iraqi J. Comput. Sci. Math.*, vol. 4, no. 2, pp. 202–214, 2023.

[2] A. Arias, Y. Mattos, J. Heredia, and D. Heredia, "Minería de texto como una herramienta para la búsqueda de artículos científicos para la investigación," *Rev. I+D en TI*, vol. 7, no. 1, pp. 14–20, 2017.

[3] A. Zanasi, "Virtual Weapons for Real Wars: Text Mining for National Security," in *Proceedings of the International Workshop on Computational Intelligence in Security for Information Systems CISIS'08*, 2009, vol. 53, pp. 53–60.

[4] R. Bridgelall, "An Application of Natural Language Processing to Classify What Terrorists Say They Want," *Soc. Sci.*, vol. 11, no. 1, pp. 1–15, 2022.

[5] Jufri and M. Thamrin, "Political Influence Analysis Social Media Text Mining for Public Opinion: Case Study Makassar City," in *2021 3rd International Conference on Cybernetics and Intelligent System* (ICORIS), 2021, pp. 1–5.

[6] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger, "Pulse: Mining Customer Opinions from Free Text," in *Advances in Intelligent Data Analysis* VI. IDA 2005, 2005, pp. 121–132.

[7] S. Jardim and C. Mora, "Customer reviews sentiment-based analysis and clustering for market-oriented tourism services and products development or positioning," *Procedia Comput. Sci.*, vol. 196, no. 2021, pp. 199–206, 2021.

[8] D. Mittal and S. R. Agrawal, "Determining banking service attributes from online reviews: text mining and sentiment analysis," *Int. J. Bank Mark.*, vol. 40, no. 3, pp. 558–577, 2022.

[9] S. Chatterjee, D. Goyal, A. Prakash, and J. Sharma, "Exploring healthcare/health-product ecommerce satisfaction: A text mining and machine learning application," *J. Bus. Res.*, vol. 131, no. October 2020, pp. 815–825, 2021.

[10] M. C. Barrera, "Minería de texto en la clasificación de material bibliográfico," *Biblios*, no. 64, pp. 33–43, 2016.

[11] R. Ferreira-Mello, M. André, A. Pinheiro, E. Costa, and C. Romero, "Text mining in education," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 6, 2019.

[12] J. Villalón, P. Kearney, R. A. Calvo, and P. Reimann, "Glosser: Enhanced feedback for student writing tasks," in *Proceedings - The 8th IEEE International Conference on Advanced Learning Technologies, ICALT 2008*, 2008, no. 1, pp. 454–458.

[13] E. Hossain *et al.*, "Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review," *Comput. Biol. Med.*, vol. 155, pp. 1–24, 2023.

[14] G. Aciar, S. Aciar, and C. González, "Analítica del aprendizaje: método automático para identificar sentencias que contienen información positiva y negativa utilizando técnicas de minería de texto," in *VIII Jornadas Internacionales de Campus Virtuales (JICV'18)*, 2018.

[15] V. Mercado, A. Villagra, and M. Errecalde, "El Proceso de Extracción de Conocimiento en la Determinación del Perfil del Autor y la Atribución de Autoría," in *XIX Workshop de Investigadores en Ciencias de la Computación (WICC 2017, ITBA, Buenos Aires)*, 2017, pp. 261–265.

[16] M. Ramírez, J. Carillo, and M. Somodevilla, "Atribución de autoría combinando información léxico-sintáctica mediante representaciones holográficas reducidas," *Res. Comput. Sci.*, vol. 88, pp. 103–113, 2014.

[17] K. Thakur and V. Kumar, "Application of Text Mining Techniques on Scholarly Research Articles: Methods and Tools," *New Rev. Acad. Librariansh.*, vol. 28, no. 3, pp. 279–302, 2022.

[18] I. Valero, "Técnicas estadísticas en Minería de Textos," Universidad de Sevilla, 2017.

[19] A. A. Jalal and B. H. Ali, "Text documents clustering using data mining techniques," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 1, pp. 664–670, 2021.

[20] S. Selva Birunda and R. Kanniga Devi, *A review on word embedding techniques for text classification*, vol. 59. Springer Singapore, 2021.

[21] M. Ruiz, "Implementación de un sistema de diálogo automático como asistente en el proceso administrativo del examen de traductor e intérprete oficial de la Universidad de Antioquia," Universidad de Antioquia, 2020.

[22] G. Liberatore, A. Vuotto, and G. Fernández, "Desarrollo de una herramienta para el análisis y representación semántica de colecciones documentales a través del factor TF-IDF," in *Jornadas Temas Actuales en Bibliotecología*, 2018.

[23] A. Cardoso, L. Talame, M. Amor, and A. Monge, "Aplicación de técnicas avanzadas de aprendizaje automático para identificar emociones en textos," in XXIII *Workshop de Investigadores en Ciencias de la Computación*, 2021, pp. 73–77.

[24] G. Naidu, T. Zuva, and E. M. Sibanda, *A Review of Evaluation Metrics in Machine Learning Algorithms*, vol. 724 LNNS. Springer International Publishing, 2023.

[25] S. Sarica and J. Luo, "Stopwords in technical language processing," *PLoS One*, vol. 16, no. 8 August, pp. 1–13, 2021.

[26] Z. Abidin, A. Junaidi, and Wamiliana, "Text Stemming and Lemmatization of Regional Languages in Indonesia: A Systematic Literature Review," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 10, no. 2, pp. 217–231, 2024.

[27] P. Pico-Valencia, O. Vinueza-Celi, and J. A. Holgado-Terriza, "Bringing Machine Learning Predictive Models Based on Machine Learning Closer to Non-technical Users," in *Advances in Intelligent Systems and Computing*, 2021, vol. 1273 AISC, pp. 3–15.

[28] D. G. Pereira, A. Afonso, and F. M. Medeiros, "Overview of Friedman's Test and Post-hoc Analysis," *Commun. Stat. - Simul. Comput.*, vol. 44, no. 10, pp. 2636–2653, 2015.