Twenty-One* Pseudo-Chrysostoms and more: authorship verification in the patristic world

Thibault Clérice, Anthony Glaise

Twenty-One* Pseudo-Chrysostoms and more: authorship verification in the patristic world

Thibault Clérice[1,*,†], Anthony Glaise[‡]

[1]ALMAnaCH, Inria Paris, France

Abstract

As the most prolific of the Church Fathers, John Chrysostom (344–407 CE) has a vast textual mass and theological importance that has led to a significant misattribution of texts, resulting in the existence of a second corpus known as the pseudo-Chrysostomian corpus. Like many Greek-language Church Fathers' works, this corpus comprises anonymous texts, which scholars have attempted to reattribute or group together based on factors such as the person's function, biography, ideology, style, etc. One survey conducted by Voicu in 1981 explored potential groupings of such texts and produced a critical list of 21 Pseudo-Chrysostom works identified by scholars, including Montfaucon (1655–1741), one of the first modern editors of Chrysostom's writings. In this paper, we present a novel approach to addressing pseudonymous work in the context of chrysostomian studies. We propose to employ siamese networks within an authorship verification framework, following the methodology commonly used in recent computational linguistic competitions. Our embedding model is trained using commonly used features in the digital humanities landscape, such as the most frequent words, affixes, and POS trigrams, utilizing a signal-to-noise ratio distance and pair mining. The results of our model show high AUCROC scores (0.855). Furthermore, the article concludes with an analysis of the pseudo-Chrysostoms proposed by Voicu. We validate a significant portion of the hypotheses found in Voicu's survey while also providing counter-arguments for two Pseudo-Chrysostoms. This research contributes to shedding light on the attribution of ancient texts and enriches the field of chrysostomian studies.

Keywords

Patristic Studies, Ancient Greek, Stylometry, Siamese Networks, Authorship verification

1. Introduction

Late Antiquity literature in Latin and Ancient Greek bears a profound influence from Christian literature, shaping the era's literary landscape. Patristic studies focus on the examination of the earliest Christian authors, spanning from the 1st century CE to the 7th century CE, with some scholars extending the period up to Jean Damascene in the middle of the 8th century CE. Within the realm of Church Fathers, two prominent figures stand out for their enduring

CHR 2023: Computational Humanities Research Conference, December 6 – 8, 2023, Paris, France

*Corresponding author.

†Contribution on all domain.

‡Contribution on the pseudo-chryostomian context and evaluation of results in regards to the "traditional" patristic academic work.

§All authors participated in the writing of the paper.

thibault.clerice@inria.fr (T. Clérice)

https://ponteineptique.github.io/ (T. Clérice)

0000-0003-1852-9204 (T. Clérice); 0000-0003-4715-5184 (A. Glaise)

impact on the corpora: Augustine (354–430) in Latin and John Chrysostom (344–407) in Ancient Greek. Despite their significance, both Augustine and John Chrysostom encountered a common challenge in Late Antiquity. Numerous works were falsely attributed to them, possibly with the intention of elevating the popularity of these writings, whose actual authors were less renowned. This misattribution has led to the creation of an extensive body of texts with uncertain authorship. According to Voicu [1], there are "more than a thousand" pseudo-chrysostomian works whose true originators remain unknown. To address this issue, scholars of patristics have endeavored to categorize some of these spurious works, associating them with a single person, who may be anonymous or identified with a specific historical figure.

The task of distinguishing anonymous authors within the pseudo-Chrysostomian corpus is a significant and essential endeavor. By doing so, historians, patristicians, and philologists can work with a more manageable collection of texts. This process offers valuable insights into the construction of Christian theological frameworks, allowing the identification of ideological or thematic clusters within authorship groups. Establishing authorship for texts with at least one proposed date enables scholars to study the evolution of Christian ideology more accurately. Moreover, in rare cases, dating or attributing a specific anonymous text (especially when milestones are present) can provide new dates to other texts, as they might be cited or reused [2]. This effort holds immense potential for advancing our understanding of the historical and intellectual landscape of Late Antiquity, contributing to a deeper comprehension of the development of the Christian tradition.

In traditional patristics, authorship attribution largely relies on the identification of rare patterns [3], as well as extra-textual clues, such as historical events serving as *terminus post quem* or *ante quem* markers, and the analysis of ideological traits or the use of specific quotations. Authors' tendencies to rely on particular parts of the *Bible* or cite other authors also serve as *de facto terminus ante quem*. In 1981, Voicu [1] presented a comprehensive analysis of attributions using such methods in the context of the pseudo-chrysostomian corpus. He identified 21 text groups for which a common authorship has been proposed by himself or other scholars. However, for some of these pseudo-Chrysostoms, Voicu and the scholars he cites sometimes express hesitancy regarding the attribution or hold differing perspectives on the matter.

In such a context, automatic authorship verification can offer new insights that either support or challenge previous hypotheses. Two other fields that are actively engaged in authorship verification are computational linguistics and digital humanities (DH). In the realm of stylometrical analysis, DH scholars tend to treat texts as bag-of-words and utilize statistically significant features, such as the most frequent words of a corpus [4], part-of-speech 3-grams, and character n-grams [5]. On the other hand, computational linguistics (CL) has recently shown a preference for treating text as a sequence of words, employing masked language models [6]. In terms of approach, DH papers lean towards classification (utilizing SVM or similar models) or unsupervised clustering methods. In contrast, CL seems to be dominated by siamese neural networks, regardless of the type of input [7].

In our research, we propose a novel methodology that combines approaches from both computational linguistics and digital humanities to analyze the 21 potential Pseudo-Chrysostoms identified by Voicu and his predecessors. Specifically, we aim to evaluate the effectiveness of different features, including most frequent words, affixes, and POS 3-grams, within the context of a Siamese network using a linear projection in N dimensions. To maximize the potential of

our corpus, we introduce Easy-SemiHard Pair Mining[8] to our batch learning process and utilize a signal-to-noise ratio distance[9] to distinguish and separate our texts.

In summary, the contributions of our paper are as follows:

- Introducing a new approach to authorial verification for ancient texts, incorporating Siamese networks and easy-semihard pair mining into the landscape of stylometry within computational humanities.
- Introducing the signal-to-noise ratio distance as a novel distance and loss for authorship verification.
- Conducting an in-depth analysis of the results obtained from applying our approach to the 21* pseudo-Chrysostoms' texts identified by Voicu.

The remaining of the paper is organised as follows. Section 2 provides background on the article of Voicu and in general on the issue of the pseudo-Chrysostomian corpus, as well as a deeper background on authorship verification and stylometry in general. Section 3 (Proposed Methods) provides details about the architecture used for the experiments. Section 4 (Experimental Setup) provides insight on the corpus, the feature selection and the metrics. Section 5 provides an evaluation of the results on independent test sets. Section 6 reuses the models built and to provide insight on the PC corpus.

2. Background and Related Work

2.1. Background

Regarding the pseudo-Chrysostomian corpus, a non-specialist might notice the lack of recent discussions concerning various attribution hypotheses. Most of these "old" hypotheses trace their origins back to the 18th century, thanks to the diligent efforts of Bernard de Montfaucon, a Benedictine monk who dedicated his work to publishing the works of Athanasius (1698) and John Chrysostom (1718). Montfaucon's editions and commentary laid the groundwork for later comprehensive editions of numerous patristic texts [10], including the renowned *Patrologia Graeca* compiled by Migne, which is still widely used today as it is now in the public domain. Fortunately, in 1981, Voicu [1] provided the most recent comprehensive summary of the authorship hypotheses within the pseudo-Chrysostomian corpus. This summary encompassed both refutations and ongoing debates surrounding various authorship attributions. Since then, some new hypotheses have emerged, but there hasn't been a comparable effort to Voicu's in terms of summarizing the existing scholarship.

In his paper, Voicu provides a summary regarding authorship clustering of 88 texts, grouped under the pseudo-identity of 21 different authors by various scholars (PC1 to PC21). On top of the 21 base PC, another group of text is hypothetically attributed to one of them: PC20 is accompanied by PC20b, for which he drafts a possible ensemble without confirmation that they might be of the same authors (see Table 1). The arguments regarding the attributions can vary, and we summarise them in four different categories:

- **Theological arguments** are based on ideological (in)compatibilities between texts, in the context of a non-unified Christian religion in which subgroups can be found. *E.g.*

PC16 is categorized as a moderate Antiochian, which makes them incompatible with an Alexandrian ideology.

- **Sequential arguments** are based on the obvious continuity between two texts, with established narrative links in both senses (Text A builds on Text B and vice-versa).
- **Stylistic arguments** are based on the style of the authors. They range from the use of what is perceived as "bad Greek" (PC6 and 7) to citation habits regarding the scriptures.
- **Extra-textual arguments** are mostly based on events or texts referenced within the texts, which gives them a common *terminus ante quem* or *post quem*, or actually makes them incompatible. It also refers to transmission proofs, such as the constant grouping of the same texts in their transmission history (PC17).

Cluster	Number of texts	Original Hypothesis	Confidence	Additional analysis	Confidence	Argument type
1	2	Montfaucon		Altendorf	Refuted	theological
2	2	Montfaucon		Voicu	Confirmed	continuity
3	3	Montfaucon		Voicu	Refuted	
4	7	Montfaucon		Voicu	Low	
5	3	Montfaucon		Voicu	Partially refuted	
6	3	Montfaucon		Voicu	Low	
7	7	Montfaucon		Voicu	Possible	
8	2	Montfaucon		Voicu	Refuted	stylistic
9	5	Marx		Voicu	Refuted	
10	2	Weyer				
11	3	Nautin				
12	2*	Liébart				theologic
13	3	Leroy		Voicu	Refuted	
14	4	Voicu	High			
15	5	Voicu	Mostly high			
16	5	Voicu	High	Wenger	Partially Possible	
17	2	Voicu	High			
18	2	Voicu	Possible			stylistic,theologic
19	2*	Rilliet	High			
20	5	Datema	High	Voicu		
20b	12	Datema	Possible	Voicu		
21	2	Voicu	High			extra-textual

Table 1
List of pseudo-Chrysostoms and the scholars behind the hypotheses. Confidence concerning the grouping of texts under a single pseudonymous author is provided based on Voicu's commentary of each scholar's analysis, including his own. We present a simple typology of the arguments when provided. All the cited scholarship, except for Montfaucon, was published between 1940 (Marx [11]) and 1981 (Voicu).

Most of the cited texts are available in digital formats, specifically in the *Thesaurus Lingua Graeca*, which is unfortunately closed access[1]. However, one PC could not be tested in our framework: PC19 only refers to one text in Ancient Greek, the other text being a Syriac translation. PC12 has been produced using the OCR available on Google Books, which has been lightly post-corrected.

[1]All the feature extraction process is shared within the repositories, and the features themselves are made available.

2.2. Related work in stylometry and authorship verification

The present work is related to two fields: computational linguistics (CL) and digital humanities (DH). The existing literature reveals distinct approaches to the problem of authorship verification or authorship identification, primarily influenced by the attributes of each field's corpus and expectations for explainability. We focus on the latest approaches applied in both fields, emphasizing the common technical approaches they have shared in the past.

Computational Linguistics CL offers a clear landscape thanks to PAN, a "series of scientific events and shared texts on digital text forensics and stylometry." PAN's shared tasks have provided recurring competitions since as early as 2011, focusing on authorship attribution (2011, 2012, 2018, 2019) or verification (2013–2015, 2020–2023). As in most other computational linguistics tasks, deep learning has seen a rise in popularity, leading to improved scores but lower explainability. An insight into some of the approaches taken in authorship verification since 2015 reveals the following methods and features used:

- In 2015, Word-based n-grams, sentence length, word frequencies, punctuation frequencies, POS frequencies, and POS n-grams were shared features across different papers [12, 13]. These features were eventually fed into standard classifiers such as Random Forest or SVMs.
- In the 2018 PAN authorship attribution task [14], features continued to focus on characters and word n-grams, with various weighting and normalization methods. While SVMs were commonly used, some early neural network approaches also appeared.
- In the 2020 PAN authorship verification task [15], methods employing neural networks made an appearance, particularly in the form of Siamese neural networks. Features mainly remained classical stylometric features, such as normalized frequencies of tokens or POS. The winning paper utilized a Siamese network with extracted "linguistic embedding vectors" using an LSTM network with attention to produce document embeddings.
- In 2021 [7], features and approaches from 2019 were carried over, with a paper using BERT and another one using Siamese networks ranking first and second place, respectively, in the large dataset competition on overall scores, with the first one winning the competition on all provided metrics.
- In 2022 [6], all competitors moved away from SVM and instead used either masked language models' embeddings or previous existing features with Siamese network approaches or fully connected neural networks.

With this summary, we observe that CL is gradually moving away from explainability and increasingly employing text sequences (using RNNs, CNNs, or transformers like BERT or T5) rather than treating texts as bags of words (BoW). Scoring is conducted using various metrics, including AUROC and a modified F1-Score $F_{0.5u}$ [16], which "emphasizes correctly-answered same-author cases and rewards non-answers". Siamese networks gained prominence in authorship verification tasks in 2020 and 2021 but mostly disappeared in 2022.

In contrast to previous competitions, 2022 departed from 2021 in the type of dataset used. While 2021 and previous shared tasks utilized datasets sharing a common domain (such as fanfictions for 2021), 2022 focused on *cross-discourse authorship verification*, encompassing "emails,

essays, texts messages, and business memos." This shift resulted in a significant performance drop compared to previous years. Additionally, regarding the dataset, each competition relied on fixed pairings of texts, with limited or no information about the author in the metadata of the documents.

Digital Humanities Unlike CL, DH has largely preferred using feature selection and treats most texts in an authorship attribution framework rather than authorship verification. The three most frequently recognized features are: most frequent words [17, 18], function words, affixes (especially for languages with significant variations in spelling or flexions) [19], and POS, with the latter often used as n-grams [5]. Some other features, such as rhymes [5], meters [20], and even treebank syntactic tags for Ancient Greek [21], have demonstrated usefulness and stylometrical importance but are less commonly used[2].

In terms of technologies, tools offering explainability (particularly feature weight) are highly favored, particularly SVMs or, in an unsupervised setting, hierarchical clustering using distances such as Manhattan, Cosine, Burrows' Delta [22], or Eder's Delta [23], etc. We hypothesize that the need for explainability arises due to the fundamentally different objectives pursued by computational linguistics (CL) and digital humanities (DH) in the context of authorship verification and identification. While CL is primarily concerned with developing robust models at scale, DH seeks to employ the results as potential evidence in scholarly research. Consequently, the ability to interpret the decisions made by a model becomes invaluable for detecting "invisible biases to the human eye," such as the presence of unique characters in specific author editions, which might lead the model to overfit and yield erroneous conclusions. This risk is significantly amplified in the context of the frequent diachronic peculiarities prevalent in most DH inquiries or the constraints posed by a limited pool of available historical documents. Conversely, CL typically concentrates on short-term, cohesive, and large corpora, such as social network messages or fan fiction, where interpretability might not be as crucial.

In 2016, General Imposters (GI) methods [24], an authorship verification method, were introduced to DH [25]. Unlike authorship verification methods of CL, which rely on learning to recognize a true pair of texts from the same author and a pair of texts from different authors, GI introduces impostor authors alongside candidate authors in a classification task. It randomly removes features (50% by default in the *stylo* package [26]) over n experiments (100 by default in *stylo*) and proposes, for each of these experiments, the author of the closest text as a potential match. If one of the candidates has an overwhelming presence in the closest texts over all experiments, it is proposed as the verified author of the text.

3. Proposed Method

We propose to bridge the gap between DH and CL practices by introducing supervised authorship verification using text pairing validation through a siamese network. However, Voicu's anonymous PCs should not be used in the context of supervised training and authorship attribution. This is because supervised training relies on having ground truth data to identify the anonymous authors, and for some of Voicu's PCs, they may not even be processed through

[2]Probably due to the cost of said annotations and the lack of well-performing automatic models for such tasks.

the general impostors method as they only offer a single pair of texts, which is insufficient for effective training and verification.

Features Following most of the DH literature and PAN approaches until 2020, we select the relative frequency of the 1000 most frequent words (MFW), POS tri-grams, and trigram character affixes within each text. We reject the use of any punctuation-based features, as punctuation in our texts is the result of the editorial task and can be editor or period dependent (an editor from the 18th century might not punctuate like a contemporaneous one). As our texts in the Pseudo-Chrysostomian corpus are rather small, we follow Rebora's recommendation [27] to use a 1000-word sample of our texts for MFW and apply the same limit for affixes trigrams. For POS tri-grams, we resolve to use the 100 most frequent POS tri-grams (MFP). We concatenate the relative frequencies of these features into a single vector, denoted as features T_i, consisting of 2100 dimensions representing the text sample i.

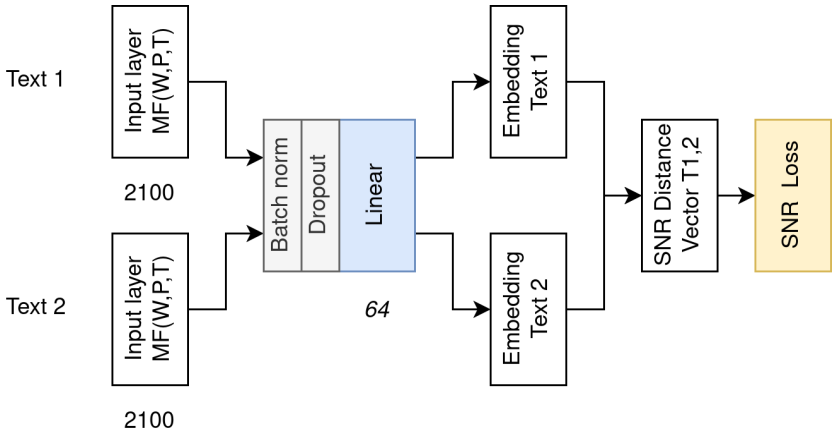


Figure 1: Architecture of our models: vectors are passed in parallel in the embedding projection layer and a distance vector is then computed.

Model structure The model follows a typical architecture of linear layers in the context of Siamese networks (see Figure 1): feature vectors are passed through the same projection layers, paired, and a distance-based decision is proposed. We use linear layers to reduce T_i to a given dimension m, resulting in an embedding representation E_i of the text. For the loss and distance metric, we employ the "Signal-to-Noise Ratio" distance [9] (SNR-Distance). The SNR-Distance considers the noise between an anchor embedding E_i and a compared embedding E_j, represented as $N_{ij} = E_j - E_i$, and uses its variance as an informative measure (variance of values across dimensions). The SNR for pair i, j is defined as $SNR_{i,j} = \frac{var(E_i)}{var(N_{ij})}$, and the SNR-Distance is $\text{SNRD}_{ij} = \frac{1}{\text{SNR}_{ij}}$. Unlike distance metrics such as the Euclidean or Manhattan distance, $\text{SNRD}_{ij} \neq \text{SNRD}_{ji}$, which encourages more comparisons.

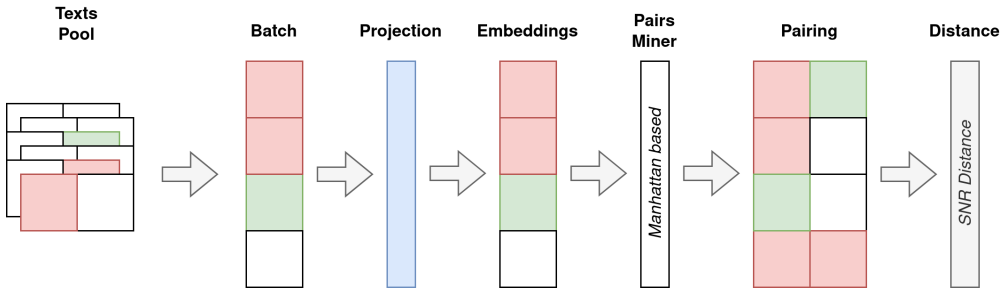


Figure 2: Learning pipeline using miner to provide pairs within the model.

Learning methods Unlike PAN tasks, we can build our own corpus without pre-established pairs, allowing us to utilize pair mining methods, which are particularly useful in the context of Siamese networks (see Figure 2). We use Easy–Semi-Hard (ESH) triplet mining [28, 8] using SNR distance. Given embeddings E and their classes K, $ESH(E, K)$ computes the distance between all embeddings and provides all positive pairs ($K_i == K_j$), as well as semi-hard and hard negative pairs. Negative pairs are considered semi-hard when "they are further away from the anchor than the positive exemplar, but still hard because the squared distance is close to the anchor-positive distance". We use this miner for both the training step and the evaluation step. In their recent paper on the state of authorship verification, Tyo et al. [29] showed that ESH mining could help feature-based models be as good as sequential models using transformers or similar layers.

4. Experimental Setup

Datasets Unlike social networks content or fan-fictions, or even 19th-century literature, Ancient Greek literature spans from 9 BCE to at least 9 CE in the context of our problem. We are dealing with variation in corpus genre, linguistic changes, ideological changes, etc. In order to address this at the training and evaluation step, we design a corpus that is focused on Christian and theological texts using the TLG CD-ROM [30] and their XML export through Diogenes [31]. Unfortunately, the raw dataset is not shareable, and no current open dataset provides a quantitatively large enough dataset for this period of Ancient Greek[3]. For POS tagging, we use Singh et al. [33]'s Ancient Greek BERT-based tagger, which is, to our knowledge, the only Ancient Greek tagger trained with both Medieval Greek and classical Greek.

Within the available corpus, we removed:

- any *dubia* or *spuria* (de-attributed texts) or anonymous texts;
- any text from John Chrysostom, as the sheer mass of his work and his style seemed to impact the model too much in early experiments, as well as both the *Old* and *New Testaments*;

[3]We hope, however, that the Patristic Text Archive [32] will reach one day this kind of milestone for reproducibility purposes. We provide the annotated and processed samples as well as the processing pipeline.

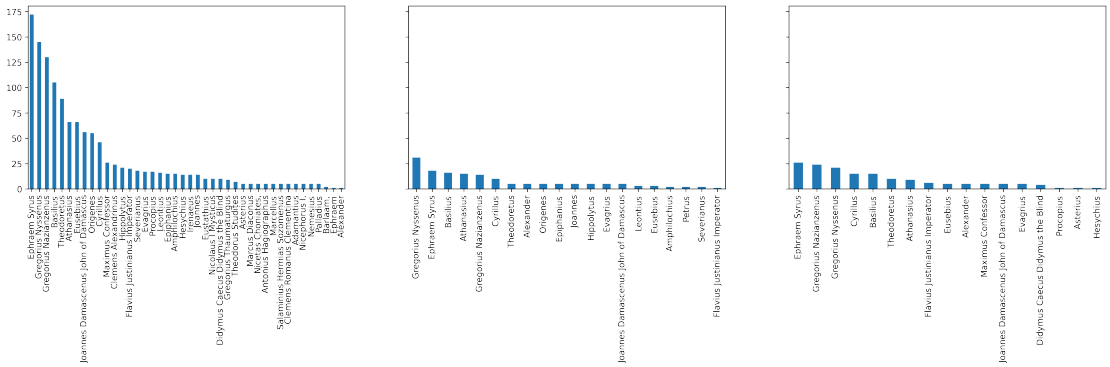


Figure 3: Sample count per author in the training (left), development (centre), and testing set (right).

- texts marked as commentaries, *scholia*, fragments, codices transcriptions, or specific *recensio*;
- any text containing a single line of poetry[4].

However, all of our most frequent features were extracted from the full Christian corpus.

To counter the reduced mass of texts and address the imbalanced state of the dataset, we produced 1000-word samples for each text using the following rules:

- if the text was shorter than 2000 words, we used the 1000 words in the middle of the text, in order to avoid introductions and conclusions.
- If the text is larger than 2000 words, we used up to 5 samples of 1000 words, randomly selected within the text except for the first 500 words and the last 500 words.

We then split the corpus in an 80-10-10 ratio along the titles of the various works, so that authors can span in different sets but samples of the same title remain inside the same text (see Figure 3). The final corpus is diverse in genre but is composed of two genres which make up around two-thirds of its texts (Homelia and Treatise) while Oratio, Letters, Hagiography, and other non-frequent genres rapidly decrease in numbers of titles (see Figure 4).

The PC corpus was not sampled, but relative frequencies were issued using the same process.

Hyper-parameters and metrics We chose to evaluate the general performance of the model based on the Area Under the Curve of the Receiver Operating Characteristics (AUROC), which allows for measuring the relationship between the False Positive Rate (FPR) and the True Positive Rate (TPR). This metric also enables us to select a task-oriented threshold by minimizing the FPR while maintaining a high enough TPR.

The hyperparameters used for the experiment were as follows: Adam optimizer, learning rate of $1e^{-4}$, embedding size of 64, no batch sampling, batch size of 64, 30% dropout of features, class sampling of 2, and a minimum of 100 epochs for training. Training was stopped after 20 consecutive bad epochs, using the dev loss as an indicator.

[4]Detected through the presence of TEI-1 tags in the XML files.

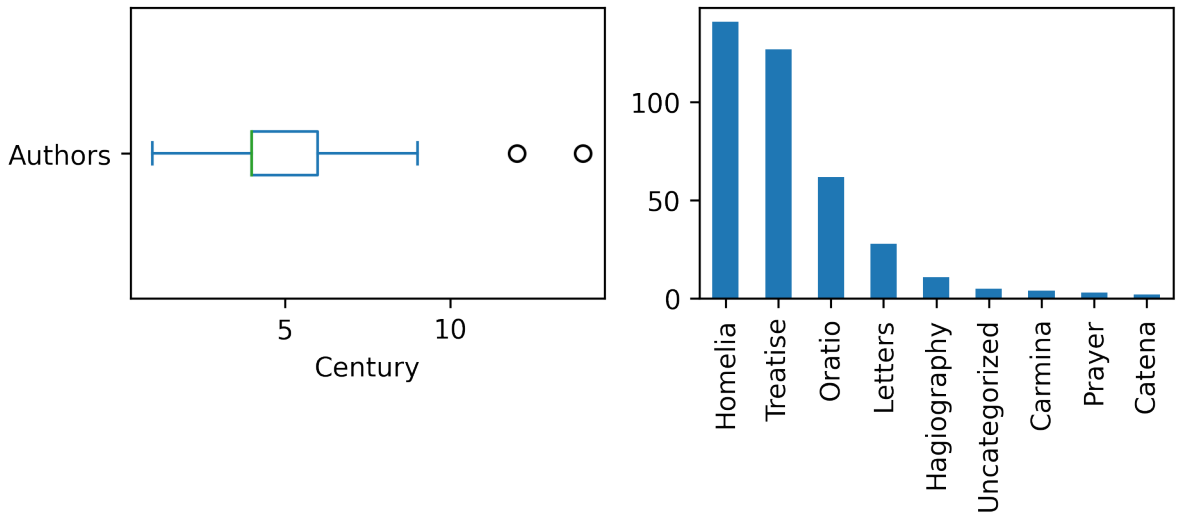


Figure 4: On the left, authors' century of activity, on the right, distribution of genres across works. Each author's century of activity is either provided by the century they are known to live in (uncertain dating) or the century for which they lived most of their adulthood (adulthood being considered as 20 years old and more).

Software implementation The experiment was implemented using the following software and hardware: it was run on an nVidia RTX 3090 GPU with 24GB of RAM, using `torch` [34], `torchmetrics` [35] for AUROC metrics, `pytorch-lightning` [36] for training, and `pytorch-metric-learning` [37] for the mining operations. The same process was also successfully run on a CPU (AMD Ryzen 5700 with 32GB RAM).

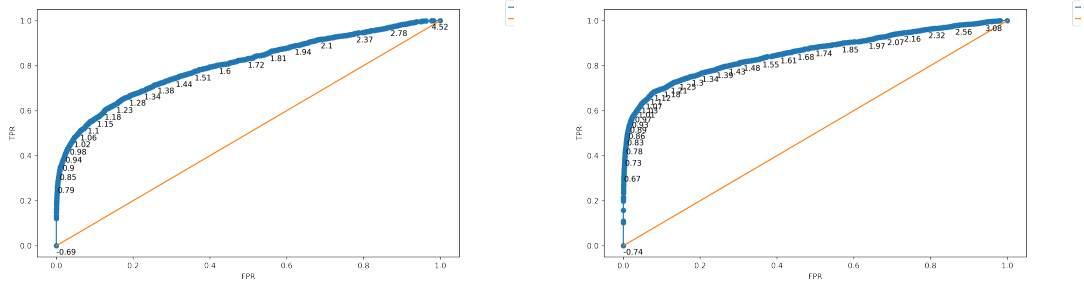
5. Results

	Authors	Texts	Samples	AucROC	1st FP Dist.	1st FP Classes
Train	42	390	1266	93.82	0.320	Origenes - Theodoretus
Dev	20	52	157	*80.16	0.597	Greg. Nyssen. - Leontius
Test	17	45	158	85.51	0.515	Basilius - Greg. Nazian.

Table 2
Details on the train, dev, and test sets along with their corresponding scores. "First FP Dist." stands for First False-Positive distance, which indicates the distance at which the first false positive is encountered. The AUCROC is computed on all pairs, such that most text have more negative pairs than positive pairs overall. The metrics are computed using A-B and B-A pairs.

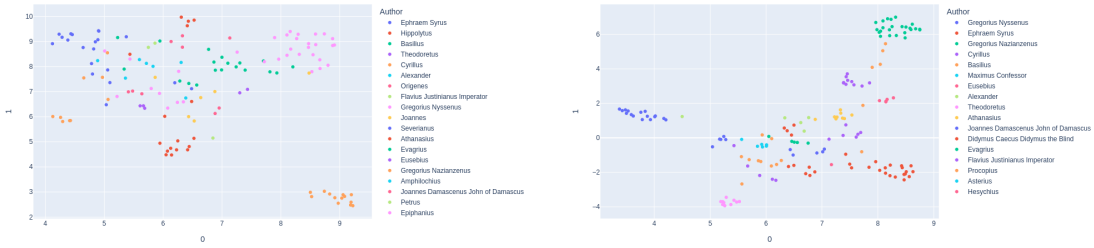
The results achieved in this study fall within the range of the current state of authorship verification. Tyo et al. [29]'s implementation of the N-Grams based model of Weerasinghe et al. [38] and ESH mining achieved between 77% to 91% AUCROC. The proposed model converges towards 79.5% AUCROC on the ESH mined pairs after 206 epochs and provides an 80.16%

AUCROC on all pairs of the development set (see Table 2). Surprisingly, the model reaches a slightly higher score on the test set with an 85.51% AUCROC. The difference in performance between the development set and the test set may be attributed to the random dispatch of works, resulting in a more diverse and potentially more challenging development set for the model. We also would like to note that due to ESH mining, the pairs used in the dev set for loss might change from one iteration to the others, as the embeddings are used to decide which pairs to use, resulting in a different kind of over-fitting than with a traditional fixed development set.



(a) Development set (b) Test set

Figure 5: ROC curves on both the development set and the test set. Y axis is the True Positive Rate, X axis is the False Positive Rate. Marks on the curve represent the distance threshold used to determine the positives.



(a) Development set (b) Test set

Figure 6: UMAP 2 dimension projections of the embeddings.

The ROC curves (see Figure 5) exhibit a compelling shape, with a significant slope in the initial percentages of the False Positive Rate (FPR), followed by a more gradual increase until reaching 100% True Positive Rate (TPR). This suggests that the model effectively distinguishes between positive and negative pairs, especially in the early stages of classification. UMAP [39] (Figure 6), while still serving as a proxy for the complexity of the embeddings' dimension, provides a clear depiction of the discrepancy between both sets. In the test set, embeddings of texts from the same authors are notably more distant from each other, indicating that the model's representations are better disentangled in this set. On the other hand, the development set

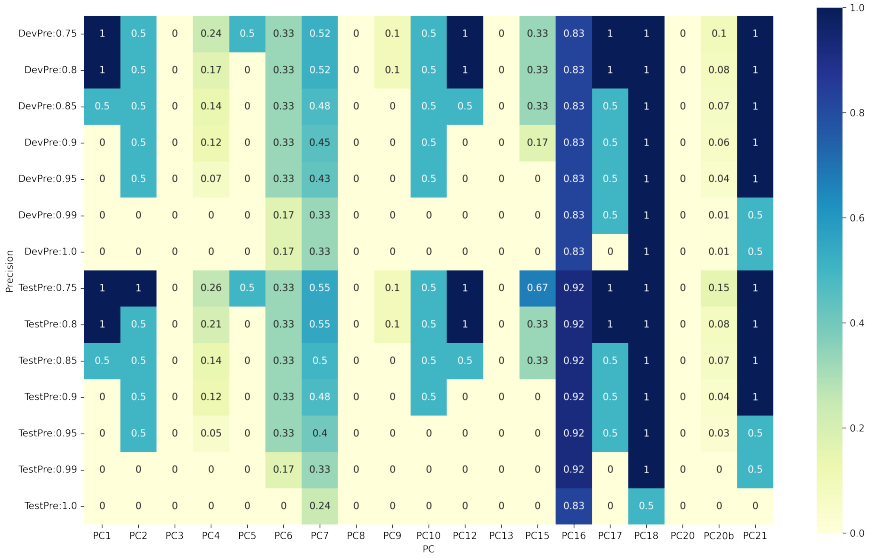


Figure 7: Heatmap of the precision thresholds for all pseudo-chrysostomian corpora. The percentage given in each cell gives the amount of pairs (A-B and B-A) of each PC that are found to be positive given the precision threshold.

displays more overlap and mixing of embeddings, suggesting that the model's representations are less distinct and might have captured some biases from the data splitting process.

6. Application on PCs

6.1. Method and general results

To apply our model to the pseudo-Chrysostomian corpus, we need to provide a confidence score to interpret the distance between pairs of texts. We choose to compute the precision of our model at a given distance threshold SNRD_{AB} for each pair A-B (and its reverse B-A) on both the development set and the test set. This threshold produces a "precision threshold" (PT) that allows us to estimate the probability of a false positive at any given distance. For example, if $\text{SNRD}_{AB} = 0.4$ and the precision of the model for distances less than or equal to 0.4 is 0.9, then $PT_{AB} = 0.9$, indicating a high probability that A and B are from the same author.

To provide a quick overview of all pseudo-Chrysostoms, we calculate the percentage of pairs within each pseudo-Chrysostomian corpus that are positively connected based on a relatively high precision threshold (Precision ≥ 0.75, see Figure 7). The overall results are remarkably consistent with the summary proposed by Voicu. Notably, the PC 3, 4, 5, 8, 9, and 13 are not confirmed as single authors by our model and are either indicated as low probability clusters or refuted groupings in Voicu's article. On the other hand, only the 20 and 20b clusters are

completely ignored by our model despite having a high chance of being from the same authors according to both Datema[2] and Voicu. Among the highly connected clusters, PC 1, 2, 12, 16, 18, and 21 are all marked as highly possible or confirmed by Voicu and are found at a precision threshold of 85% or higher.

6.2. Closer look at some pseudo-Chrysostoms

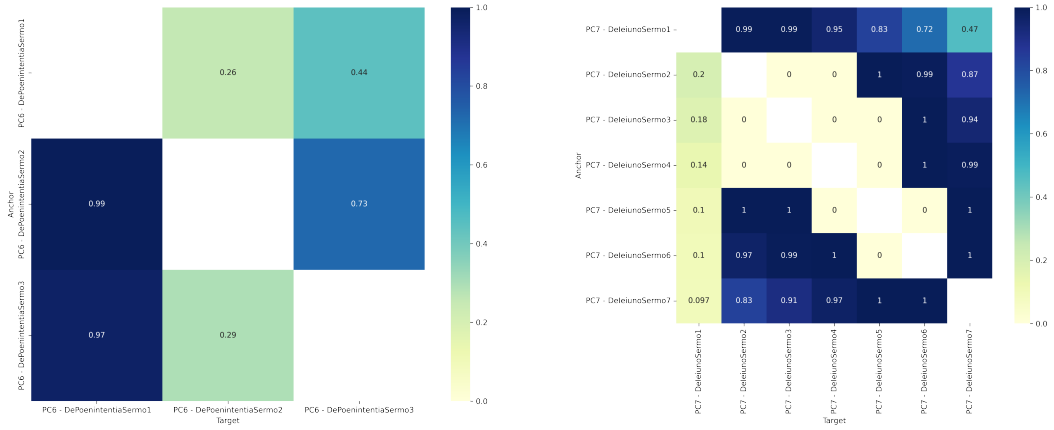
With PC 1–4, 8–9, 12–13, 16, 18, and 21 showing clear ranges of precision, we turn our attention to the remaining pseudo-Chrysostoms: PC 6–7, 10–11, 14–15, 17, 20, and 20b.

Montfaucon's pseudo-Chrysostoms 6 and 7 Both pseudo-Chrysostom 6 and 7 are hypotheses made by Montfaucon in the 18th century. PC6 is considered to be written by bad (and stupid) ancient Greek speakers (*inepti graeculi*), with quite repetitive and *ad nauseam* use of repetitions and epithets, while ignoring common rules of grammar (PG, 60, 681-682). Regarding PC7 and the *De jejunio* sermons, the only arguments of Montfaucon (PG, 60, 711-712) were again his absence of knowledge of basic Greek (*imperiti Graeculi*) and his ability to write mostly nonsensical things (*plerumque nugacis*).

Looking at the precision threshold matrices in Figure 8, we see that the matrix behaves in a non-reciprocating way, such that some texts are deemed to be of the same author in the sense A–B but not in the sense B–A. Moreover, for PC7, we see that consecutive sermons are unconnected between each other (*e.g. De jejuno 3* and *De jejuno 4* are not deemed to be of the same author at any precision) but are connected with subsequent or previous sermons (*e.g. De jejuno 3* and *De jejuno 4* are deemed to be of the same author as *De jejuno 5*). Given the scores, we might be tempted to consider PC6 and PC7 as two distinct authors, but we cannot confirm this without reasonable doubt. Our hypothesis, though untested, suggests that if PC6 and PC7 employ such an unconventional form of Ancient Greek compared to the majority of the corpus, they could prove challenging to categorize uniformly.

The case of Pseudo-Chrysostoms 13, 14, and 16 PC13, 14, and 16 are a particular case in Voicu's survey, as PC13 is refuted by Voicu and its texts are dispatched into two other subcorpora: two texts (*Contra Iudaeos, Gentiles et Haereticos* and *In uenerabilem*) are thought to go with PC14's *De Eleemonsyna*, while *De Epiphania* would go with PC16. Our precision matrix threshold does confirm the unity of PC16 except for the inclusion of PC13's text (see Figure 10a). However, while the pairs connections are low for PC13 and PC14, they seem to be rather high when connected to PC16. Our model seems to recommend the following composition for PC16:

- *De Eleemosyna* (PC14)
- *In Psalmum Homilia 50*, 1–2 (PC16)
- *In Illud Sufficit Tibi Gratia Mea* (PC16)
- *In Illud Si Qua Christo Nova Creatura* (PC16)
- (Lower probability) *Contra Iudaeos, Gentiles et Haereticos* (PC13*)
- (Lowest probability) *In Venerabilem Crucem* (PC13*)



(a) Pseudo-Chrysostom 6 (b) Pseudo-Chrysostom 7

Figure 8: Matrix of the precision thresholds for authors whose bad prose is considered a proof of authorship.

Pseudo-Chrysostom 20 and 20b: What happened? The most significant disagreement between traditional scholarship and our model arises with PC20, which is classified as unconfirmed by our model (see Figure 9). Voicu mentions PC20 as the outcome of "on-going research" and comprises 6 base texts (PC20) and 12 potential others (PC20b). Unfortunately, some of these texts are unedited, and we did not have access to them. As of now, none of the texts in PC20 and PC20b form prominent clusters. Moreover, the precision thresholds for most pairs with high values are not reciprocated when considering the inverse direction:

- 0.98 *In Drachmam ...–De Remissionem...*, with a 55-point drop in the other direction;
- 0.97 *De Turture Seu ...–In Decollationem S. Ioannis*, with a 26-point drop in the other direction;
- 0.97 *In Rachelem et in Infantes–In Decollationem S. Ioannis*, with a 79-point drop in the other direction.

The lack of clear clustering and the significant differences in precision threshold values between pairs in opposite directions raise questions about the coherence and authorship attributions within PC20 and PC20b. It is essential to recognize that authorship verification is a complex task, and discrepancies between our model's findings and traditional scholarship can be attributed to various factors, including the linguistic features used for analysis, the dataset's size and quality, and the nature of the texts themselves.

Others For other Pseudo-Chrysostom, we propose the following analysis:

- Pseudo-Chrysostom 10 (Figure 10c): This cluster falls into the group of Pseudo-Chrysostoms for which the A–B and B–A distances differ. As it consists of a single pair of texts, we would be inclined to confirm the assumption of Weyer[40] regarding their common authorship.

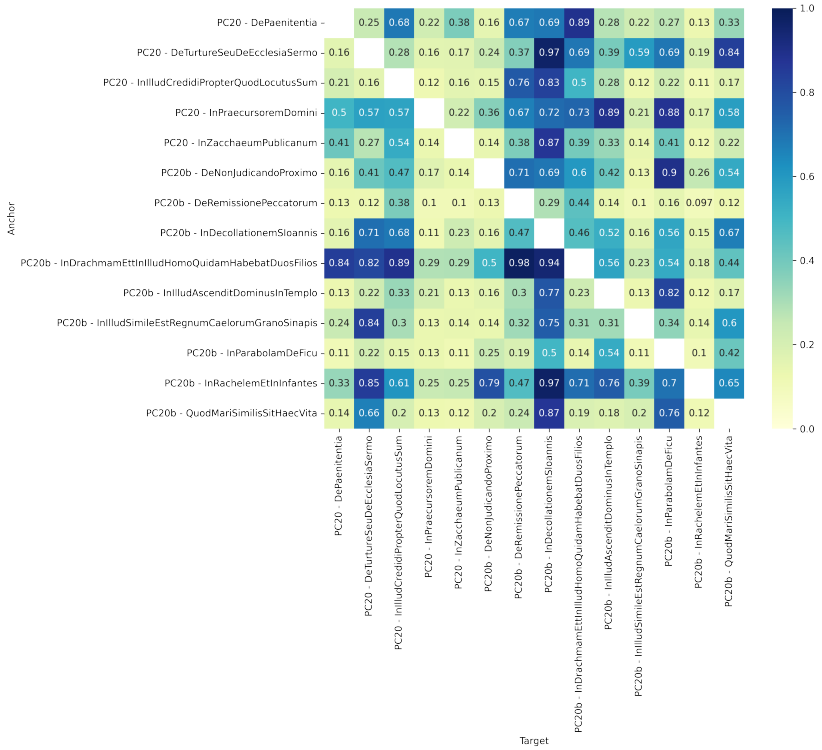
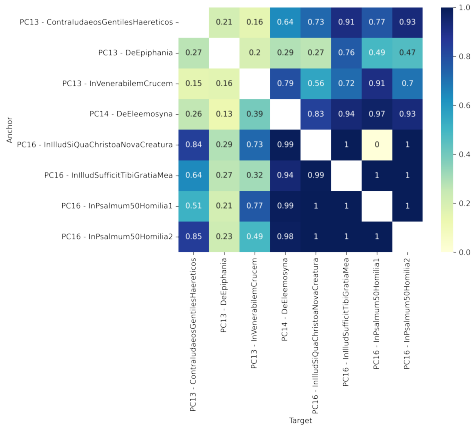


Figure 9: Matrix of the precision thresholds for the biggest unconfirmed cluster: PC20 and its extension PC20b.

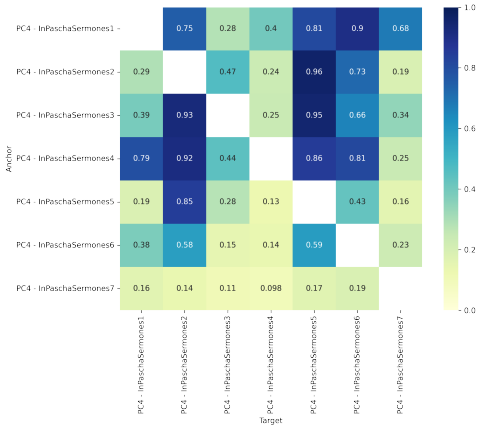
- Pseudo-Chrysostom 11 (Figure 10b): This cluster is a subset of Pseudo-Chrysostom 4 and comprises 7 sermons, of which the first three are believed to be of a single author according to Nautin[41]. While the unity of the PC4 corpus is not confirmed through our model, our results are more aligned with the PC11 hypothesis, but we would not go as far as confirming it, as the cluster suffers from unmirrored pair distances.
- Pseudo-Chrysostom 15 (Figure 10d): In this cluster, we observe a strong reciprocal connection between *De Pharisaeo* and *Ignem Veni Mittere*, while *In Illud Hominis...* shows high variation depending on the pair directions.
- Pseudo-Chrysostom 17 (Figure 10e): This cluster exhibits significant variation (0.32 vs. 0.87 precision) depending on the direction of its pairs.

7. Conclusion

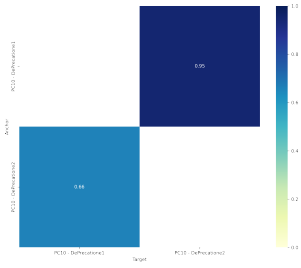
Authorship verification is a crucial task in the field of computational humanities and the humanities in general, as it offers a new approach to validate older hypotheses made using traditional philological methods, such as transmission study or stylistic analysis. This is particularly important in patristic studies, where pseudo-author corpora, such as those attributed to Augustine



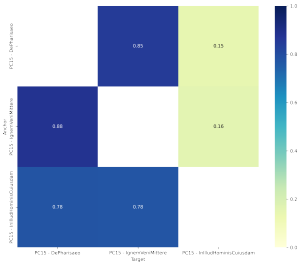
(a) Precision threshold matrix for Pseudo-Chrysostoms 13, 14, and 16.



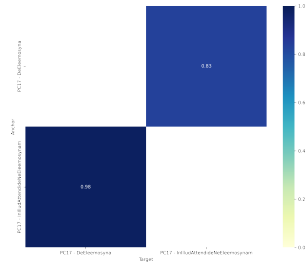
(b) Pseudo-Chrysostom 4, of which *sermos* 1-3 are though to be of the same author by Nautin[41].



(c) Pseudo-Chrysostom 10



(d) Pseudo-Chrysostom 15



(e) Pseudo-Chrysostom 17

Figure 10: Matrix of the precision thresholds.

or John Chrysostom, are significant. The challenge lies in dealing with pseudonymous corpora where most authors are unknown.

To address this, we proposed exploring the use of Siamese Networks with embeddings based on known stylometric features, such as the relative frequencies of most frequent words, affixes, and POS 3-grams. To optimize our approach, we leveraged pair mining and signal-to-noise ratio distance, both originally designed for Siamese network architectures. The results we obtained on our development and test sets are very promising, and we argue that, unlike authorship attribution problems, authorship verification problems are less susceptible to overfitting when using such tools.

Our findings mostly align with the survey conducted by Voicu in 1981, which identified 21 potential unique pseudo-authors in the corpus of the pseudo-Chrysostomian texts (see Table 3). Some pseudo-Chrysostoms are highly probable under both the framework of classical philological studies and our approach. However, we encountered discrepancies with two pseudo-Chrysostoms. First, PC1, hypothesized by Montfaucon and refuted by Altendorf [42], is mostly confirmed within our framework. Second, we were not able to confirm PC20. It's important to note that our method is designed for precision rather than recall, so this result does not necessarily refute Voicu's hypothesis but still stands out compared to the rest of our results.

PC	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	20b	21
In Voicu, 1981	R	C	R	LP	LP	LP	LP	R	R	C	C	C	R	C	C	C	C	C	*	C	C	C
Our model	HP	HP	U	U	U	LP	HP	U	U	HP	LP	C	U	U	LP	C	HP	C	*	U	U	C

Table 3
Analysis summary. C stands for confirmed, HP high probability, LP low probability, R refuted. On the second line, U stands for cluster unconfirmed, which differs from Refutation as our model is not trained for recall. PC19 could bot been studied because of the Syriac content.

In future work, we are interested in extending our approach without relying on a learned embedding space and instead using probabilistic tools for better explainability. This kind of work has already been explored by Weerasinghe et al. [38] and would be a valuable addition to the digital humanities landscape. Further analysis on PC20 should be produced, specifically by looking at the research produced by Voicu since 1981 on this particular topic.

References

[1] S. J. Voicu, Une nomenclature pour les anonymes du corpus pseudo-chrysostomien, Byzantion 51 (1981) 297–305. URL: http://www.jstor.org/stable/44170685.

[2] C. Datema, An unedited homily of ps. chrysostom on the birth of john the baptist (bhg 843k), Byzantion 52 (1982) 72–82. URL: http://www.jstor.org/stable/44170752.

[3] M. Schatkin, The authenticity of St. John Chrysostom's de Sancto Babyla, Contra Iulianum et gentiles, volume 1 of *Kyriakon*, Aschendorff, 1970, p. 474–489.

[4] M. Eder, Style-markers in authorship attribution: a cross-language study of the authorial fingerprint, Studies in Polish Linguistics 6 (2011).

[5] F. Cafiero, J.-B. Camps, Why molière most likely did write his plays, Science advances 5 (2019) eaax5489.

[6] E. Stamatatos, M. Kestemont, K. Kredens, P. Pezik, A. Heini, J. Bevendorff, B. Stein, M. Potthast, Overview of the authorship verification task at pan 2022, in: CEUR workshop proceedings, volume 3180, 2022, pp. 2301–2313. URL: https://ceur-ws.org/Vol-3180/paper-184.pdf.

[7] M. Kestemont, E. Manjavacas, I. Markov, J. Bevendorff, M. Wiegmann, E. Stamatatos, M. Potthast, B. Stein, Overview of the cross-domain authorship verification task at pan 2021, in: Working notes of CLEF 2021-Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania, volume 2936, 2021, pp. 1743–1759. URL: https://ceur-ws.org/Vol-2936/paper-147.pdf.

[8] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.

[9] T. Yuan, W. Deng, J. Tang, Y. Tang, B. Chen, Signal-to-noise ratio: A robust distance metric for deep metric learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4815–4824.

[10] J.-L. Quantin, Du chrysostome latin au chrysostome grec: une histoire européenne (1588-1613), Chrysostomosbilder in 1600 Jahren (2008) 267–346.

[11] B. Marx, Procliana: Untersuchung über den homiletischen nachlass des patriarchen proklos von konstantinopel, Münsterische Beiträge zur Theologie 23 (1940).
[12] M. L. Pacheco, K. Fernandes, A. Porco, Random forest with increased generalization: A universal background approach for authorship verification., in: CLEF (Working Notes), 2015.
[13] S. Nikolov, D. Tabakova, S. Savov, Y. Kiprov, P. Nakov, Su@ pan'2015: Experiments in author verification., in: CLEF (Working Notes), 2015.
[14] M. Kestemont, M. Tschuggnall, E. Stamatatos, W. Daelemans, G. Specht, B. Stein, M. Potthast, Overview of the author identification task at pan-2018: Cross-domain authorship attribution and style change detection, in: Working notes of CLEF 2018-Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018., volume 2125, 2020, pp. 1–14. URL: https://ceur-ws.org/Vol-2125/invited_paper_2.pdf.
[15] M. Kestemont, E. Manjavacas, I. Markov, J. Bevendorff, M. Wiegmann, E. Stamatatos, M. Potthast, B. Stein, Overview of the cross-domain authorship verification task at pan 2020, in: Working notes of CLEF 2020-Conference and Labs of the Evaluation Forum, 22-25 September, Thessaloniki, Greece, volume 2696, 2020, pp. 1–14. URL: https://ceur-ws.org/Vol-2696/paper_264.pdf.
[16] J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Generalizing unmasking for short texts, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 654–659. URL: https://aclanthology.org/N19-1068. doi:10.18653/v1/N19-1068.
[17] M. Eder, Rolling stylometry, Digital Scholarship in the Humanities 31 (2015) 457–469. URL: https://doi.org/10.1093/llc/fqv010. doi:10.1093/llc/fqv010. arXiv:https://academic.oup.com/dsh/article-pdf/31/3/457/21517799/fqv010.pdf.
[18] S. Rebora, J. B. Herrmann, G. Lauer, M. Salgaro, Robert Musil, a war journal, and stylometry: Tackling the issue of short texts in authorship attribution, Digital Scholarship in the Humanities 34 (2018) 582–605. URL: https://doi.org/10.1093/llc/fqy055. doi:10.1093/llc/fqy055. arXiv:https://academic.oup.com/dsh/article-pdf/34/3/582/29212867/fqy055.pdf.
[19] J.-B. Camps, T. Clérice, A. Pinche, Noisy medieval data, from digitized manuscript to stylometric analysis: Evaluating paul meyer's hagiographic hypothesis, Digital Scholarship in the Humanities 36 (2021) ii49–ii71.
[20] B. Nagy, Metre as a stylometric feature in Latin hexameter poetry, Digital Scholarship in the Humanities 36 (2021) 999–1012. URL: https://doi.org/10.1093/llc/fqaa043. doi:10.1093/llc/fqaa043. arXiv:https://academic.oup.com/dsh/article-pdf/36/4/999/41027260/fqaa043.pdf.
[21] R. Gorman, Author identification of short texts using dependency treebanks without vocabulary, Digital Scholarship in the Humanities 35 (2019) 812–825. URL: https://doi.org/10.1093/llc/fqz070. doi:10.1093/llc/fqz070. arXiv:https://academic.oup.com/dsh/article-pdf/35/4/812/34084883/fqz070.pdf.
[22] J. Burrows, 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship, Lit-

erary and Linguistic Computing 17 (2002) 267–287. URL: https://doi.org/10.1093/llc/17.3.267. doi:10.1093/llc/17.3.267. arXiv:https://academic.oup.com/dsh/article-pdf/17/3/267/2743069/170267.pdf.

[23] M. Eder, Taking stylometry to the limits: Benchmark study on 5,281 texts from patrologia latina, in: Digital humanities 2015: conference abstracts, 2015, pp. 1919–1924.

[24] M. Koppel, Y. Winter, Determining if two documents are written by the same author, Journal of the Association for Information Science and Technology 65 (2014) 178–187.

[25] M. Kestemont, J. Stover, M. Koppel, F. Karsdorp, W. Daelemans, Authenticating the writings of julius caesar, Expert Systems with Applications 63 (2016) 86–96.

[26] M. Eder, J. Rybicki, M. Kestemont, Stylometry with r: a package for computational text analysis, The R Journal 8 (2016).

[27] Short texts with fewer authors. Revisiting the boundaries of stylometry, Zenodo, 2023. URL: https://doi.org/10.5281/zenodo.7961822. doi:10.5281/zenodo.7961822.

[28] H. Xuan, A. Stylianou, R. Pless, Improved embeddings with easy positive triplet mining, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 2474–2482.

[29] J. Tyo, B. Dhingra, Z. C. Lipton, On the state of the art in authorship attribution and authorship verification, 2022. arXiv:2209.06869.

[30] L. Berkowitz, K. A. Squitier, M. Pantelia, Thesaurus linguae graecae. canon of greek authors and works., 2020.

[31] P. Heslin, Diogenes, 2023.

[32] A. von Stockhausen, Die modellierung kritischer editionen im digitalen zeitalter, Zeitschrift für Antikes Christentum/Journal of Ancient Christianity 24 (2020) 123–160.

[33] P. Singh, G. Rutten, E. Lefever, A pilot study for bert language modelling and morphological analysis for ancient and medieval greek, in: 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, co-located with EMNLP 2021, Association for Computational Linguistics, 2021, pp. 128–137.

[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[35] Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, William Falcon, TorchMetrics - Measuring Reproducibility in PyTorch, 2022. URL: https://github.com/Lightning-AI/torchmetrics. doi:10.21105/joss.04101.

[36] W. Falcon, The PyTorch Lightning team, PyTorch Lightning, 2019. URL: https://github.com/Lightning-AI/lightning. doi:10.5281/zenodo.3828935.

[37] K. Musgrave, S. J. Belongie, S.-N. Lim, Pytorch metric learning, ArXiv abs/2008.09164 (2020).

[38] J. Weerasinghe, R. Singh, R. Greenstadt, Feature vector difference based authorship verification for open-world settings., in: CLEF (Working Notes), 2021, pp. 2201–2207.

[39] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 (2018).
[40] J. Weyer, De homiliis quae Joaenni Chrysostomo falso attribuuntur, Ph.D. thesis, Bonn, 1952.
[41] P. Nautin, Homélies pascales: II. Trois homélies dans la tradition d'Origene, volume 36, Sources chrétiennes, 1953.
[42] H. Altendorf, Untersuchungen zu Severian von Gabala, Ph.D. thesis, Tübingen, 1957.

8. Online Resources

Code and data for the Voicu experiment are available at https://github.com/PonteIneptique/Chryso-Voicu.