Authorship Attribution with Rejection Capability in Challenging Contexts of Limited Datasets

Pedro M. Oliveira and **Joaquim F. Silva**
NOVA LINCS, NOVA School of Science and Technology, 2829-516, Caparica, Portugal
pmr.oliveira@fct.unl.pt and jfs@fct.unl.pt

Abstract

Attributing authorship to text can be a complex problem for both specialists and AI systems. This difficulty arises from challenges like capturing distinct writing styles and authors, handling texts from the same era and languages, or distinct heteronyms of the same writer, or identifying the author's gender. Traditionally, solutions for authorship attribution have required the extraction of numerous attributes, frequently obtained through specialized linguistic tools, coupled with the availability of extensive training documents. The advent of Deep Learning transformers has further amplified this reliance on data quantity.

Classic classification approaches usually assign a class to documents to be classified, even if they are too strange concerning the classes learned in the training phase. However those strange texts should be rejected based on founded approaches, in order to enhance the classifiers reliability.

This paper proposes a language independent approach to authorship attribution with the capability to reject strange samples, in challenging contexts, achieving high accuracies for all tested datasets. By assessing the discriminating ability of each attribute, the final set of features can be strongly reduced.

1 Introduction

The Attribution of Authorship (AA) with high Accuracy presently find significant utility in areas such as plagiarism detection, copyright protection, and cybercrime investigation. Over the years, various approaches have emerged to tackle this challenge, with efforts aimed at achieving more promising results (Koppel et al., 2003; Potha and Rao, 2018; Keskin and Adali, 2019). Despite these advancements, a comprehensive solution that can attribute authorship to documents within challenging contexts, without relying on linguistic tools and the need to infer the language, prevails to be found.

Developing a universal authorship attribution solution faces challenges due to the absence of cross-linguistic capabilities, and achieving clear differentiation among known authors is a priority. However, a common limitation is the failure to assess attribute discriminative potential on a per-dataset basis, hindering automation efforts.

The primary objective of this paper is to propose a supervised classification language-independent system tailored for challenges like capturing distinct writing styles and authors, handling texts from the same era and languages, or distinct heteronyms of the same writer, or identifying the author's gender. The approach uses no linguistic tools and assesses the discriminating ability of the potential attributes. Furthermore, our proposal includes a mechanism to reject unknown documents, being useful for cases where confident classification is impractical but essential. The subsequent sections delve into the specifics of our proposed approach, encompassing the methodology, experimental setup, insightful results and conclusions.

2 State of the Art

Text classification is an extensively researched area, with recent focus on AA and author gender classification (Koppel et al., 2003; Potha and Rao, 2018; Keskin and Adali, 2019). There's no universal feature set applicable to all contexts (Iqbal et al., 2010). Studies (Elmanarelbouanani and Kassou, 2013; Gamon, 2004) highlight that the AA classification process depends on various indicators, including *corpus* size, document size, class count, as well as author characteristics like age, nationality, and gender. In this context, we emphasize the necessity of acquiring attributes that effectively discriminate among authors. While some methods (Zipf, 1932; Iqbal et al., 2010; Abbasi and Chen, 2008) identify document similarities to group them, these approaches may struggle with small datasets. Alternatively, graph-based methods (Gomez Adorno et al.,

2015) represent documents as graphs, extracting features for similarity calculations. However, these techniques might not be language-independent and could falter with limited author-document samples.

Statistical approaches, as seen in (Kešelj et al., 2003; Howedi, 2014), gather attributes for classification, yielding up to 90% F-measure. Yet, they often treat attributes equally significant regardless of the dataset, leading to suboptimal performance in challenging scenarios. Evaluating the discriminant power of attributes is crucial for successful classification (Stamatatos, 2009; Ouamour and Sayoud, 2012), but can demand large training texts/documents.

Although AA involving heteronyms arises extra challenging complexity as texts are penned by a single writer, in (Teixeira and Couto, 2015) authors tackled this problem with attributes from different techniques. While achieving high Accuracy, the study dealt with only two heteronyms, whereas popular cases involve more. They collected 8941 attributes, later reduced to 4398, underlining the challenge of handling numerous attributes.

In author gender identification, linguistic distinctions (Argamon et al., 2003) and ensemble learning (Garg et al., 2018; Zhao and Li, 2018) have shown promise. Although, these studies achieved 80% to 92.5% Accuracy, by being based on vocabulary and syntax, they are language-dependent. Deep Learning transformers have gained traction in text classification, but high Accuracy often demands substantial *corpora* sizes (Glorot and Bengio, 2010). (Rodrigues et al., 2023) developed the Albertina, an encoder which can potentially be used for text classification, although, it is suited for just one language (Portuguese).

Some traditional classifiers provide output confidence scores when classifying samples, which are commonly used to empirically set a threshold to decide about the rejection of strange samples (Gritsenko and Smirnov, 2008). However, this is not a founded method as this threshold may vary with the context where the samples lie and the number of the classes. So, this is a problem that requires a reasoned approach.

	Alberto Caeiro	Álvaro de Campos
Original	*Todos dias agora acordo com alegria e pena.*	*No tempo em que festejavam o dia dos meus anos,*
	Antigamente acordava sem sensação nenhuma; acordava.	*Eu era feliz e ninguém estava morto.*
	Tenho alegria e pena porque perco o que sonho	*Na casa antiga, até eu fazer anos era uma tradição de há séculos,*
	E posso estar na realidade onde está o que sonho.	*E a alegria de todos, e a minha, estava certa com uma religião qualquer.*
	Não sei o que hei-de fazer das minhas sensações.	*No tempo em que festejavam o dia dos meus anos*
	Não sei o que hei-de ser sozinho.	*Eu tinha a grande saúde de não perceber coisa nenhuma,*
	Quero que ela me diga qualquer coisa para eu acordar de novo.	*De ser inteligente para entre a família,*
	Quem ama é diferente de quem é	*E de não ter as esperanças que os outros tinham por mim.*
	É a mesma pessoa sem ninguém.	*Quando vim a ter esperanças, já não sabia ter esperanças. (...)*
English	Every day now I wake up with joy and sorrow.	In the days when they celebrated my birthday,
	I used to wake up with no feeling; I woke up.	I was happy and nobody was dead.
	I have joy and sorrow because I lose what I dream.	In the old house, until I turned years old,
	And I can be in the reality where what I dream is.	It was a centuries-old tradition,
	I don't know what to do with my sensations.	And everyone's joy, and mine, was certain with any religion.
	I don't know what to be alone.	When they celebrated my birthday
	I want her to tell me something to wake me up again.	I had the great health of not noticing anything,
	Whoever loves is different from who is.	From being smart to among the family,
	It is the same person without anyone.	And not having the hopes that others had for me.(...)

Table 1: Example of two documents produced by two heteronyms of the same writer, *Fernando Pessoa*.

Feature/Attribute	Description
9-char	Relative frequency of words per document whose length is greater than or equal to nine
6-char	Relative frequency of words per document whose length is greater than or equal to six
3-char	Relative frequency of words per document whose length is less than three
5-char	Relative frequency of words per document whose length is between three and five
2-char	Relative frequency of words per document whose length is two
1-grams	Relative frequency of the most repeated 1-grams
2-grams	Relative frequency of the most repeated 2-gram
3-grams	Relative frequency of the most repeated 3-grams
4-grams	Relative frequency of the most repeated 4-grams
Syllabic variance	Syllabic variance of text blocks
Commas	Relative frequency of comma usage
Periods	Relative frequency of period usage
Hyphen	Relative frequency of hyphen usage
Non-ascii	Relative frequency of non-ascii characters in the document
Capital letters	Relative frequency of uppercase character usage in the document
Average word length	Average length of each word
Average block length	Average length of each text block
Exclamation	Relative frequency of exclamation point usage
Question mark	Relative frequency of question mark usage
Semicolon	Relative frequency of semicolon usage normalized
Text between commas	Average number of words between two consecutive commas
Text between question marks	Average number of words between consecutive question marks
Text between exclamation points	Average number of words between two consecutive exclamation poitns
Text between periods	Average number of words within two consecutive periods
Q	Normalized occurrence of the char Q
K	Relative frequency of the char K
Different words	Normalized number of different words per document
&	Relative frequency of the character & usage normalized

Table 2: Potentially discriminant attributes used in the proposed solution.

3 Feature Extraction

Finding features with sufficient discriminant power that can characterize and differentiate authors, can prove to be a difficult task, since the writing patterns between authors can be very tenuous. In fact, considering a context such as the identification of several heteronyms of the same writer (see example in Table 1), where the differences in the writing of the two heteronyms can be very subtle, it would be possible to capture some of these differences, eventually through sentiment analysis or sentence polarity. However, these tools are language-dependent.

3.1 The Nature of the Features

With the aim of implementing a supervised and language-independent text classification approach for challenging contexts, the collected attributes will be statistical in nature. Table 2 presents a comprehensive list of potentially discriminant attributes used to address the requirements of several contexts, which include identifying heteronyms, determining the authors from the same or different epochs, and identifying the gender of the authors.

The meaning of most attributes in Table 2 is implicit in its name or in the *Description* column. By *Relative frequency* (in the beginning of the name of some attributes) we mean the absolute frequency of the feature in the document, divided by its size (number of words). By *text block* we mean the text between two consecutive *newline* characters. Attribute *Different words* corresponds to the number of distinct words divided by the document size. Some attributes, such as *3-grams* and *4-grams*, showed to be helpful on capturing the text *fingerprint* of authors who repeat groups of words in their poems. Features such as *Average block length* and *Text between commas* help on discriminating different writing styles. Other attributes, e.g. *Text between question marks* and *Text between exclamation points* may help to distinguish dialog/non-dialog texts. *Different words* feature is endowed with the vocabulary richness of the texts. The relative frequency of characters Q, K and & in each document, show to have some discriminant power. Concerning the *Syllabic variance* attribute, it is computed by

$$SV(D) = \frac{1}{\|S(D)\|} \sum_{s \in S(D)} (Syl(s) - AvSyl(D))^2$$

where $s = (w_1 \dots w_n)$ belongs to the set of sentences of document D and $AvSyl(D) = \frac{1}{\|S(D)\|} \sum_{s \in S(D)} Syl(s)$. $Syl(s)$ is the total number of syllables in words $(w_1 \dots w_n)$ of sentence s, which can be calculated from the number of vowels in those words minus the number of cases where two contiguous vowels form a diphthong — notice that there is no diphthong if one of the two contiguous vowels has an acute or grave accent —. This simple rule works for the vast majority of cases. However, although there are almost perfect alternative methods for calculating the number of syllables, they rely on language-specific morphosyntactic information, which we prefer to avoid. Thus, by measuring how variable the number of syllables in the sentences of a document is, *Syllabic variance* attribute tends to separate the group of authors of poems where the fixed number of metric syllables predominates, from the other authors.

Thus, we can see that the set of features in Table 2 is not intended to identify any specific author, but groups/classes of authors.

3.2 Measuring the *Discriminating Ability* of the Features

After gathering the features/attributes, the question arises as to how discriminating each potential feature A is in the context of each dataset. To help answer this question, a metric based on the ANOVA (Fisher, 1925), here called $D(A)$, was then used.

$D(A)$ computes the ability of feature A to discriminate classes via the quotient of the variance of the mean relative frequency of the attribute per document within the same class, to the mean variance of the relative frequency of the attribute per document within the same class, as shown in (1).

$$D(A) = \frac{\frac{1}{\|G\|} \sum_{g \in G} (M(A,g) - M(A,.))^2}{\frac{1}{\|G\|} \sum_{g \in G} \frac{1}{\|g\|} \sum_{d \in g} (f_r(A,d) - M(A,g))^2} \tag{1}$$

where G is the set of dataset classes and $M(A,g)$ represents the mean relative frequency of A in documents of class g, which is given by $M(A,g) = \frac{1}{\|g\|} \sum_{d \in g} f_r(A,d)$, being $f_r(A,d)$ the relative frequency of A in document d of class g. $M(A,.)$ is the mean value of $M(A,g)$ per class. It is calculated by $M(A,.) = \frac{1}{\|G\|} \sum_{g \in G} M(A,g)$.

Thus, the higher the $D(A)$ value, the greater the discriminating power of A measured by the *Discriminating Ability*. It is important to note that $D(A)$ may vary for the same attribute A, depending on the dataset. Table 3 shows examples of $D(A)$ values, for a subset of the attributes in Table 2, reflecting the ability of each one to discriminate among two genders/classes. To this end, 30 documents of each author were collected.

3.3 The Training and Classification Phases

After the attributes collection phase and the respective assessment of their *Discriminating Ability*, the training and classification phases begin. For this, the following classifiers were considered: Support Vector Machines (SVM) (Vapnik, 1999); Gaussian Naive Bayes (Bouman and van der Wurff, 1986); Decision Tree (Breiman et al., 1984); Bagging Classifier (Breiman, 1996); Random Forest (Breiman, 2001); Ada Boost (Freund and Schapire, 1997); k-NN (Cover and Hart, 1967). Concerning the training phase, the input data delivered to the classifiers is a matrix C where each line corresponds to a document d_i and each column to one of the attributes A_j. In our approach, each cell of C, $x(A_j, d_i)$, reflects the relative frequency of A_j in d_i, weighted by $D(A_j)$, given by (1), that is, $x(A_j, d_i) = f_r(A_j, d_i) \times D(A_j)$. Matrix C contains only the columns corresponding to attributes

Feature	$D(A)$
Exclamation	39.74459
Text between commas	5.828643
Uni-grams	4.898106
Different words	3.938020
Non ascii	3.781476
K	3.533615
Text between exclamation points	3.292565
2-grams	1.979188
3-grams	1.612895
3-char	1.310317
4-grams	1.123285
Q	1.054863
5-char	1.026168
9-char	0.923611
Average word length	0.787612
Text between points	0.777084

Table 3: Sorted table by descending values of $D(A)$, reflecting the *Discriminating Ability* of a subset of features from Table 2, for a dataset formed by 30 documents of each gender.

A_j where $D(A_j) >$ *Threshold*, as weakly discriminating attributes are usually useless. *Threshold* values are tuned according to each dataset.

3.4 Document Rejection Phase

In general, approaches using well-known classifiers (SVM, Naïve Bayes, K-NN, among others) assign one of the learned classes to the element being classified, usually the one with the most similar characteristics. However, sometimes the element being classified is dissimilar to all classes. For instance, if a classifier is trained to recognize documents written in English, French, and Portuguese, classifying a document written in Spanish would likely be assigned to Portuguese due to higher relative proximity. Although there's a weak resemblance to one of the classes in this case, in reality, this document should be rejected as it does not belong to any of the trained languages. Classic classifiers lack such an automatic rejection capability. In real-world scenarios, this behavior is often undesirable. Thus, we propose to equip the classification process with the ability to reject strange documents.

3.5 A New Criterion for Classification

To address the issue presented in the previous subchapter, we can utilize the theory that, if the distribution associated with data in each *cluster* is Gaussian/multivariate Gaussian, it is valid to perform a $\mathcal{X}^2$ test. This test relates the hypothesis of an element belonging to a class represented by a *cluster* with the squared Mahalanobis distance of the element to the centroid of that *cluster*. The core idea is to establish a sufficiently high probability to accept that the element should still belong to the *cluster*. There is a Mahalanobis distance threshold associated with this probability. For distances greater than this threshold, we reject the hypothesis that the element belongs to the class represented by the *cluster*. Therefore, it is possible to use a $\mathcal{X}^2$ test to reject the authorship of a document or assign it to one of the learned classes (authors) in the learning phase, according to the following hypothesis:

H_0 *: Let* p *be a document to be classified, represented by the vector* $\vec{p}$ *that belongs to class* k_i *portrayed by a cluster whose mean values of each attribute in the class and its features covariance matrix are respectively centroid* $\vec{\mu_i}$ *and* $\vec{\Sigma_i^{-1}}$*. Thus, applying a test with a confidence level of* α*, we can assert that* H_0 *will not be rejected if and only if:*

$$M^2(\vec{p}, \vec{\mu_i}, \vec{\Sigma_i^{-1}}) \leq \mathcal{X}^2_{df}(\alpha) \ . \tag{2}$$

$M^2(\vec{p}, \vec{\mu}, \vec{\Sigma^{-1}}) = (\vec{p} - \vec{\mu_i})^T \vec{\Sigma_i^{-1}} (\vec{p} - \vec{\mu_i})$ is the squared Mahalanobis distance and df, the degrees of freedom, is given by the number of features under study. Thus, by using a cumulative $\mathcal{X}^2$ table and the squared Mahalanobis distance from vector $\vec{p}$ to centroid $\vec{\mu_i}$, we can decide whether the document is close enough to assign authorship to one of the learned classes (authors), or if it is dissimilar enough to allow us to reject the authorship. Thus, we propose the following classification criterion:

$$\begin{gathered}\text{If } \exists k_i : M^2(\vec{p}, \vec{\mu_i}, \vec{\Sigma_i^{-1}}) = \min_{j \in \mathcal{K}} M^2(\vec{p}, \vec{\mu_j}, \Sigma_j^{-1}) \\ \wedge\, In(\vec{p}, \vec{\mu_i}, \vec{\Sigma_i^{-1}}, \alpha) \text{ then } p \in k_i \text{ class ,} \\ \text{otherwise } p \text{ belongs to an unknown class.}\end{gathered} \tag{3}$$

Predicate $In(\vec{p}, \vec{\mu_i}, \vec{\Sigma_i^{-1}}, \alpha)$ is true if and only if the condition represented in (2) is satisfied. $\mathcal{K}$ is the set of clusters. The inverse of the covariance matrix $\vec{\Sigma_i^{-1}}$ is associated with the features that characterize documents of a given class, typically the author or gender. $\vec{\Sigma_i}$ is estimated by the covariance matrix $\vec{E_i}$, based on the sample taken from the documents (the training documents) of cluster i, as follows:

Where $\|F\|$ is the number of features, and a generic element of $\vec{E_i}$ is described as follows:

$$\vec{E_i} = \begin{bmatrix} E_{1,1} & E_{1,2} & \cdots & E_{1,\|F\|} \\ E_{1,2} & E_{2,2} & \cdots & E_{2,\|F\|} \\ \vdots & \vdots & \ddots & \cdots \\ E_{1,\|F\|} & E_{2,\|F\|} & \cdots & E_{\|F\|,\|F\|} \end{bmatrix}$$

$$E_{l,t} = \frac{\sum\limits_{d \in g_i} \left(x(l,d) - x(l,.)\right)\left(x(t,d) - x(t,.)\right)}{\|g_i\|}$$

Here, g_i corresponds to the group of documents of class i, and $x(l,.)$ is the average value of component/feature l for the documents belonging to class i, that is $x(l,.) = \frac{1}{\|g_i\|} \sum_{d \in g_i} x(l,d)$.

3.6 Data Transformation to Normal

In order to use the $\mathcal{X}^2$ test in (2), data must be as close as possible to multivariate Gaussian. Thus, we leverage the Yeo-Johnson power to achieve a more Gaussian-like distribution while accommodating both positive and negative values in data. These transformations allow us to normalize skewed data in a manner that enhances the performance of subsequent classification models. The transformation is defined as follows:

$$\mathrm{Y}(\lambda) = \begin{cases} \left((1+x)^{\lambda} - 1\right)/\lambda & \text{if } x \geq 0, \lambda \neq 0 \\ \ln(1 - \lambda \cdot (-x))/(-\lambda) & \text{if } x < 0, \lambda \neq 0 \\ x & \text{if } \lambda = 0 \end{cases}$$

Where x is the original data point, λ is the parameter that optimizes the normality of the data distribution, and $Y(\lambda)$ represents the transformed value. By determining the optimal λ for each feature, see (Yeo and Johnson, 2000) for details, we can mitigate the effects of skewed distributions.

4 Results

4.1 The Datasets

In order to test our approach, several datasets were gathered, each one corresponding to a different class of problems. Each dataset uses documents from books of specific authors. In other words, each book is divided into documents. This way, documents can be used as samples for training or classification purposes. Table 4 shows the complete set of books used in the different datasets. However, this table does not include heteronym texts since they were not found available in books. Although heteronym documents were also included in our experiments and tests.

Book name	**Author**	**Year**
A Brusca	Agustina Bessa Luís	1967
Dentes de Rato	Agustina Bessa Luís	1987
Dicionário Imperfeito	Agustina Bessa Luís	2008
Sibila	Agustina Bessa Luís	1954
A relíquia	Eça de Queirós	1887
O Mistério da Estrada de Sintra	Eça de Queirós	1870
Os maias	Eça de Queirós	1888
S. Cristóvão	Eça de Queirós	(1890-1900)
História do Descobrimento e Conquista da Índia	Fernão Lopes de Castanhede	1554
Peregrinação	Fernão Mendes Pinto	1614
Textos de quatros Heterónimos	Fernando Pessoa	(1914-1934)
Desamparo	Inês Pedrosa	2015
Fazes-me falta	Inês Pedrosa	2002
Fica comigo esta noite	Inês Pedrosa	2003
Nas tuas mãos	Inês Pedrosa	1997
Catarina de Bragança	Isabel Stilwell	2008
D.Amélia	Isabel Stilwell	2010
D.Teresa	Isabel Stilwell	2015
Inclita Geração	Isabel Stilwell	2016
As intermitências da morte	José Saramago	2005
Caim	José Saramago	2009
Ensaio sobre a cegueira	José Saramago	1995
O homem duplicado	José Saramago	2002
As Naus	Lobo Antunes	2000
Auto dos danados	Lobo Antunes	1992
Explicação aos pássaros	Lobo Antunes	1981
O arquipélago da insónia	Lobo Antunes	2008
Sermão de São Pedro	Padre António Vieira	1644
Sermão de Santo António	Padre António Vieira	1654
Sermão de Todos os Santos	Padre António Vieira	1643

Table 4: Books used to form the different datasets (heteronym texts are not included).

4.1.1 Identification of Authorship of Contemporary Writers (19th and 20th Century).

This dataset, which we call *Contemporary*, aims to gather authors whose works were written within a time frame of less than about 100 years, specifically contemporary authors. Therefore, it is expected that morphological and syntactic patterns remain unchanged, overall. The set of selected authors and their works (from Table 4) are the following: Agustina Bessa Luís (ABL); Eça de Queirós (EQ); Inês Pedrosa (IP); Isabel Stilwell (IS); José Saramago (JS); Lobo Antunes (LA).

4.1.2 Author Gender Identification

Another dataset, called *Gender*, contains exactly the same authors as the previous one, but the classes are altered in order to form two groups (classes) corresponding to the authors' gender. For the study in question, only masculine and feminine genders are used. As in any other classification problem in challenging contexts, particularly in the present case where the attributes are purely statistical, it is necessary that there are actually differentiated writing patterns by gender, which is not guaranteed, therefore making the problem more difficult to solve. Thus, the classes are defined as follows:

$$\text{Classes} = \begin{cases} \text{EQ, JS, or LA} & \rightarrow \text{Masculine} \\ \text{ABL, IP, or IS} & \rightarrow \text{Feminine} \end{cases}$$

4.1.3 Identification of Authorship of Writers from Different Eras

Languages change over time, so documents from different eras will have distinct syntax and structure. Another dataset, *Different eras*, includes documents from authors of two main different eras. Training and classifying within this context is still challenging, since language-specific tools are not used in order to maintain language-independence, and authors from the same era are still to be distinguished. This dataset contains the works from Table 4 of the following authors: Fernão Mendes Pinto (FMP); Fernão Castanhede (FC); Padre António Vieira (PAV); Lobo Antunes (LA); Inês Pedrosa (IP); Isabel Stilwell (IS).

4.1.4 Identification of Authorship of Heteronyms of the same Writer

This task can be difficult, specially if there are several heteronyms, since the writer, being the same person, may repeat part of the style in every document. Despite that, there are differences in their writing patterns that can be detected through attributes such as *Syllabyc variance* and *Average block length*, as can be seen in Sect. 4.2, Table 5. A dataset called *Heteronyms* was built from a repository[1] and used for this study, including documents in Portuguese and English.

4.2 Evaluation (without Rejection Ability)

The aforementioned approach was then tested on the datasets referred in Sect. 4.1. For every dataset, except for the *Heteronyms*, 50 document samples were used for each class. Then, leave-one-out criterion was used in order to mitigate the relatively small number of samples. For the *Heteronyms* case, 127, 504, 307 and 397 document samples were used for Ricardo Reis (RR), Alberto Caeiro (AC), Álvaro de Campos (AdC) and Bernardo Soares (BS) heteronyms (classes), respectively; leave-one-out was also used here.

Based on experiments, it was found that values of $D(A)$ (1) tend to differ for the same attribute, depending on the dataset in question. As a result, the *Threshold* utilized to choose the optimal features using $D(A)$ may also differ. The resulting Table 5 showcases the group of features that offer the highest classification Accuracy for each dataset. Additionally, Table 6 illustrates which classifier produced the best Accuracy for each dataset.

Dataset	Features
Heteronyms	Q; 9-char; 5-char; Syllabic variance; Average block length
Contemporary	2-char; 5-char, Exclamation
Gender	Text between commas; Exclamation
Different eras	&; Periods; Text between commas

Table 5: For each dataset, the set of features that yielded the highest classification Accuracy.

Dataset	Classifier	Accuracy
Heteronyms	Bagging classifier	0.94
Contemporary	Gaussian Naïve Bayes	0.99
Gender	Random Forest	0.99
Different eras	Random Forest	0.98

Table 6: Classification Accuracy and optimal classifier for each dataset, as determined by the features outlined in Table 5.

From the confusion matrix in Table 7, we can read that Precision and Recall is not 1 ($100\,\%$) for all cases: for classes ABL and IS, Precision is $50/(50+1) \approx 0.98$ for both; for classes IP and LA, Recall is $49/(49+1) = 0.98$ for both. This is reflected by a global Accuracy of $1 - 2/(50 \times 4 + 2 \times 49 + 2) \approx 0.99$ for the *Contemporary* classes. Table 8 shows a high global Accuracy $(1 - 1/100 = 0.99)$ for the identification of classes of *Gender*. Also, the six classes (authors) from the *Different eras* dataset were classified achieving a global Accuracy of $1 - 5/300 \approx 0.98$, see Table 9.

Global Accuracy for the *Heteronyms* dataset reached $1 - (2 + 13 + 24 + 7 + 17 + 11 + 11 + 1)/1335 \approx 0.94$. This result confirms this as a highly challenging dataset by containing four heteronyms.

Actual/Predicted	**ABL**	**EQ**	**IP**	**IS**	**LA**	**JS**
ABL	50	0	0	0	0	0
EQ	0	50	0	0	0	0
IP	1	0	49	0	0	0
IS	0	0	0	50	0	0
LA	0	0	0	1	49	0
JS	0	0	0	0	0	50

Table 7: Confusion Matrix for the *Contemporary* dataset.

[1] https://www.kaggle.com/datasets/luisroque/the-complete-literary-works-of-fernando-pessoa?resource=download

Actual/Predicted	Masculine	Feminine
Masculine	49	1
Feminine	0	50

Table 8: Confusion Matrix for the *Gender* dataset.

Actual/Predicted	IS	IP	FC	FMP	LA	PAV
IS	48	2	0	0	0	0
IP	0	48	1	0	0	1
FC	0	0	49	0	0	1
FMP	0	0	0	50	0	0
LA	0	0	0	0	50	0
PAV	0	0	0	0	0	50

Table 9: Confusion Matrix for the *Different eras* dataset.

4.2.1 Comparative Analysis of Authorship Attribution Methods: Traditional vs. Deep Learning Approaches

Here we present a comparative analysis of the results obtained by our authorship attribution method in contrast to those achieved using two prominent pre-trained language models, BERT and RoBERTa, considering the highest challenging *Heteronyms* dataset.

Training Procedure: We fine-tuned these models with the AdamW optimizer, employing two different learning rate: 10^{-5} and 3×10^{-5} . The loss function adopted for training was cross-entropy, aligning with the classification nature of our task.

Validation and Early Stopping: To monitor model performance and avoid overfitting, we consistently evaluated the models on the validation set. Early stopping, with a *tolerance* = 3 based on validation loss, was employed to halt training.

Performance Metrics: Throughout the training process, we systematically assessed the models' performance. Key metrics, particularly Accuracy, were tracked for both training and validation sets, providing valuable insights into model progress.

Testing and Evaluation: Following model training, a rigorous evaluation was conducted using a separate test dataset. To this end, 30% of each author's documents were used for testing, 55% for training and 15% for validation.

Actual/Predicted	RR	AC	AdC	BS
RR	126	0	0	1
AC	0	476	17	11
AdC	0	24	272	11
BS	2	13	7	375

Table 10: Confusion Matrix for the *Heteronyms* dataset.

Model	Max Length	Lr	Accuracy
bert-base-cased	64	1e-5	0.75
bert-base-cased	128	1e-5	0.82
bert-base-cased	256	1e-5	0.86
roberta-base	64	1e-5	0.80
roberta-base	128	1e-5	0.87
roberta-base	256	1e-5	0.85
bert-base-cased	64	3e-5	0.38
bert-base-cased	128	3e-5	0.58
bert-base-cased	256	3e-5	0.38
roberta-base	64	3e-5	0.79
roberta-base	128	3e-5	0.85
roberta-base	256	3e-5	0.86

Table 11: Model comparison of the Accuracy obtained per Model and parameters using the *Heteronyms* dataset, where Max Length means the maximum number of tokens that can be processed in a single input sequence, and Lr is the learning rate.

As shown in the Table 11, the results obtained using transformers are generally inferior in terms of Accuracy when compared to our approach in this paper. This is likely because Deep Learning methods often perform better with large datasets, leading to higher Accuracy.

4.3 Evaluation (with Rejection Ability)

In order to provide the classification process with the ability to reject unkown documents, using the method mentioned above in Section 3.5, a different training phase has to be done. It consists on building the several pairs of $\vec{\Sigma}_i^{-1}$ matrix and centroid $\vec{\mu_i}$ (one pair per class), to be used later in the classification phase involving the Mahalanobis distance, see criterion defined in (3). These training and classification phases also followed the same methodology where documents are characterized by the frequency they have for each feature weighted by its *Discriminating Ability*, as described in Section 3.3.

To evaluate this new classifier (defined in criterion in (3)), two different tests were made: *a*) using documents belonging to classes known by the training phase; *b*) using only unknown ones. Concerning test *a*), leave-one-out approach were used with documents from all classes. Tables 14, 15 and 16 show the confusion matrices for the datasets indicated. Table 12 contains the corresponding global Accuracy values for test *a*). Thus, we can see from Table 15, as an example, that this criterion missclassify 14 of 300 documents from dataset

Dataset	Test	Accuracy
Different eras	*a)*	0.94
	b)	0.94
Contemporary	*a)*	0.95
	b)	0.99
Heteronyms	*a)*	0.84
	b)	0.90

Table 12: Classification results for the classifier proposed and defined in criterion (3): evaluation for tests *a)* and *b)* defined in Subsec. 4.3.

Dataset	Features
Different eras	'Different words', 'Text between periods' 'Text between commas'
Contemporary	'Different words', 'Average word length', '5-char'
Heteronyms	'Different Words', 'Capital letters', 'Non ascii', 'Average block length', 'Syllabic variance', 'Average word length'

Table 13: Features used in different datasets for the classifier proposed and defined in criterion (3).

Contemporary , therefore 1 - 14/300 ≈ 0.95. From these 14, 13 were wrongly rejected as unknown (Unk).

For test *b*), leave-one-out were also used but each training iteration did not include the documents of the class of the document to be classified. This way, it was possible to assess the ability of the classifier to reject unknown documents. Thus, for *Different eras* dataset, 17 in 300 documents were wrongly classified as belonging to one of the known authors, instead of being rejected, which corresponds to 94% Accuracy. For *Contemporary* and *Heteronyms* datasets, the wrong cases were 2 in 300, and 135 in 1335, which corresponds to 99% and 90% respectively, as shown in Table 12.

Table 13 shows the features used for each dataset in the context of the classifier we propose.

Table 12 also shows that the Accuracy of test *a)*, for example for *Contemporary* dataset (0.94), is lower than the the Accuracy obtained with Guassian Naïve Bayes classifier (0.99), see Table 6. This may be the price to pay for the need to *gaussianize* data in order to use criterion defined in (3), which includes the rejection ability, as explained in Subsec. 3.6. In fact, it is a transformation that tends to smooth the relative distances between documents' representation in the vectorial space, which may slightly *smooth* the distances between clusters.

Actual/Predicted	FC	FMP	IP	IS	LA	PAV	Unk
FC	45	1	0	0	0	0	4
FMP	0	46	0	0	2	0	2
IP	0	0	47	0	0	0	3
IS	0	0	0	48	0	0	2
LA	0	0	0	0	49	0	1
PAV	0	0	0	0	0	47	3
Unk	0	0	0	0	0	0	0

Table 14: Confusion Matrix for test *a) - Different eras*.

Actual/Predicted	ABL	JS	IP	IS	LA	EQ	Unk
ABL	47	0	0	0	0	0	3
JS	1	47	0	0	0	0	2
IP	0	0	45	0	0	0	5
IS	0	0	0	49	0	0	1
LA	0	0	0	0	49	0	1
EQ	0	0	0	0	0	49	1
Unk	0	0	0	0	0	0	0

Table 15: Confusion Matrix for test *a)* - *Contemporary* dataset.

Actual/Predicted	RR	AdC	AC	BS	Unk
RR	214	23	1	69	0
AdC	11	351	2	33	0
AC	9	16	100	2	0
BS	24	23	0	457	0
Unk	0	0	0	0	0

Table 16: Confusion Matrix - *Heteronyms*

5 Conclusion

This paper presents a supervised document classification approach for authorship identification in challenging contexts, with the capability to reject documents from unknown classes. The approach is faced with the challenge of finding features that are not influenced by the morphosyntactic structure of any particular language and achieving promising classification results. To address this, we exclusively used statistical features to increase the approach's applicability across the widest possible range of languages. The features were evaluated based on their discriminating ability within the context of each dataset, and only the most effective ones were employed.

While these features empower the attainment of very high Accuracy in classification when employed alongside conventional classifiers like Gaussian Naïve Bayes or Random Forest, they lack a well-established approach to incorporate rejection capabilities. Recognizing this shortfall, we introduced a novel classification criterion based on Mahalanobis distance and a χ^2 test, ensuring a founded technique for document rejection, while maintaining high Accuracy.

In the most complex task of identifying four heteronyms written by the same writer, the approach highlighted the need for further improvement in future work.

Acknowledgements

This work is supported by NOVA LINCS (UIDB/04516/2020) with the financial support of FCT.IP.

References

Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):1–29.

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text*, 23(3):321–346.

C.A. Bouman and C.L. van der Wurff. 1986. The optimal classification rule for gaussian distributions. *Pattern Recognition*, 19(3):237–241.

L. Breiman. 1996. Bagging predictors. *Machine Learning*, 24(2):123–140.

L. Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. 1984. *Classification and regression trees*. CRC press.

T. Cover and P. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

Sara Elmanarelbouanani and Ismail Kassou. 2013. Authorship analysis studies: A survey. *International Journal of Computer Applications*, 86.

Ronald A. Fisher. 1925. Statistical methods for research workers. *Genesis*, 1:1–10.

Y. Freund and R.E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Michael Gamon. 2004. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 611–617, Geneva, Switzerland. COLING.

Neha Garg, Sumit Singla, Amandeep Kaur, Mayank Saini, Tarun Khanna, and Sumeet Kumar. 2018. Author gender classification: A comparison of different feature sets and classifiers. In *2018 International Conference on Intelligent Computing and Control Systems (ICCS)*, pages 293–298. IEEE.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*.

Helena Gomez Adorno, Grigori Sidorov, David Pinto, and Ilia Markov. 2015. A graph based authorship identification approach.

Alexey Gritsenko and Evgueni N Smirnov. 2008. Rejection strategies in support vector machines: A comparative study. *Pattern Recognition Letters*, 29(12):1737–1744.

Fatma Howedi. 2014. Text classification for authorship attribution using naive bayes classifier with limited training data.

Farkhund Iqbal, Hamad Binsalleeh, Benjamin CM Fung, and Mourad Debbabi. 2010. Mining writeprints from anonymous e-mails for forensic investigation. *digital investigation*, 7(1-2):56–64.

Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING*, volume 3, pages 255–264.

Özge Fırat Keskin and Sarp Adali. 2019. Turkish authorship attribution based on linguistic features. *PloS one*, 14(11):e0224682.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2003. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 54(4):344–357.

Siham Ouamour and Halim Sayoud. 2012. Authorship attribution of ancient texts written by ten arabic travelers using a smo-svm classifier. In *2012 International Conference on Communications and Information Technology (ICCIT)*, pages 44–47. IEEE.

Naga Potha and Pasupuleti Rao. 2018. Efficient machine learning algorithms for authorship attribution. *International Journal of Computer Applications*, 180(39):37–43.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of portuguese with transformer albertina pt. *arXiv preprint arXiv:2305.06721*.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Joao F Teixeira and Marco Couto. 2015. Automatic distinction of fernando pessoas' heteronyms. In *Portuguese Conference on Artificial Intelligence*, pages 783–788. Springer.

Vladimir N Vapnik. 1999. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999.

Robert A Yeo and Robert J Johnson. 2000. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959.

Zhiqing Zhao and Liang Li. 2018. Gender classification of chinese microblog authors using ensemble learning. In *2018 International Conference on Asian Language Processing (IALP)*, pages 14–17. IEEE.

George Kingsley Zipf. 1932. Selected studies of the principle of relative frequency in language.