Hunter Distance in Authorship Verification

Octavian-Mihai Radu
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
ORCID: 0009-0002-1634-5065

Bogdan Ionescu
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
bogdan.ionescu@upb.ro

Ruxandra Tapu
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
ruxandra.tapu@upb.ro

***Abstract*—The inherent idiosyncrasies of individual writing styles pose significant challenges to developing scalable authorship verification (AV) systems. These challenges are even more difficult to address for less-resourced languages, such as Romanian, where there is a scarcity of annotated datasets and advanced language tools. This article introduces a novel way to evaluate the styling vector distance, that can successfully replace the standard Euclidian distance used in most classification cases for authorship verification. In a low-dimensional world, the Euclidean distance, which is the straight-line path between two points, is a reliable and intuitive measure of distance. However, in the case of high-dimensional vectors, this metric can pose challenges, and the problem lies in the counter-intuitive geometry of high-dimensional spaces. The Hunter distance presented in this article is trying to solve high dimensionality case issue.**

***Index Terms*—Euclidian distance, Hunter Theorem, CHS polynomials**

I. Introduction and Related Work

Authorship verification (AV) is about figuring out if the same person wrote two different pieces of text. Traditionally, linguists did this by looking for an author's unique writing habits in anonymous texts to identify who wrote them [1]. Today, much of the algorithms and research on identifying authors relies on a core idea called the "Stylome Hypothesis." This idea, first clearly described by van Halteren et al. [2], suggests that everyone has a unique "fingerprint" in their writing style. This fingerprint is made up of consistent features that can be found if you have enough of their writing. While the Stylome Hypothesis is a useful idea to work with, it's difficult to show it's true, and even harder to prove it definitively. The aim of authorship verification is to estimate the function:

$$\eta : \mathcal{D}_{known}\ x\ D_{test} \rightarrow 0, 1$$

where $\mathcal{D}_{known}$ is a set of documents of known authorships and D_{test} is a document of unknown or questioned authorship. If $D_1 \in \mathcal{D}_{known}$ and $\eta\,(D_1,\ D_{test}) = 1$, then the author of D_1 is also the author of D_{test}. Similar, if $\eta\,(D_1,\ D_{test}) = 0$, then the author of D_1 is not the same as the author of D_{test}. Authorship verification (AV) is a notably active domain within computational linguistics, marked by a substantial volume of research and published findings in recent years. The methodologies employed can be broadly categorized into four main paradigms: classic machine learning models, deep learning architectures, Transformer-based models, and, more recently, architectures leveraging large language models.

2nd Workshop on Artificial Intelligence for Multimedia, AI Multimedia Lab, CAMPUS Research Institute, POLITEHNICA Bucharest, Bucharest, Romania, 15 July, 2025

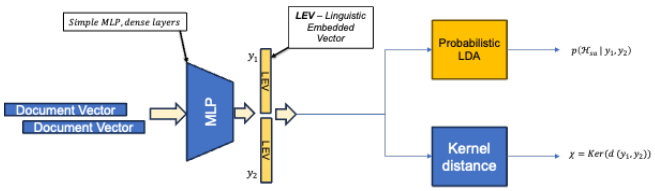

Fig. 1: Decision block of an Authorship Verification System.

In Figure 1 is presented the high level architecture of Letera system for the authorship verification. In this article we will focus only on the kernel distance, the module that measure the distance in high dimensional space between the two document (linguistic) vectors. Let $d\,(y_1, y_2)$ denote the distance between the linguistic embedded vectors y_1 and y_2, which correspond to the input texts. We will evaluate two distance metrics: the conventional Euclidean distance and a novel approach we term 'Hunter distance.' To transform these distance measures into a normalized similarity function bounded between 0 and 1, where a value approaching 1 signifies a high degree of similarity, we employ a kernel function defined as:

$$Ker(d) = -e^{\alpha d^{\beta}}$$

The behavior of this kernel is governed by two learnable parameters: a scaling parameter, α, and a shape parameter, β. It is worthy to note that setting $\beta = 2$ results in a Gaussian kernel, while $\beta = 1$ yields a Laplacian kernel. Both α and β are optimized during the training phase. For our experiments, we initialize these parameters with values of α $\alpha = 0.095$ and $\beta = 1.5$.

The most common approach is to employ the Euclidean distance, defined as the $L2$ norm of the difference between the vectors:

$$d(y_1, y_2) = ||y_1 - y_2||_2$$

When this specific distance metric is incorporated into the kernel function, the expression for the similarity score becomes:

$$Ker(d) = -e^{\alpha ||y_1 - y_2||_2^{\beta}}$$

where the kernel directly maps the Euclidean distance between the vectors to a normalized similarity value. This type of distance is also used in by Boenninghoff et al. in [3], being a very common approach. In a low-dimensional world, the

Euclidean distance, which is the straight-line path between two points, is a reliable and intuitive measure of distance. However, in the case of high-dimensional vectors, this metric can pose challenges, and the problem lies in the counter-intuitive geometry of high-dimensional spaces. As the linguistic vectors in our network are high dimensions ($D_{LEV} \in [60; 150]$), we introduced a new metric called the Hunter distance. In this case, considering the $H(\mathbf{x})$ the distance between the linguistic embedded vectors y_1 and y_2 using Hunter distance, the kernel distance becomes:

$$Ker(d) = -e^{\alpha H(y_1 - y_2)^\beta}$$

In the next section we will define mathematically the Hunter distance.

II. Hunter Distance

To define the Hunter distance we need first to understand what are CHS polinomials and what is the Hunter's theorem.

A. Complete Homogeneous Symmetric (CHS) Polynomials

The complete homogeneous symmetric (CHS) polynomial of degree k in n variables is the sum of all distinct degree-k monomials in the given variables. Formally, CHS is defined as:

$$f_k(x_1, x_2, \dots x_n) = \sum_{1 \le i_1 \le \dots \le i_k \le n} x_{i_1} x_{i_2} \dots x_{i_k} \quad (1)$$

Here are few examples for n=2 and k=1,2,4:

$$\begin{cases} f_1(x_1, x_2) = x_1 + x_2 \\ f_2(x_1, x_2) = x_1^2 + x_1 x_2 \ + \ x_2^2 \\ f_4(x_1, x_2) = x_1^4 + x_1^3 x_2 + x_1^2 x_2^2 + \ x_1 x_2^3 + \ x_2^4 \end{cases}$$

B. Hunter's Theorem

We will start from a remarkable theorem of D.B. Hunter, proved in 1977 in [4], that asserts that the CHS polynomials of even degree are positive definite functions.

Hunter's Positivity Theorem - Let $k \ge 0$ be an even integer. Then the CHS polynomial of degree k, $f_k(x_1, x_2, \dots, x_n)$ is a positive definite function in $\mathbb{R}^n$, that is $f_k(x_1, x_2, \dots, x_n) > 0$ for all $\mathbf{x} \neq 0$.

Proof First we will by evaluating first the following integral (which is basically the Laplace transform):

$$I_{n,k} = \int_{[0,\infty)^n} (t_1 x_1 + \dots + t_n x_n)^k e^{-(t_1 + \dots t_n)} dt_1 \dots dt_n \quad (2)$$

By looking at the first two terms of the series that are simply to evaluate we notice that:

$$\begin{cases} I_{1,k} = k! f_k(x_1) \\ I_{2,k} = k! f_k(x_1, x_2) \end{cases}$$

where f_k is defined in equation (1). By induction it is simple to prove that the general form is:

$$I_{n,k} = k! f_k(x_1, x_2, \dots, x_n)$$

$$\implies f_k(x_1, x_2, \dots, x_n) = \frac{1}{k!} I_{n,k}$$

By looking at the definition of $I_{n,k}$, now we have the proof that for k even we have that:

$$f_k(x_1, x_2, \dots, x_n) > 0$$

Even more, Hunter established a lower bound for the even degree CHS polynomials to prove positive definiteness cite1.

Hunter's Norm - If $f_k(x_1, x_2, \dots, x_n)$ the CHS polynomial of degree k with k an even integer, then $\sqrt[k]{f_k(x_1, x_2, \dots, x_n)}$ is a norm on $\mathbb{R}^n$.

Proof By definition, a norm $||\cdot|| : R^n \to R$ must have the following properties:

$$\begin{cases} ||x|| \ge 0, \ for\ all\ x \in \mathbb{R}^n \\ ||\lambda x|| = ||\lambda|| \cdot ||x||, \ for\ all\ \lambda \in \mathbb{R}, x \in \mathbb{R}^n \\ ||x + y|| \le ||x|| + ||y||, \ for\ all\ x, y \in \mathbb{R}^n \end{cases}$$

For those familiar with topology, any norm $||\cdot||$ on a vector space gives a metric distance on the vector space defined by the formula $d(x, y) = ||x - y||$. For Hunter's norm we will call it the Hunter distance.

Property 1) of the norm is already demonstrated as part of Hunter's positivity theorem.

Property 2) is also very simple to prove:

$$||\lambda x|| = \sqrt[k]{\sum_{cyclic} x_1^{\alpha_1} \lambda_1^{\alpha_1} \dots x_n^{\alpha_n} \lambda_n^{\alpha_n}}$$

$$\implies ||\lambda x|| = |\lambda| \sqrt[k]{\sum_{cyclic} x_1^{\alpha_1} \dots x_n^{\alpha_n}} = ||\lambda|| \cdot ||x||$$

Regarding the property 3), it becomes:

$$\left(\sum_{cyclic} (x_1 + y_1)^{\alpha_1} (x_2 + y_2)^{\alpha_2} \dots (x_n + y_n)^{\alpha_n} \right)^{\frac{1}{k}} \le$$

$$\le \left(\sum_{cyclic} x_1^{\alpha_1} x_2^{\alpha_2} \dots x_n^{\alpha_n} \right)^{\frac{1}{k}} + \left(\sum_{cyclic} y_1^{\alpha_1} y_2^{\alpha_2} \dots y_n^{\alpha_n} \right)^{\frac{1}{k}}$$

which is ensured by the Minkowski's inequality. Having all the three properties of a norm, we proved that:

$$H(\mathbf{x}) = \sqrt[k]{f_k(x_1, x_2, \dots x_n)} \quad (3)$$

is a convex function and represents a norm on $\mathbb{R}^n$.The norm defined in (3) we will call the Hunter norm. A more general approach on norms on complex matrices induced by complete homogeneous symmetric polynomials can be found in [5].

III. RESULTS

To evaluate the efficacy of our novel Hunter's distance we are using our own system for authorship verification - Letera. A series of experiments was conducted on the Romanian Story (RoSto) corpus, a benchmark dataset comprising texts from 19 Romanian authors. All experiments were performed on both full-text (FT) and paragraph-level (PP) subsets of the corpus to assess model robustness to variations in text length. For all cases we also used the Euclidian distance as base for the comparison.

TABLE I: RoSto Corpus, 19 Authors, Full Texts

Model	*F1 Score*	*AUC*	*F0.5*	*Brier*
Kernel Euclidian Distance	0.881	**0.963**	**0.912**	**0.919**
Kernel Hunter Distance	**0.883**	0.962	0.910	0.918

In the first evaluation dataset we have 19 authors with a total of 1263 of texts. As it can be seen in Table 1, few metrics were used to valuate the performance of Euclidian distance versus Hunter distance: F1 score, Area Under Curve (AUC), F0.5 and Brier score.

TABLE II: RoSto Corpus, 19 Authors, Paragraphs

Model	*F1 Score*	*AUC*	*F0.5*	*Brier*
Kernel Euclidian Distance	0.842	0.919	0.842	0.882
Kernel Hunter Distance	**0.851**	**0.920**	**0.843**	**0.885**

The second evaluation was done on shorter texts (paragraphs) from the same authors. In this case we have 19 authors with a total of 12516 texts of various lengths.

TABLE III: RoNews Corpus, 767 Authors

Model	*F1 Score*	*AUC*	*F0.5*	*Brier*
Kernel Euclidian Distance	0.831	**0.921**	**0.849**	**0.884**
Kernel Hunter Distance	0.831	0.919	0.844	0.882

The third evaluation was done on Romanian news texts we collected from the internet. In this dataset we have 767 authors with a total of 46703 texts (all the news that have more than 200 characters). It is important to contextualize the slight decrease in performance observed on the RoNews corpus. Datasets created through automated web-scraping are inherently susceptible to label noise, a common challenge in real-world authorship analysis. In the context of online journalism, this noise can manifest in several ways; for instance, a single journalist may publish articles under different pseudonyms, while conversely, multiple individuals may contribute articles under a single, generic attribution. Such ambiguities in the ground-truth author labels introduce a level of irreducible error for any classification model.

REFERENCES

[1] Ehrhardt, S. – "Authorship attribution analysis" - Handbook of Communication in the Legal Sphere. pp. 169–200. de Gruyter, Berlin/Boston (2018)

[2] H. van Halteren, H. Baayen, F. Tweedie, M. Haverkort, A. Neijt – "New machine learning methods demonstrate the existence of a human stylome" - Journal of Quantitative Linguistics, 2005

[3] B. Boenninghoff, R. M. Nickel, D. Kolossa – "O2D2: Out-of-distribution detector to capture undecidable trials in authorship verification" – Computing Research Repository (CoRR), 2021

[4] David B. Hunter – "The positive-definiteness of the complete symmetric functions of even order" - Math. Proc. Cambridge Philos. Soc. 82 (1977)

[5] K. Aguilar, A. Chávez, S.Garcia, and J. Volčič – "Norms on complex matrices induced by complete homogeneous symmetric polynomials" - Bull. Lond. Math. Soc. 54, 2022