MedLatinEpi and MedLatinLit: Two Datasets for the Computational Authorship Analysis of Medieval Latin Texts

SILVIA CORBARA, Scuola Normale Superiore, Italy
ALEJANDRO MOREO, Consiglio Nazionale delle Ricerche, Italy
FABRIZIO SEBASTIANI, Consiglio Nazionale delle Ricerche, Italy
MIRKO TAVONI, Università di Pisa, Italy

We present and make available MedLatinEpi and MedLatinLit, two datasets of medieval Latin texts to be used in research on computational authorship analysis. MedLatinEpi and MedLatinLit consist of 294 and 30 curated texts, respectively, labelled by author; MedLatinEpi texts are of epistolary nature, while MedLatinLit texts consist of literary comments and treatises about various subjects. As such, these two datasets lend themselves to supporting research in authorship analysis tasks, such as authorship attribution, authorship verification, or same-author verification. Along with the datasets we provide experimental results, obtained on these datasets, for the authorship verification task, i.e., the task of predicting whether a text of unknown authorship was written by a candidate author or not. We also make available the source code of the authorship verification system we have used, thus allowing our experiments to be reproduced, and to be used as baselines, by other researchers. We also describe the application of the above authorship verification system, using these datasets as training data, for investigating the authorship of two medieval epistles whose authorship has been disputed by scholars.

CCS Concepts: • **Computing methodologies** → Supervised learning by classification; • **Applied computing** → **Arts and humanities**.

Additional Key Words and Phrases: Authorship Analysis, Authorship Verification, Medieval Latin

1 INTRODUCTION

(Computational) *Authorship Analysis* is the task of inferring the characteristics of the author of a text of unknown or disputed authorship. Authorship Analysis has several subtasks of practical use; examples include *gender detection* (i.e., predicting whether the text was written by a woman or a man [11]), or *native language identification* (i.e., predicting the native language of the author of the text [18]).

Many subtasks of authorship analysis have actually to do with the prediction of the *identity* of the author of the text. The one such subtask that has the longest history is *Authorship Attribution* (AA) [7, 13, 23], which consists of predicting who, among a set of *k* candidate authors, is the real (or the most likely) author of the text. A task that has gained prominence more recently is *Authorship Verification* (AV) [12, 24], the task of predicting if a certain candidate author is or is not the author of the text. Finally, the task that has been introduced latest in

The order in which the authors are listed is purely alphabetical; each author has given an equally important contribution to this work.

Authors' addresses: Silvia Corbara, Scuola Normale Superiore, 56126, Pisa, Italy, silvia.corbara@sns.it; Alejandro Moreo, Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, 56124, Pisa, Italy, alejandro.moreo@isti.cnr.it; Fabrizio Sebastiani, Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, 56124, Pisa, Italy, fabrizio.sebastiani@isti.cnr.it; Mirko Tavoni, Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa, 56126, Pisa, Italy, mirko.tavoni@unipi.it.

1556-4673/2022/5-ART $15.00
https://doi.org/10.1145/3485822

this field is *Same-Authorship Verification* (SAV) [14], the task of predicting whether two texts d' and d'' are by the same author or not.

Nowadays, authorship analysis tasks are usually tackled as *text classification* tasks [1], and thus solved with the help of machine learning algorithms: for instance, an authorship verification task is solved as a *binary classification* problem, i.e., as the problem of classifying the disputed text into one of the two classes {Yes, No}, where Yes (resp., No) indicates that the text is (resp., is not) by the candidate author. In order to do so, a machine learning algorithm trains a {Yes, No} classifier from a training set of labelled texts, where the training examples labelled Yes are texts by the candidate author and the training examples labelled No are texts by other authors, usually closely related to the candidate author.

Authorship analysis is useful for many applications, ranging from cybersecurity (the field that addresses the design of techniques for preventing crimes committed via digital means) [22], to computational forensics (the field concerned with the study of digital evidence for investigating crimes that have already occurred) [3, 15, 18, 19]. Another important application is related to philology, and has to do with inferring the identity of the unknown authors of texts of literary and historical value. In the case of modern texts, this often has to do with the attempt to disclose the identity of authors who originally wanted to remain anonymous, or to disguise as someone else, while in the case of ancient texts this usually has to do with texts whose authorship has *become* unknown, or uncertain, in the course of history [8, 10, 21, 25, 27].

After reviewing some related work (Section 2), we here present and make available (Section 3) two datasets of texts of the latter type, i.e., texts written in medieval Latin, mostly by Italian literates, mostly dating around the 13th and 14th century.[1] We believe this to be an important contribution for at least two reasons. The first is that the datasets bring together (in preprocessed form for use by authorship analysis researchers) a set of texts that were not readily available to these researchers, since some of these texts were not available in digital form, while others lay scattered across different electronic formats and different digital libraries. The second is that there are many documents in medieval Latin from this historical period whose authorship is disputed by scholars,[2] and this makes an authorship analysis system trained on these datasets an important tool for philologists and historians of language alike.

Aside from describing the two datasets, we make available the source code of MedieValla,[3] a software tool for running authorship verification experiments on medieval Latin texts, and we present (Section 4) the results of our experiments using MedieValla on these datasets. The availability of both the datasets and the tool we have used on them, will allow other researchers to replicate our results and, hopefully, to develop and test improved authorship verification methods for medieval Latin.

In Section 5 we present two example applications of MedieValla on MedLatinEpi and MedLatinLit. In the first we verify if the *Epistle to Cangrande*, an epistle traditionally attributed to Dante Alighieri, but which several scholars have conjectured to be a forgery, is actually by Dante or not. In the second we verify if an epistle traditionally attributed to Cangrande della Scala, but which has recently been conjectured to be by Dante himself, is indeed by Dante or not. In both cases our authorship verifier rejects the hypothesis that Dante may be the author, and does so with high confidence.

[1] Medieval Latin is different from classical Latin in a number of ways, e.g., it is more generous than classical Latin in its use of prepositions and conjunctions, and it uses a more regular syntax.

[2] Examples include the *Epistle to Cangrande* [2], *Cangrande's Epistle to Henry VII* [17], and the *Quaestio de aqua et terra* [26], just to mention ones that some scholars attribute to Dante Alighieri while some others do not. The first two will be discussed more fully in Section 5.

[3] The name MedieValla is a combination of "medieval" and the last name of Lorenzo Valla (1407–1457), one of the first (human) authorship verifiers recorded in history. Lorenzo Valla is well-known for proving that the so-called "Donation of Constantine" (a decree attributed to 4th-century Roman emperor Constantine in which he supposedly conferred authority over Rome and the western part of the Roman Empire to the Pope) was a forgery.

2 RELATED WORK

In our cultural heritage, documents of unknown or disputed authorship are rather common, especially in centuries-old traditions, where the testimony of the true author may easily have been lost or altered. In particular, a number of recent works have tackled problems of authorship analysis for the Latin language.

In [9], Kestemont et al. address an authorship attribution task characterised by two disputed documents written in medieval Latin, and three possible authors – the well-known Christian mystic Hildegard of Bingen, her secretary Guibert of Gembloux, and Bernard of Clairvaux. They employ a PCA-based approach on the frequencies of 65 function words. In a later work [10], Kestemont et al. tackle another authorship attribution task concerning parts of the *Corpus Caesarianum*, including in the candidate set Caesar, his general Aulus Hirtius, and three other unidentified authors. The methodologies they adopt is based on comparing an author's profile (where an author's profile is defined as the centroid of the vectors corresponding to that author's known texts) with the document of disputed authorship. Two different techniques are employed in this work, namely, the distance between the vectors representing the author's profile and the disputed document, and a generic implementation of Koppel and Winter's "impostors method" [14]. With both techniques the authors use word unigrams and character *n*-grams as features, and test their systems on the datasets from the Authorship Verification track at PAN2014 and on a corpus of historic Latin authors. They perform experiments with various distance metrics and vector space models for both techniques.

An approach that similarly exploits the concept of author profile can be found in [27], a study regarding the authenticity of one of Pliny the Younger's letters. In particular, the author employs the "simplified profile intersection", a similarity measure that uses the size of the intersection among the profile of the unknown document and that of the target author's production, which is computed by counting the *n*-grams in common between them. In order to find the model with the best discriminating power between Pliny's and non-Pliny's writing, additional fragments of letters from Cicero and Seneca are employed.

Stover et al.'s [25] is yet another work that employs Koppel and Winter's "impostors method" [14]. Here, a newly found Latin document is investigated in a same-authorship verification setting, where word unigrams and bigrams are used as features. Ultimately, the only textual pair that receives a satisfying positive score is the one consisting of the disputed document and *De Platone* by Apuleius, hence strongly supporting the hypothesis that Apuleius may be the author of the document.

Vainio et al.'s [28] is the only study, among the ones we consider here, that uses a deep-learning algorithm. In particular, the authors train both an SVM and a CNN for an authorship verification task, consisting of recognising Cicero's written style against the styles of the background authors, and then use the two trained classifiers to classify four disputed documents. They conduct various experiments with POS-grams and character 5-grams. This dataset, which is freely downloadable, is the one with the largest number of authors among the works discussed in this section, counting 44 authors, including anonymous and pseudo-authors; this is thanks to the wide timeframe considered, which goes from the 1st century BC to the 5th century AD.

Kabala's [8] is the only work that, like the present paper, focuses on *medieval* Latin, although dating from an earlier period than the one we consider. It performs same-authorship verification on two texts, the *Translatio s. Nicolai* and the *Gesta principum polonorum.* In particular, through the studies on four different datasets, the author seeks to understand whether the alleged authors of the two documents, the so-called Monk of Lido and Gallus Anonymous, are actually the same person. The study is conducted by classifying both texts with respect to the author classes within each dataset, using 9 distance metrics and logistic regression. Each dataset counts between 39 and 116 texts dating from the 10th to the 12th centuries, written by between 15 and 22 different authors. These are the only datasets of medieval Latin texts that are freely available to the public among the ones we have surveyed in this section.

Table 1. Main characteristics of published works on authorship analysis for the Latin language reviewed in Section 2. “AA” stands for authorship attribution, “AV” stands for authorship verification, while “SAV” stands for same-author verification. The works are in alphabetical order by first author.

	Task	Number of authors in dataset	Methods used	Features used	Dataset	Makes the dataset available
Forstall et al. [6]	AV	7	SVM	Functional *n*-grams (on text and metric) and low-probability *n*-grams	Transcriptions and Tesserae	No
Kabala [8]	SAV	15–22	Distance metrics and Logistic regression	250 most frequent words	Patrologia Latina and Latin Library	Yes
Kestemont et al. [9]	AA	3	PCA	65 function words	Brepols Publishers	No
Kestemont et al. [10]	AA	29 (dev) 3 (test)	Distance metrics on author’s profile and Impostors method	Word unigrams and char *n*-grams	Latin Library	Yes
Stover et al. [25]	SAV	36	Impostors method	Word unigrams and bigrams	Brepols Publishers and Latin Library and Patrologia Latina	Partially
Tuccinardi [27]	AV	3	Simplified Profile Intersection	Character *n*-grams	[unspecified]	No
Vainio et al. [28]	AV	44	SVM and CNN	Pos-tags, word and char *n*-grams	Latin Library and Bibliotheca Augustana	Yes

While the above works focus on cases of uncertain paternity, such methodologies might also be applied to documents of certain authorship, e.g., in order to study possible stylistic influences among authors. In Forstall et al.’s work [6], for example, the goal is to verify a supposed influence by Catullo on the poetry of Paul the Deacon. Forstall et al.’s idea is to train an SVM with samples of Catullo’s writings (in a typical authorship verification setting), employing various kinds of *n*-grams as features. A document highly influenced by Catullo, thus bearing many similarities to his style, will then receive a high classification score by the AV system.

In Table 1 we summarise the works discussed in this section, specifying the task being tackled, the number of authors in the dataset, the method of analysis and the features employed, the dataset sources, and whether the dataset is publicly available or not.

In general, it should be noted that the authors do not subject their datasets to a thorough cleaning from information extraneous to the author’s production. In particular, citations of other authors (i.e., pieces of text that are by someone other than the author of the citing text) are seldom removed (in some cases, only the most extensive ones are); this may hamper authorship analysis, since cited text “contaminates” the citing text, at least as far as authorship analysis is concerned. This is unlike the present paper, where cited text is scrupulously removed.

3 THE DATASETS

3.1 Origin of the datasets

Our two datasets originated in the context of an authorship verification research work [4, 5] that we carried out in order to establish, using an approach based on machine learning, whether the *Epistle to Cangrande*, originally attributed to Dante Alighieri, is actually a forgery or not, a fact which is intensely debated among philologists today [2]. The *Epistle to Cangrande* is traditionally listed as the 13th of Dante's epistles that have reached us; hereafter we will thus refer to it as Ep13.

Ep13 is written in medieval Latin and addressed to Cangrande I, ruler of the Italian cities of Verona and Vicenza at the beginning of the 14th century. Scholars traditionally divide it into two portions that are distinct in purpose and, consequently, style: the first portion (paragraphs 1–13, hereafter: Ep13(I)) is the dedicatory section, with proper epistolary characteristics, while the second portion (paragraphs 14–90, hereafter: Ep13(II)) contains an exegesis (i.e., analysis) of Alighieri's *Divine Comedy*, and in particular a commentary of the first few lines of its third part, the *Paradise*. Scholars are not unanimous as whether Dante Alighieri is the true author of Ep13: some of them consider both portions authentic, some consider both portions the work of a forger, while others consider the first part authentic and the other a forgery.

Since it is unclear whether the two portions are by the same author or not, we tackled our AV problem as two separate AV sub-problems, one for Ep13(I) and one for Ep13(II). Because of the different nature of the two portions, we built two separate training sets, one for Ep13(I) and one for Ep13(II); we will refer to them as MedLatinEpi (where "Epi" refers to the epistolary nature of the texts contained therein) and MedLatinLit (where "Lit" stands for literary), respectively.

In both MedLatinEpi and MedLatinLit Dante Alighieri is, of course, the author of some of the labelled texts. The texts attributed with certainty to Alighieri and written in Latin are few and well known; we have thus included all of them.[4] Concerning other authors, the approach we have chosen is to select literates who are as "close" (culturally and stylistically) to Dante Alighieri as possible, i.e., authors whose production is characterised by linguistic features similar to Alighieri's. The reason for this choice, of course, is that, if the non-Dantean texts used for training were very different from Dante's training texts, any text even vaguely similar to Dante's production would be recognised as Dantean, the classifier being untrained to make subtle distinctions. Instead, one can expect better results if the classifier is trained to spot minimal differences. We have thus done a large-scale screening of authors who have written in Latin around the same historical period of Dante's, and who have written works of either an epistolary or literary nature; since the included authors are close to each other, in the above-mentioned cultural-stylistic sense, the two resulting datasets are challenging ones for computational authorship analysis systems.

While we used MedLatinEpi and MedLatinLit as training sets for our Ep13 work, of course they can be used as datasets for medieval Latin AV research that does not necessarily involve Ep13 (we will discuss such an example in Section 5), or as datasets for other authorship analysis tasks that address medieval Latin, or as benchmarks for general-purpose, language-agnostic authorship verification systems. This is the reason why we make them available to the research community.

3.2 Composition and preprocessing of the datasets

The composition of our two datasets is described in detail in Tables 2 and 3.

MedLatinEpi is composed of texts of epistolary genre (given that this is the nature of Ep13(I)) mostly dating back to the 13th and 14th centuries, for a total of 294 epistles; the average length of these epistles is 378 words.

[4] We have not included the *Quaestio de aqua et terra*, a work traditionally attributed to Dante Alighieri, exactly because its authorship is currently disputed. Other works by Alighieri, such as his masterpiece *Divina Commedia*, are not included because they are written not in Latin but in the Florentine vernacular, the language that would later form the basis of the Italian language.

Most of the texts are actually entire collections of epistles; we consider each epistle as a single training text. Note that, concerning the epistles by Guido Faba and Pietro della Vigna (rows 4 and 5 of Table 2), we have not used the entire collections available from [**? ?**], but only parts of them. One reason is that some such epistles are extremely short in length (sometimes even a single sentence), and hence they would not have conveyed much information to the training process. The second reason is that, as can be seen in Table 2, Guido Faba and Pietro della Vigna are the two authors for whom we have the highest number of epistles anyway, and including the collections in their entirety would have made the dataset even more imbalanced than it already is.

MedLatinLit contains instead (given the similar nature of Ep13(II)) texts of a non-epistolary nature, especially exegetic comments on literary works and treatises, also dating to the 13th and 14th centuries, for a total of 30 texts; the average length of these texts is 39,958 words, i.e., about 100 times longer (on average) than those of MedLatinEpi. Some of these texts are not included in their entirety. In these cases, the portions excluded mainly consist of lengthy *explicit* citations to other authors' works; as already mentioned in Section 2, we have removed explicit citations since they provide noise, rather than information, to an authorship analyser.

All of the texts included in the two datasets are such that their authorship is certain, i.e., is not currently disputed by any scholar.[5] Some of the texts were already available in .txt format, and their inclusion in the dataset has thus posed no major problem. Some other texts were only available in .pdf format, or only on paper; in these cases, we converted the .pdf or the scanned images into .txt format via an optical character recognition software[6], and thoroughly corrected the output by hand.

We have subjected all texts to a number of preprocessing steps necessary for performing accurate authorship analysis; these include

- Removing any meta-textual information that has been inserted by the curator of the edition, such as titles, page numbers, quotation marks, square brackets, etc; this cleans the documents from obvious editorial intervention.
- Marking explicit citations in Latin with asterisks, and explicit citations in languages other than Latin (mostly Florentine vernacular) with curly brackets; this is both to allow ignoring them in the computation (since they are the production of someone different than the author of the text) or to use them as a potential authorial-related feature (i.e., the usage of citations in different languages), at the discretion of the researcher.
- Replacing every occurrence of the character "v" with the character "u"; the reason for this lies in the different approaches followed by the various editors of the texts included, regarding whether to consider "u" and "v" as the same character or not.[7]

The two datasets are available for download at https://doi.org/10.5281/zenodo.4298503; a `readme` file is also included that explains the structure of the archive.[8]

4 BASELINE AUTHORSHIP VERIFICATION RESULTS

In [5] we briefly describe some authorship verification experiments that we have run on MedLatinEpi and MedLatinLit. For the present paper we have rerun the experiments completely, revising and correcting the experimental protocol that we had followed in [5].[9] As a consequence, there are slight differences between the

[5] Note that from Petrus de Boateriis' collection (see last row of Table 2) we have removed the epistle allegedly written by Cangrande della Scala to Henry VII, since it has recently been suggested (see Footnote 2 and Section 5) that it may have been written by Dante Alighieri.

[6] FreeOCR, available at http://www.paperfile.net/ .

[7] In medieval written Latin there was only one grapheme, represented as a lowercase "u" and a capital "V", instead of the two modern graphemes "u-U" and "v-V".

[8] Zenodo is an open-access repository that provides free and permanent access to the resources stored on it; see https://about.zenodo.org/.

[9] In [5] we had performed both feature selection and parameter optimisation on the entire dataset, and we had subsequently estimated the accuracy of the system by applying the leave-one-out protocol. This means that, when a document was used as the test document, it had

Table 2. Composition of the MedLatinEpi dataset; the 3rd column indicates the approximate historical period in which the texts were written, the 4th and 5th columns indicate the number of texts and the number of words that the collection consists of, while the 7th and 8th columns indicate the F_1 value and the *Acc* ("vanilla accuracy") value obtained in the experiments of Section 4 by the authorship verifier trained via logistic regression for the specified author.

Author	Text (or collection thereof)	Period (approx.)	#d	#w	Ed.	F_1	*Acc*
Clara Assisiensis	*Epistola ad Ermentrudem*	1240-1253	1	249	[?]	1.000	1.000
	Epistolae ad sanctam Agnetem de Praga I, II, III	1234-1253	3	1,842	[?]		
Dante Alighieri	Epistles	1304-1315	12	6,061	[?]	0.857	0.990
Giovanni Boccaccio	Epistles and letters	1340-1375	24	25,789	[?]	0.980	0.997
Guido Faba	Epistles	1239-1241	78	7,203	[?]	0.946	0.973
Pietro della Vigna	The collected epistles of Pietro della Vigna	1220-1249	146	65,004	[?]	0.986	0.986
(Various authors)	Epistles from the collection of Petrus de Boateriis	1250-1315	30	5,056	[?]	—	—

Table 3. Composition of the MedLatinLit dataset; the meanings of the columns are as in Table 2.

Author	Text	Period	#w	Ed.	F_1	*Acc*
Bene Florentinus	*Candelabrum*	1238	41,078	[?]	—	—
Benvenuto da Imola	*Comentum super Dantis Aldigherij Comoediam*	1375-1380	105,096	[?]	0.800	0.967
	Expositio super Valerio Maximo	1380	3,419	[?]		
	Glose Bucolicorum Virgilii	1380	3,912	[?]		
Boncompagno da Signa	*Liber de obsidione Ancone*	1198-1200	7,821	[?]	0.333	0.867
	Palma	1198	5,022	[?]		
	Rota Veneris	ante 1215	4,632	[?]		
	Ysagoge	1204	8,550	[?]		
Dante Alighieri	*De Vulgari Eloquentia*	1304–1306	11,384	[?]	0.500	0.933
	Monarchia	1313–1319	19,162	[?]		
Filippo Villani	*Expositio seu comentum super Comedia Dantis Allegherii*	1391-1405	31,503	[?]	—	—
Giovanni Boccaccio	*De vita et moribus d. Francisci Petracchi*	1342	1,884	[?]	0.800	0.967
	De mulieribus claris	1361-1362	49,242	[?]		
	De Genealogia deorum gentilium	1360-1375	198,508	[?]		
Giovanni del Virgilio	*Allegorie super fabulas Ovidii Methamorphoseos*	1320	25,131	[?]	0.000	0.933
	Ars dictaminis	1320	2,376	[?]		
Graziolo Bambaglioli	A Commentary on Dante's Inferno	1324	41,104	[?]	—	—
Guido da Pisa	*Expositiones et glose. Declaratio super Comediam Dantis*	1327-1328	87,822	[?]	—	—
Guido de Columnis	*Historia destructionis Troiae*	1272-1287	82,753	[?]	—	—
Guido Faba	*Dictamina rhetorica*	1226-1228	16,982	[?]	—	—
Iacobus de Varagine	*Chronica civitatis Ianuensis*	1295-1298	53,864	[?]	—	—
Iohannes de Appia	*Constitutiones Romandiolae*	1283	4,068	[?]	—	—
Iohannes de Plano Carpini	*Historia Mongalorum*	1247-1252	20,145	[?]	—	—
Iulianus de Spira	*Vita Sancti Francisci*	1232-1239	12,396	[?]	—	—
Nicola Trevet	*Expositio Herculis Furentis*	1315-1316	33,017	[?]	1.000	1.000
	Expositio L. Annaei Senecae Agamemnonis	1315-1316	19,873	[?]		
Pietro Alighieri	*Comentum super poema Comedie Dantis*	1340-1364	186,608	[?]	—	—
Ryccardus de Sancto Germano	*Chronicon*	1216-1243	36,525	[?]	—	—
Raimundus Lullus	*Ars amativa boni*	1290	82,733	[?]	—	—
Zono de' Magnalis	Life of Virgilio	1340	2,136	[?]	—	—

accuracy values reported in [5] and those reported here. In order to ease the task of researchers wishing to replicate and/or to outperform the results we have obtained, we make available at https://doi.org/10.5281/zenodo.3903235

already participated both in the feature selection process and in the parameter optimisation process, which are parts of the training process; this is not legitimate. Thanks to two anonymous reviewers for pointing this out.

the source code of MedieValla, the authorship verification tool that we have developed and used in order to obtain these results.

For these experiments, first of all we remove explicit citations, either in Latin or other languages, and we segment each resulting text into shorter texts, so as to increase the overall number of labelled texts, while reducing their average size. This is necessary because machine learning processes require a significant number of training examples, regardless of their length. In particular, for each text:

- we identify the sentences that make up the text (using the NLTK package, available at https://www.nltk.org/); if a sentence is shorter than 8 words, we merge it with the next sentence (or the previous sentence, if it is the last sentence of the text);
- we create sequences of 3 consecutive sentences (hereafter: "segments"), consider each of these sequences as a labelled text, and assign it the author label of the text from which it was extracted.

Following this process, we use as labelled texts both the original texts in their entirety *and* the segments. Thus, the number of labelled texts has increased from 294 to 1,310 for MedLatinEpi and from 30 to 12,772 for MedLatinLit.

For our experiments, we lower-case the entire text, remove punctuation marks, and convert each labelled text into a vector of features. The reason why we ignore punctuation marks is that they were not inserted by the authors (punctuation was absent or hardly coherent in ancient manuscripts, and such marks have been introduced into texts by editors).

The set of features we use is subdivided into six subsets of different feature types:

(1) Character n-grams ($n \in \{3, 4, 5\}$);
(2) Word n-grams ($n \in \{1, 2\}$);
(3) Function words (from a list of 74 Latin function words);
(4) Verbal endings (from a list of 245 regular Latin verbal endings);
(5) Word lengths (from 1 to 23 characters);
(6) Sentence lengths (from 3 to 70 words).

and the vector space results from the union of all of these features. In order to deal with the high dimensionality of the feature space we subject the features resulting in a sparse distribution (character n-grams and word n-grams) to a process of dimensionality reduction. First, we perform feature selection via the Chi-square function (see e.g., [29]), where probabilities are interpreted on the event space of documents; in other words, $\Pr(t_k, a_j)$ represents the probability that, for a random document that belongs to class a_j (i.e., that was written by author a_j), feature t_k appears in the document. In our experiments we select the best 10% character n-grams and the best 10% word n-grams. We then perform feature weighting via the tfidf function in its standard "ltc" variant (see e.g., [20]). For MedLatinEpi the number of resulting features is 16,101, while for MedLatinLit this number is instead 86,924.

The six subsets of features described above have very different cardinality: the numbers of features contained in sets (1) and (2) depend on the dataset, but is in general very high (in both cases it typically ranges in the tens – or hundred – thousands features), while the numbers of features contained in sets (3), (4), (5), (6) are fixed (there are 74, 245, 23, 68, features in each of these groups, respectively), and are much smaller than the two previous ones. This means that the latter groups may end up being overwhelmed, in terms of their contribution to the verification process, by the former groups. In order to avoid this, we individually normalise each of the six feature subsets via L2-normalisation, so that each of the six vectors subspaces they define have unit norm.[10]

[10] This means that the contribution of, say, a character n-gram, ends up being smaller than the contribution of, say, a word length, because there are more character n-grams than word lengths. This does not prevent the classifier from uncovering which among the features are the most important (these might well include some character n-grams) or least important (these might well include some word lengths), though, since the classifier attempts to find the linear combination of feature weights that best classifies the documents.

As the learning mechanism we use *logistic regression*, as implemented in the `scikit-learn` package.[11] We train each binary classifier by optimising hyperparameter C (the inverse of the regularisation strength) via stratified 10-fold cross-validation (10-FCV), using a grid search on the set $\{0.001, 0.01, \ldots, 100, 1000\}$. We use a variant of stratified 10-FCV called "grouped" stratified 10-FCV, that prevents different segments from the same document ("group") to end up in different folds; in this way, the classifier never unduly benefits from testing on segments of a document, other segments of which have been seen during training. There are two main reasons why we have used logistic regression. One is the fact that it generates classifiers that have proven very effective across a broad spectrum of text classification scenarios. A second reason is the fact that, together with a binary classification decision, for each document d it returns a "confidence score" (i.e., a measure of the confidence that the classifier has in the correctness of its own decision) in the form of a probability value (called a "posterior probability"), and that these probability values tend to be *well calibrated* (i.e., reliable probability values) [16].

We also briefly report on some additional experiments for which we have used other learning algorithms, i.e., SVMs (for which we have optimised hyperparameter C via grid search on $\{0.001, 0.01, \ldots, 100, 1000\}$) and multinomial naive Bayes (for which we have optimised parameter $\alpha = \{10^i\}$ for $i \in \{-7, -6, \ldots, -1, 0\}$).

We subject the resulting MedieValla system to a "leave-one-out" validation test, which consists of predicting, for each dataset $D \in \{$MedLatinEpi, MedLatinLit$\}$, for each author a in the set of authors $\mathcal{A}$ represented in D, and for each document $d \in D$, whether a is the author of d or not, where the prediction is issued by an "a vs. (NOT a)" binary classifier trained on all labelled texts (i.e., segments *and* entire documents) from $D \setminus \{d\}$. This means that all labelled texts from documents in $D \setminus \{d\}$ originating from author a are used as positive training examples while all labelled texts from documents in $D \setminus \{d\}$ originating from authors other than a are used as negative training examples. Note that

- In order to faithfully reproduce the operating conditions of an authorship verifier, as test examples we use only entire documents, i.e., we use segments and entire documents for training purposes but only entire documents for testing purposes.
- In order to avoid any overlap between training examples and test examples, when document d is used as a test document we exclude from the training set all the segments derived from d.
- In order to avoid any overlap between the training phase and the test phase, both the feature selection step and the parameter optimisation step are performed not on the entire dataset D, but on $D \setminus \{d\}$. This means that the entire cycle (feature selection + parameter optimisation + classifier training) is repeated for each document $d \in D$, for both MedLatinEpi and MedLatinLit.
- We have not generated classifiers for authors for which we have only one text in D, since this would entail experiments in which the author is not present both in the training and in the test set;[12] as a result, the texts of these authors are used only as negative examples in experiments centred on other authors. Ultimately, this means that we have trained binary classifiers for 5 authors of MedLatinEpi (all authors except those from the collection of Petrus de Boateriis, since this collection is a miscellanea of authors) and 6 authors for MedLatinLit; this leads to 5×294=1470 predictions for MedLatinEpi and 6×30=180 predictions for

[11] https://scikit-learn.org/stable/index.html

[12] For the very same reason, we bypass the parameter optimisation phase in cases in which we only have 2 positive documents and one of them is acting as the held-out document. This causes the training set to have only one positive document (plus fragments) and this eventually forces one of the trainings (as generated via 10-fold cross-validation) to be devoid of any positive example (since in the "grouped" variant of stratified 10-FCV the fragments of a document are always within the same fold as the full document, for reasons already discussed). In those (few) cases, we resort to a logistic regressor that is moderately regularised (we set $C = 0.1$) in order to avoid overfitting the one and only positive document; likewise, for SVMs we also set $C = 0.1$ and for multinomial naive Bayes we set $\alpha = 0.001$.

Table 4. Summary results of our AV experiments on the MedLatinEpi and MedLatinLit datasets.

	MedLatinEpi			MedLatinLit		
Learner	F_1^M	F_1^μ	*Acc*	F_1^M	F_1^μ	*Acc*
LR	0.954	0.969	0.989	0.572	0.615	0.944
SVM	0.944	0.969	0.989	0.383	0.435	0.928
MNB	0.760	0.933	0.976	0.310	0.357	0.900

MedLatinLit, where each prediction is the result of a different cycle consisting of feature selection + parameter optimisation + classifier training.[13]

In order to evaluate the performance of a binary AV system we use, as customary, the F_1 function, defined as

$$F_1 = \begin{cases} \dfrac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} & \text{if TP} + \text{FP} + \text{FN} > 0 \\ 1 & \text{if TP} = \text{FP} = \text{FN} = 0 \end{cases} \tag{1}$$

where TP, FP, FN, represent the numbers of true positives, false positives, false negatives, generated by the binary AV system. F_1 ranges between 0 (worst) and 1 (best). In order to compute F_1 across an entire dataset, for which several binary AV systems need to be deployed (5 for MedLatinEpi and 6 for MedLatinLit), we compute its *macroaveraged* variant (denoted by F_1^M) and its *microaveraged* variant (denoted by F_1^μ). F_1^M is obtained by first computing values of F_1 for all $a_j \in \mathcal{A}$ and then averaging them. F_1^μ is obtained by (a) computing the author-specific values TP_j, FP_j, FN_j for all $a_j \in \mathcal{A}$; (b) obtaining TP as the sum of the TP_j's (same for FP and FN), and then (c) applying Equation 1. For completeness we also report effectiveness results in terms of the so-called "vanilla accuracy" measure, defined as

$$Acc = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{2}$$

i.e., as the ratio between the number of correct predictions and the number of predictions. In order to compute *Acc* across different binary AV systems, either the microaveraged or the macroaveraged version of *Acc* can be computed, along the same lines as for F_1. Unlike for F_1, though, the microaveraged and the macroaveraged versions of *Acc* are demonstrably the same measure, which we will thus simply indicate as *Acc*, without μ or M superscripts.

Our experimental results are reported in Table 4.[14] The last columns of Tables 2 and 3 report the F_1 and *Acc* values we have obtained for the individual authors for which we have generated binary AV systems; from these it is easy to compute the F_1^M values and average *Acc* values of Table 4 by simply averaging them.

[13] Since we use 10-fold cross validation for parameter optimisation and explore a grid of 7 parameters, or experimentation consists of roughly 115,000 trainings per learner (we report experiments for 3 learners).

[14] Two further reasons why these results slightly differ from the ones reported in [5] are that (a) some `scikit-learn` libraries that we use are now available in updated versions, different from the ones we had used in [5]; (b) the stratified 10-fold cross-validation that we use for optimizing hyperparameter C splits the data into 10 folds randomly, and this random component can introduce small fluctuations in the final results. Overall, these fluctuations are noticeable but not substantial from a qualitative point of view. The results we report in this paper should be exactly reproducible (barring changes in `scikit-learn` libraries) by anyone who downloads the code and the datasets, also thanks to the fact that we have now "seeded" the stratified 10-fold cross-validation process, thus eliminating the above-mentioned random component.

Note that, as evident from the F_1 and *Acc* columns of Tables 2 and 3, there is a lot of variability in the scores (especially for F_1) across different authors for the same dataset. There are at least three possible explanations for this:

- For some authors there are more (positive) training data than for other authors. Since authorship verification consists of a different binary classification task for each author, this means that it will be easier (other things being equal) to conduct authorship verification for the former authors than for the latter.
- Some large differences in F_1 values are due to the idiosyncrasies of the F_1 measure. For instance, the authorship verifier for Giovanni del Virgilio (see Table 3), when asked to verify the 30 texts in MedLatinLit, returns 2 false negatives and 28 true negatives. Despite having correctly predicted 28 out of 30 times (the "vanilla accuracy" result is Acc=28/30=0.933), the verifier obtains an F_1 value of 0 because (see Equation 1) there are no true positives, i.e., none of the two texts actually by Giovanni del Virgilio were correctly predicted as by him.
- Even if we had the same quantity of training data for each author, we might obtain different accuracy results for different authors because some authors may inherently be more difficult to identify, from a stylistic point of view, than others.

At https://doi.org/10.5281/zenodo.4298503 we provide, in spreadsheet form, the list of all ⟨author, document⟩ classification decisions as taken by MedieValla, as well as the F_1^{μ} results that are also reported in Table 4 and the author-specific F_1 values also reported in Tables 2 and 3.

Interestingly enough, an analysis of these individual classification decisions shows that *there are no systematic mistakes*, but just a few, scattered individual ones. More in particular, it never happens that there are two or more incorrectly classified documents with the same true author A_1 *and* with the same predicted author A_2, with $A_1 \neq A_2$; in other words, there are no systematic mistakes that would indicate an extreme similarity in style between two authors A_1 and A_2. One of the reasons for this is that the mistakes made by our verifiers are very few, i.e., only 26 out of 1650 verification decisions (16 out of 1470 for the MedLatinEpi experiments and 10 out of 180 for the MedLatinLit experiments) are incorrect.

5 TWO DISPUTED EPISTLES

5.1 The Epistle to Cangrande

In Section 3.1 we mentioned that the original reason for developing these two datasets was the attempt to solve the puzzle of the *Epistle to Cangrande*, i.e., verifying if the letter addressed to Cangrande della Scala was indeed written by Dante Alighieri. After running the experiments described in Section 4, for each of the two datasets we have retrained the authorship verifier for author Dante Alighieri (i.e., the one that whose Yes label indicates authorship by Dante and whose No label indicates authorship by someone other than Dante), rerunning the entire cycle "feature selection + parameter optimisation + classifier training" on the entire dataset; we have then applied the classifier derived from MedLatinEpi to the first portion of the epistle (Ep13(I)) and the classifier derived from MedLatinLit to the second portion (Ep13(II)).

The results of the application of the two classifiers are reported in Table 5. These results show that out authorship verifiers believe that both portions of the Epistle to Cangrande are the work of a malicious forger.

Once applied to Ep13(I), the "Dante vs. Not Dante" verifier trained on MedLatinEpi returns a posterior probability of 0.367: this means that the verifier believes that Ep13(I) is not by Dante (since this probability is <0.500), and is moderately confident about this fact (its "degree of confidence" being (1-0.367)=0.633). As from Table 2, this verifier has also proved very accurate (F_1 = 0.857, Acc = 0.990) once tested on MedLatinEpi via

Table 5. Results of the application of the two authorship verifiers “Dante vs. Not Dante” to the two portions of the Epistle to Cangrande. Columns 4 and 5 recall (from Tables 2 and 3) the F_1 and *Acc* values that the “Dante vs. Not Dante” verifiers have obtained in the experiments of Section 4.

	Binary decision	Posterior probability	F_1	*Acc*
Ep13(I)	No	0.367	0.857	0.990
Ep13(II)	No	0.022	0.500	0.933

Table 6. Result of the application of the authorship verifier “Dante vs. Not Dante” to the Epistle to Henry VII. Column 4 recalls (from Table 2) the F_1 value that the “Dante vs. Not Dante” verifier has obtained in the experiments of Section 4.

	Binary decision	Posterior probability	F_1	*Acc*
EpHenryVII	No	0.026	0.857	0.990

leave-one-out. These two facts, altogether, make a fairly convincing case for the non-Dantean authorship of Ep13(I).

Concerning Ep13(II), instead, once applied to it, the “Dante vs. Not Dante” verifier trained on MedLatinLit returns a posterior probability of 0.022: this means that the verifier believes that Ep13(II) is also not by Dante (since this probability is <0.500), and is extremely confident about this fact (its degree of confidence being (1-0.022)=0.978). As from Table 3, this verifier has proved reasonably accurate (F_1 = 0.500, *Acc* = 0.933) once tested on MedLatinLit via leave-one-out. These two facts support the hypothesis that also Ep13(I) is not by Dante.[15]

5.2 The Epistle to Henry VII

While we were carrying out our research on Ep13 that led to the creation of MedLatinEpi and MedLatinLit, a paper appeared [17] whose object was an epistle addressed to emperor Henry VII and signed by Cangrande della Scala. The author of [17], based on an analysis of the contents of the epistle, conjectured that its author could be Dante Alighieri himself. Since we had already trained a “Dante vs. not Dante” authorship verifier on MedLatinEpi, and since the texts contained in MedLatinEpi have also an epistular nature, it seemed natural to preprocess the epistle to Henry VII in the same way as described in Section 4, and apply to it the verifier trained on MedLatinEpi. The results of the application are described in Table 6.

Our authorship verifier rejects the hypothesis that the epistle to Henry VII may have been written by Dante, and is extremely confident in its own prediction (i.e., it believes that the epistle is by someone other than Dante with probability (1-0.026)=0.974). Together with the fact that this verifier has shown very high accuracy (F_1 = 0.857, *Acc* = 0.990) in the experiments of Section 4, this makes us decidedly lean towards the hypothesis that the epistle is not the work of Dante.

[15] Note that a classifier that obtains F_1 = 0.500 is *not* equivalent to a classifier that returns random decisions: in fact, a completely clueless classifier for which half of the positives are true positives while the other half are false negatives, and half of the negatives are true negatives while the other half are false positives, on MedLatinLit would obtain a value of $F_1 = (2\mathrm{TP})/(2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}) = (2 \cdot 1)/(2 \cdot 1 + 14 + 1) = 0.058$. The F_1 = 0.500 result for the “Dante vs. Not Dante” authorship verifier is the result of generating, on dataset MedLatinLit, 1 true positive, 1 false positive, 1 false negative, and 27 true negatives, i.e., 28 correct predictions out of 30 total predictions.

6 CONCLUSION

We have described MedLatinEpi and MedLatinLit, two new datasets of cultural heritage texts written in medieval Latin by 13th- and 14th-century (mostly Italian) literates and labelled by author, that we make publicly available to researchers working on computational authorship analysis. These datasets can be valuable tools for researchers investigating techniques for authorship attribution, authorship verification, or same-authorship verification, especially for texts written in Latin or medieval Latin.

We also make available the source code of MedieValla, an authorship verification tool that we have built in order to work on an important case study, i.e., the real paternity of the "Epistle to Cangrande", allegedly written by Dante Alighieri but believed by some to be a forgery. We also describe in detail experiments (corrected versions of the ones which we had reported in [5]) in which we have applied MedieValla to MedLatinEpi and MedLatinLit. We hope that the availability of the datasets (and of our authorship verification tool) will allow researchers interested in authorship verification to replicate our results, and possibly to outperform them via improved AV techniques.

ACKNOWLEDGMENTS

We would like to thank Gabriella Albanese and Paolo Pontari for helping us to identify the medieval Latin texts that we have incorporated into our datasets; Patrick Juola, Moshe Koppel, Vincenzo Mele, and Efstathios Stamatatos, for suggesting important bibliographical references; and Carlo Meghini for giving the initial impetus to this research and for stimulating discussions on the topics covered by this article. Thanks to the anonymous reviewers for useful feedback and for spotting two problems in the experimental protocol used in [5].

REFERENCES

[1] Charu C. Aggarwal and ChengXiang Zhai. 2012. A survey of text classification algorithms. In *Mining Text Data*, Charu C. Aggarwal and ChengXiang Zhai (Eds.). Springer, Heidelberg, DE, 163–222.

[2] Alberto Casadei (Ed.). 2020. *Atti del seminario "nuove inchieste sull'epistola a Cangrande".* Pisa University Press, Pisa, IT.

[3] Carole E. Chaski. 2005. Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence* 4, 1 (2005).

[4] Silvia Corbara. 2019. *The Epistle to Cangrande through the lens of computational authorship verification.* Master's thesis. Department of Philology, Literature, and Linguistics, University of Pisa, Pisa, IT.

[5] Silvia Corbara, Alejandro Moreo, Fabrizio Sebastiani, and Mirko Tavoni. 2019. The Epistle to Cangrande through the lens of computational authorship verification. In *Proceedings of the 1st International Workshop on Pattern Recognition for Cultural Heritage (PatReCH 2019).* Trento, IT, 148–158. https://doi.org/10.1007/978-3-030-30754-7_15

[6] Christopher W. Forstall, Sarah L. Jacobson, and Walter J. Scheirer. 2011. Evidence of intertextuality: Investigating Paul the Deacon's Angustae Vitae. *Literary and Linguistic Computing* 26, 3 (2011), 285–296.

[7] Patrick Juola. 2006. Authorship attribution. *Foundations and Trends in Information Retrieval* 1, 3 (2006), 233–334. https://doi.org/10.1561/1500000005

[8] Jakub Kabala. 2020. Computational authorship attribution in medieval Latin corpora: The case of the Monk of Lido (ca. 1101–08) and Gallus Anonymous (ca. 1113–17). *Language Resources and Evaluation* 54, 1 (2020), 25–56. https://doi.org/10.1007/s10579-018-9424-0

[9] Mike Kestemont, Sara Moens, and Jeroen Deploige. 2015. Collaborative authorship in the twelfth century: A stylometric study of Hildegard of Bingen and Guibert of Gembloux. *Digital Scholarship in the Humanities* 30, 2 (2015), 199–224. https://doi.org/10.1093/llc/fqt063

[10] Mike Kestemont, Justin A. Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. 2016. Authenticating the writings of Julius Caesar. *Expert Systems with Applications* 63 (2016), 86–96. https://doi.org/10.1016/j.eswa.2016.06.029

[11] Moshe Koppel, Shlomo Argamon, and Anat R. Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing* 17, 4 (2002), 401–412. https://doi.org/10.1093/llc/17.4.401

[12] Moshe Koppel and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. In *Proceedings of the 21st International Conference on Machine Learning (ICML 2004).* Banff, CA. https://doi.org/10.1145/1015330.1015448

[13] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology* 60, 1 (2009), 9–26. https://doi.org/10.1002/asi.20961

[14] Moshe Koppel and Yaron Winter. 2014. Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology* 65, 1 (2014), 178–187. https://doi.org/10.1002/asi.22954

[15] Samuel Larner. 2014. *Forensic authorship analysis and the World Wide Web*. Springer, Heidelberg, DE.

[16] Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML 2005)*. Bonn, DE, 625–632. https://doi.org/10.1145/1102351.1102430

[17] Paolo Pellegrini. 2018. La quattordicesima epistola di dante alighieri: Primi appunti per una attribuzione. *Studi di Erudizione e di Filologia Italiana* 7 (2018), 5–20.

[18] Ria Perkins. 2015. Native language identification (NLID) for forensic authorship analysis of weblogs. In *New Threats and Countermeasures in Digital Crime and Cyber Terrorism*, Maurice Dawson and Marwan Omar (Eds.). IGI Global, Hershey, US, 213–234. https://doi.org/0.4018/978-1-4666-8345-7.ch012

[19] Anderson Rocha, Walter J. Scheirer, Christopher W. Forstall, Thiago Cavalcante, Antonio Theophilo, Bingyu Shen, Ariadne Carvalho, and Efstathios Stamatatos. 2017. Authorship attribution for social media forensics. *IEEE Transactions on Information Forensics and Security* 12, 1 (2017), 5–33. https://doi.org/10.1109/TIFS.2016.2603960

[20] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management* 24, 5 (1988), 513–523.

[21] Jacques Savoy. 2019. Authorship of Pauline epistles revisited. *Journal of the Association for Information Science and Technology* 70, 10 (2019), 1089–1097. https://doi.org/10.1002/asi.24176

[22] Michael R. Schmid, Farkhund Iqbal, and Benjamin C. M. Fung. 2015. E-mail authorship attribution using customized associative classification. *Digital Investigation* 14, 1 (2015), S116–S126. https://doi.org/10.1016/j.diin.2015.05.012

[23] Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60, 3 (2009), 538–556. https://doi.org/10.1002/asi.21001

[24] Efstathios Stamatatos. 2016. Authorship verification: A review of recent advances. *Research in Computing Science* 123 (2016), 9–25.

[25] Justin A. Stover, Yaron Winter, Moshe Koppel, and Mike Kestemont. 2016. Computational authorship verification method attributes a new work to a major 2nd century African author. *Journal of the American Society for Information Science and Technology* 67, 1 (2016), 239–242. https://doi.org/10.1002/asi.23460

[26] Paget Toynbee. 1918. Dante and the "cursus": A new argument in favour of the authenticity of the Quaestio de Aqua et Terra. *The Modern Language Review* 13, 4 (1918), 420–430.

[27] Enrico Tuccinardi. 2017. An application of a profile-based method for authorship verification: Investigating the authenticity of Pliny the Younger's letter to Trajan concerning the Christians. *Digital Scholarship in the Humanities* 32, 2 (2017), 435–447. https://doi.org/10.1093/llc/fqw001

[28] Raija Vainio, Reima Välimäki, Anni Hella, Marjo Kaartinen, Teemu Immonen, Aleksi Vesanto, and Filip Ginter. 2019. Reconsidering authorship in the Ciceronian corpus through computational authorship attribution. *Ciceroniana On Line* 3, 1 (2019). https://doi.org/10.13135/2532-5353/3518

[29] Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning (ICML 1997)*. Nashville, US, 412–420.

Received September 2021