MedLatin1 and MedLatin2: 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 MedLatin1 and MedLatin2, two datasets of medieval Latin texts to be used in research on computational authorship analysis. MedLatin1 and MedLatin2 consist of 294 and 30 curated texts, respectively, labelled by author, with MedLatin1 texts being of an epistolary nature and MedLatin2 texts consisting 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.

ACM Reference Format:
Silvia Corbara, Alejandro Moreo, Fabrizio Sebastiani, and Mirko Tavoni. 2020. MedLatin1 and MedLatin2: Two Datasets for the Computational Authorship Analysis of Medieval Latin Texts. 1, 1 (June 2020), 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

(Computational) *Authorship Analysis* is the task of inferring the characteristics of the author of a text of unknown or disputed paternity. 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 [44]), or *native language identification* (i.e., predicting the native language of the author of the text [50]).

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) [41, 46, 54], which consists of predicting who, among a set of k candidate authors, is the real author of the text. A task that has gained prominence more recently is *Authorship Verification* (AV) [45, 55], 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 this field is *Same-Authorship Verification* (SAV) [47], 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 [36], and thus solved with the help of machine learning methods: for instance, an authorship verification task is

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.

XXXX-XXXX/2020/6-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

solved as a *binary classification* problem, i.e., as the problem of classifying the disputed text into one of 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 to prevent crimes committed via digital means) [53], to computational forensics (the field concerned with the study of digital evidence for investigating crimes that have already occurred) [38, 48, 50, 51]. 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, 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 [42, 43, 52, 56, 58].

We here present and make available two datasets of texts of the latter type, i.e., texts written in medieval Latin 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 before 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 paternity 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 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.

2 THE DATASETS

2.1 Origin of the datasets

Our two datasets originated in the context of an authorship verification research work [39, 40] 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 [37]. 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

[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* [37], the *Quaestio de aqua et terra* [57], and *Cangrande's Epistle to Henry VII* [49], just to mention ones that some scholars attribute to Dante Alighieri while some others do not.

[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 emperor Constantine in which he supposedly conferred authority over Rome and the western part of the Roman Empire to the Pope) was a forgery.

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 MedLatin1 and MedLatin2, respectively.

In both MedLatin1 and MedLatin2 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" (in the historical-linguistic sense) 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; since the included authors are close to each other, in the above-mentioned historical-linguistic sense, the resulting dataset is a challenging one for computational authorship analysis systems.

While we used MedLatin1 and MedLatin2 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, or as datasets for other authorship analysis tasks that address medieval Latin. This is the reason why we make them available to the research community.

2.2 Composition and preprocessing of the datasets

The composition of our two datasets is listed in detail in Tables 1 and 2.

MedLatin1 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. 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 1, we have not used the entire collections available from [10, 17], 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 1, 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 that it actually is.

MedLatin2 contains instead (given the similar nature of Ep13(II)) texts of a non-epistolary nature, especially comments on literary works and treatises, also dating to the 13th and 14th centuries,

[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 vulgar, the language that would later form the basis of the Italian language.

Table 1. Composition of the MedLatin1 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 column indicates the F_1 value obtained in the experiments of Section 3 by the authorship verifier for the specified author.

Author	Text (or collection thereof)	Period (approx.)	#d	#w	Ed.	F_1
Clara Assisiensis	*Epistola ad Ermentrudem*	1240-1253	1	249	[23]	0.571
	Epistolae ad sanctam Agnetem de Praga I, II, III	1234-1253	3	1,842	[23]	
Dante Alighieri	Epistles	1304-1315	12	6,061	[13]	0.957
Giovanni Boccaccio	Epistles and letters	1340-1375	24	25,789	[2]	1.000
Guido Faba	Epistles	1239-1241	78	7,203	[17]	0.980
Pietro della Vigna	The collected epistles of Pietro della Vigna	1220-1249	146	65,004	[10]	0.993
(Various authors)	Epistles from the collection of Petrus de Boateriis	1250-1315	30	5,056	[30]	—

for a total of 30 texts; the average length of these texts is 39,958 words. Some of these texts are not included in their entirety; in these cases, the portions excluded mainly consist of lengthy *explicit* citations, i.e., excerpts from other authors' works.

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 OCR software and manually corrected the output where necessary.

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.[6]

The two datasets are available for download at https://doi.org/10.5281/zenodo.3903296; a `readme` file is also included that explains the structure of the archive.

3 BASELINE AUTHORSHIP VERIFICATION RESULTS

In [40] we briefly describe some authorship verification experiments that we have run on MedLatin1 and MedLatin2. In order to ease the task of researchers wishing to replicate and/or to outperform the results we obtained, via some improved authorship verification techniques, we here repeat, in a more detailed way, the description of those experiments, and we make available

[5]Note that from Petrus de Boateriis' collection (see last row of Table 1) we have removed the epistle allegedly written by Cangrande della Scala to Henry VII, since it has recently been suggested (see Footnote 2) that it may have been written by Dante Alighieri.

[6]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".

Table 2. Composition of the MedLatin2 dataset; the meanings of the columns are as in Table 1.

Author	Text	Period	#w	Ed.	F_1
Bene Florentinus	*Candelabrum*	1238	41,078	[1]	–
Benvenuto da Imola	*Comentum super Dantis Aldigherij Comoediam*	1375-1380	105,096	[4]	1.000
	Expositio super Valerio Maximo	1380	3,419	[29]	
	Glose Bucolicorum Virgilii	1380	3,912	[21]	
Boncompagno da Signa	*Liber de obsidione Ancone*	1198-1200	7,821	[15]	0.571
	Palma	1198	5,022	[32]	
	Rota Veneris	ante 1215	4,632	[14]	
	Ysagoge	1204	8,550	[7]	
Dante Alighieri	*De Vulgari Eloquentia*	1304–1306	11,384	[33]	0.500
	Monarchia	1313–1319	19,162	[25]	
Filippo Villani	*Expositio seu comentum super Comedia Dantis Allegherii*	1391-1405	31,503	[12]	–
Giovanni Boccaccio	*De vita et moribus d. Francisci Petracchi*	1342	1,884	[11]	1.000
	De mulieribus claris	1361-1362	49,242	[35]	
	De Genealogia deorum gentilium	1360-1375	198,508	[27]	
Giovanni del Virgilio	*Allegorie super fabulas Ovidii Methamorphoseos*	1320	25,131	[8]	0.000
	Ars dictaminis	1320	2,376	[20]	
Graziolo Bambaglioli	*Commento all'Inferno di Dante*	1324	41,104	[28]	–
Guido da Pisa	*Expositiones et glose. Declaratio super Comediam Dantis*	1327-1328	87,822	[6]	–
Guido de Columnis	*Historia destructionis Troiae*	1272-1287	82,753	[19]	–
Guido Faba	*Dictamina rhetorica*	1226-1228	16,982	[18]	–
Iacobus de Varagine	*Chronica civitatis Ianuensis*	1295-1298	53,864	[24]	–
Iohannes de Appia	*Constitutiones Romandiolae*	1283	4,068	[3]	–
Iohannes de Plano Carpini	*Historia Mongalorum*	1247-1252	20,145	[9]	–
Iulianus de Spira	*Vita Sancti Francisci*	1232-1239	12,396	[23]	–
Nicola Trevet	*Expositio Herculis Furentis*	1315-1316	33,017	[34]	1.000
	Expositio L. Annaei Senecae Agamemnonis	1315-1316	19,873	[22]	
Pietro Alighieri	*Comentum super poema Comedie Dantis*	1340-1364	186,608	[5]	–
Ryccardus de Sancto Germano	*Chronicon*	1216-1243	36,525	[16]	–
Raimundus Lullus	*Ars amativa boni*	1290	82,733	[26]	–
Zono de' Magnalis	*Vita di Virgilio*	1340	2,136	[31]	–

at https://doi.org/10.5281/zenodo.3903236 the source code of MedieValla, the authorship verification tool that we have developed and used in order to obtain those results.

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

- we have identified 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 have merged it with the next sentence (or the previous sentence, if it is the last sentence of the text);
- we have created sequences of 3 consecutive sentences (hereafter: "segments"), considered each of these sequences as a labelled text, and assigned it the author label of the text from where 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 MedLatin1 and from 30 to 12,772 for MedLatin2.

For our experiments, we lower-case the entire text, remove punctuation marks, and convert each labelled text into a vector of features. The set of features we use is the following:

- Character n-grams ($n \in \{3, 4, 5\}$);
- Word n-grams ($n \in \{1, 2\}$);
- Function words (from a list of 74 Latin function words);

- Verbal endings (from a list of 245 regular Latin verbal endings);
- Word lengths (from 1 to 23 characters);
- Sentence lengths (from 3 to 70 words).

The reason why we ignore punctuation marks is that they were not inserted by the authors (punctuation was not used in medieval Latin, and such marks have been introduced into texts by editors).

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, i.e.,

$$\chi^2(t_k, a_j) = \frac{[\Pr(t_k, a_j)\Pr(\bar{t}_k, \bar{a}_j) - \Pr(t_k, \bar{a}_j)\Pr(\bar{t}_k, a_j)]^2}{\Pr(t_k)\Pr(\bar{t}_k)\Pr(a_j)\Pr(\bar{a}_j)} \tag{1}$$

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., written by author a_j), feature t_k appears in the document. In our experiments we have selected the best 10% character n-grams and the best 10% word n-grams. We have then performed feature weighting via the tfidf function in its standard "ltc" variant, i.e.,

$$\text{tfidf}(t_k, d_i) = \text{tf}(t_k, d_i) \cdot \log \frac{|D|}{\#D(t_k)} \tag{2}$$

where $\text{tfidf}(t_k, d_i)$ is the weight of feature t_k for document d_i, D is the dataset, $\#D(t_k)$ is the *document frequency* of feature t_k (i.e., the number of documents in which the feature appears at least once), and

$$\text{tf}(t_k, d_i) = \begin{cases} 1 + \log \#(t_k, d_i) & \text{if } \#(t_k, d_i) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $\#(t_k, d_i)$ is the number of occurrences of feature t_k in document d_i. For MedLatin1 the number of resulting features is 16,101, while for MedLatin2 this number is instead 86,924.

As the learning mechanism we use logistic regression, as implemented in the `scikit-learn` package.[7] We train each binary classifier by optimizing hyperparameter C (the inverse of the regularization strength) via stratified 10-fold cross-validation, using a grid search on the set {0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000}.

We subject the resulting MedieValla system to a "leave-one-out" validation test, which means predicting, for each dataset $D \in \{\text{MedLatin1}, \text{MedLatin2}\}$, 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/d. This means that all labelled texts from documents in D/d originating from author a are used as positive training examples while the ones 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.

However, 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; as a result, the texts of these authors are used only as negative examples in experiments

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

centred on other authors. Ultimately, this means that we have trained binary classifiers for 5 authors of MedLatin1 (all authors except those from the collection of Petrus de Boateriis, since it is a miscellanea by authors represented by 1 text each) and 6 authors for MedLatin2; this leads to 5×294=1470 predictions for MedLatin1 and 6×30=180 predictions for MedLatin2.

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

$$F_1 = \begin{cases} \dfrac{2TP}{2TP + FP + FN} & \text{if } TP + FP + FN > 0 \\ 1 & \text{if } TP = FP = FN = 0 \end{cases} \quad (4)$$

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 MedLatin1 and 6 for MedLatin2), 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 4.

The results are reported in the following table[8]:

MedLatin1		MedLatin2	
F_1^M	F_1^μ	F_1^M	F_1^μ
0.900	0.983	0.679	0.759

The last columns of Tables 1 and 2 report the F_1 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 in the previous table.

At http://hlt.isti.cnr.it/medlatin/Results.xls we provide, in spreadsheet form, the list of all ⟨author, document⟩ pairs for which MedieValla has returned an incorrect decision; from these it is easy to compute the F_1^μ results in the previous table, as well as the author-specific F_1 values of Tables 1 and 2.

4 CONCLUSION

We have described MedLatin1 and MedLatin2, two new datasets of literary texts written in medieval Latin by 13th- and 14th-century 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, for medieval Latin.

We also make available the source code of MedieValla, an authorship verification tool for medieval Latin, and we describe in detail experiments (which we had already reported in [40]) in which we have applied MedieValla to MedLatin1 and MedLatin2; we hope that this will

[8]These results slightly differ from the ones reported in [40]. This is due to two factors: (a) some `scikit-learn` libraries that we use are now available in updated versions, different from the ones we used in [40]; (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. We have reported the results of the experiments we have rerun at the time of submitting this paper, rather than those reported in [40], since the former should be exactly reproducible (barring changes in `scikit-learn` libraries) by anyone who downloads the code and the datasets. (We have now "seeded" the stratified 10-fold cross-validation process, thus eliminating the above-mentioned random component.)

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.

PRIMARY SOURCES

[1] Gian Carlo Alessio. 1983. *Bene Florentini Candelabrum.* Editrice Antenore, Padova, IT. https://bit.ly/2Xbl0pR (Archivio della Latinità Italiana del Medioevo), accessed 2018-05-28.

[2] Ginetta Auzzas. 1992. *Tutte le opere di Giovanni Boccaccio.* Mondadori, Milano, IT, Chapter "Epistole e lettere". https://bit.ly/2I2VLjp (Biblioteca Italiana), accessed 2018-05-28.

[3] Luigi Colini Baldeschi. 1925-1926. Le "Constitutiones Romandiolae" di Giovanni d'Appia. *Nuovi Studi Medievali* 2, 1 (1925-1926), 221–252. https://bit.ly/30RRoAg (Archivio della Latinità Italiana del Medioevo), accessed 2018-05-28.

[4] Kevin Brownlee and Robert Hollander. 2018. Benevenuti de Rambaldis de Imola Comentum super Dantis Aldigherij Comoediam, nunc primum integre in lucem editum sumptibus Guilielmi Warren Vernon, curante Jacobo Philippo Lacaita. Florentiae, G. Barbèra, 1887. https://bit.ly/2Dqj6du (Darmouth Dante Project), accessed 2018-05-28.

[5] Massimiliano Chiamenti. 1999. *I commenti danteschi dei secoli XIV, XV e XVI.* LEXIS Progetti Editoriali, Roma, IT, Chapter "Comentum super Comedie Dantis (terza ed ultima redazione del 'Comentum')". https://bit.ly/2ECcU2c (Biblioteca Italiana), accessed 2018-05-28.

[6] Vincenzo Cioffari. 1974. *Guido da Pisa's Expositiones et Glose super Comediam Dantis, or Commentary on Dante's Inferno.* State University of New York Press, Albany, US. https://bit.ly/2FRhT0x (Darmouth Dante Project), accessed 2018-05-28.

[7] Elmert Clark. 1997. Magistri Boncompagni Ysagoge. *Quadrivium* 8 (1997), 23–71. https://bit.ly/2MedBFc (Archivio della Latinità Italiana del Medioevo), accessed 2018-05-28.

[8] Valeria Cotza. 2013. *Giovanni del Virgilio, Allegorie super fabulas Ovidii Methamorphoseos. Edizione critica e introduzione.* Master's thesis. Department of Philology, Literature and Linguistics, University of Pisa, Pisa, IT.

[9] Paolo Daffinà, Claudio Leonardi, Maria Cristiana Lungarotti, Enrico Menestò, and Luciano Petech. 1989. *Giovanni di Pian di Carpine. Storia dei Mongoli.* Fondazione CISAM, Spoleto, IT. 227–333 pages. https://bit.ly/2WoZaSM (Archivio della Latinità Italiana del Medioevo), accessed 2018-05-28.

[10] Edoardo D'Angelo. 2014. *L'epistolario di Pier della Vigna.* Rubbettino Editore, Soveria Mannelli, IT.

[11] Renata Fabbri. 1992. *Tutte le opere di Giovanni Boccaccio.* Mondadori, Milano, IT, Chapter "De vita et moribus d. Francisci Petracchi". https://bit.ly/2MdBUTV (Biblioteca Italiana), accessed 2018-05-28.

[12] Francesca Ferrario. 1999. Expositio seu comentum super "Comedia" Dantis Allegherii, a cura di Saverio Bellomo. Florence: Le Lettere, 1989. https://bit.ly/2Uac1mS (Darmouth Dante Project), accessed 2018-05-28.

[13] Arsenio Frugoni and Giorgio Brugnoli. 1996. *Dante Alighieri - Opere minori.* Riccardo Ricciardi Editore, Milano, IT, Chapter "Epistole". https://bit.ly/2JIPYTp (Biblioteca Italiana), accessed 2018-05-28.

[14] Paolo Garbini. 1996. *Boncompagnus de Signa - Rota Veneris.* Salerno Editrice, Roma, IT. https://bit.ly/2wrCTVS (Archivio della Latinità Italiana del Medioevo), accessed 2018-05-28.

[15] Paolo Garbini. 1999. *Boncompagnus de Signa - Liber de obsidione Ancone.* Viella, Roma, IT. https://bit.ly/2QuMMLw (Archivio della Latinità Italiana del Medioevo), accessed 2018-05-28.

[16] Carlo Alberto Garufi. 1937. Ryccardi de Sancto Germano notarii Chronica. *Rerum Italicarum Scriptores* 7, 2 (1937). https://bit.ly/2HHrE1W (Archivio della Latinità Italiana del Medioevo), accessed 2018-05-28.

[17] Augusto Gaudenzi. 1971. *Guido Faba - Dictamina Rhetorica Epistole.* Forni Editore, Bologna, IT, Chapter "Guidonis Fabe Epistole". https://bit.ly/2WrxgWh (Archivio della Latinità Italiana del Medioevo), accessed 2018-05-28.

[18] Augusto Gaudenzi. 1971. *Guido Faba - Dictamina Rhetorica Epistole.* Forni Editore, Bologna, IT, Chapter "Guidonis Fabe Dictamina Rhetorica". https://bit.ly/2I36vhI (Archivio della Latinità Italiana del Medioevo), accessed 2018-05-28.

[19] Nathaniel E. Griffin. 1936. *Guido De Columnis - Historia destructionis Troiae.* The Mediaeval Academy of America, Cambridge, US. https://bit.ly/2EF02IR (Archivio della Latinità Italiana del Medioevo), accessed 2018-05-28.

[20] Paul O. Kristeller. 1961. Un' 'Ars dictaminis di Giovanni del Virgilio'. *Italia Medioevale e Umanistica* 4 (1961), 179–200. https://bit.ly/2MeCD7k (Archivio della Latinità Italiana del Medioevo), accessed 2018-05-28.

[21] Selene Mancuso. 2015. *Benvenuto da Imola, Glose Bucolicorum Virgilii (ad Buc. I, IV, VI, X). Studi critici ed edizione.* Master's thesis. Department of Philology, Literature and Linguistics, University of Pisa, Pisa, IT.
[22] Pietro Meloni. 1953. *Nicolai Treveti Expositio L. Annaei Senecae Agamemnonis.* Centro di Studi Filologici e Linguistici Siciliani, Palermo, IT. https://bit.ly/2KfRxYt (Biblioteca Italiana), accessed 2018-05-28.
[23] Enrico Menestò and Stefano Brufani. 1995. *Fontes Franciscani.* Edizioni Porziuncola, Assisi, IT.
[24] Giovanni Monleone. 1941. *Jacopo da Varagine e la sua Cronaca di Genova dalle origini al MCCXCVII.* FSI, Roma, IT, 3–414. https://bit.ly/2MhKlxs (Archivio della Latinità Italiana del Medioevo), accessed 2018-05-28.
[25] Bruno Nardi. 1996. *Dante Alighieri - Opere minori.* Riccardo Ricciardi Editore, Milano, IT, Chapter "Monarchia". https://bit.ly/2EEGQLg (Biblioteca Italiana), accessed 2018-05-28.
[26] Marta M. M. Romano. 2004. *Corpus Christianorum Continuatio Mediaevalis CLXXXIII.* Brepols Publisher, Turnhout, BE, Chapter "Raimundus Lullus – Ars amativa boni", 120–432. https://bit.ly/2wnBu2t (Archivio della Latinità Italiana del Medioevo), accessed 2018-05-28.
[27] Vincenzo Romano. 1951. *Giovanni Boccaccio - Genealogie deorum gentilium libri.* Laterza, Bari, IT. https://bit.ly/2ZEvcZI (Biblioteca Italiana), accessed 2018-05-28.
[28] Luca C. Rossi. 1998. *Commento all' 'Inferno' di Dante.* Scuola Normale Superiore, Pisa, IT. Reprinted in Rossi, Luca C. (ed.), "I commenti danteschi dei secoli XIV, XV e XVI", LEXIS Progetti Editoriali, Roma, IT, 1999 https://bit.ly/2EzazoX (Biblioteca Italiana), accessed 2018-05-28.
[29] Luca Carlo Rossi. 2002. 'Benevenutus de Ymola super Valerio Maximo'. Ricerca sull'Expositio. *Aevum - Rassegna di Scienze Storiche Linguistiche e Filologiche* 76 (2002), 369–423.
[30] Fedor Schneider. 1926. Untersuchungen zur italienischen Verfassungsgeschichte: Staufisches aus der Formelsammlung des Petrus de Boateriis. *Quellen und Forschungen aus italienischen Archiven und Bibliotheken* 18 (1926), 191–273.
[31] Fabio Stok. 1991. La 'Vita di Virgilio' di Zono de' Magnalis. *Rivista di Cultura Classica e Medioevale* 33, 2 (1991), 143–181.
[32] Carl Sutter. 1894. *Aus Leben und Schriften des Magisters Boncompagno.* Akademische Verlagsbuchhandlung von J.C.B. Mohr, Freiburg im Breisgau, DE. https://bit.ly/2ws07eg (Archivio della Latinità Italiana del Medioevo), accessed 2018-05-28.
[33] Mirko Tavoni. 2011. *Dante Alighieri - Opere.* Mondadori, Milano, IT, Chapter "De vulgari eloquentia". https://bit.ly/2HIfBRW (DanteSearch), accessed 2018-05-28.
[34] Vincenzo Ussani, Jr. 1959. *L. Annaei Senecae Hercules furens et Nicolai Treveti expositio.* Edizioni dell'Ateneo, Roma, IT. https://bit.ly/2KccRxN (Biblioteca Italiana), accessed 2018-05-28.
[35] Vittorio Zaccaria. 1967. *Tutte le opere di Giovanni Boccaccio.* Mondadori, Milano, IT, Chapter "De mulieribus claris". https://bit.ly/2I1VWvl (Biblioteca Italiana), accessed 2018-05-28.

REFERENCES

[36] 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.
[37] Alberto Casadei (Ed.). 2020. *Atti del Seminario "Nuove Inchieste sull'Epistola a Cangrande".* Pisa University Press, Pisa, IT.
[38] Carole E. Chaski. 2005. Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence* 4, 1 (2005).
[39] 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.
[40] 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) (Lecture Notes in Computer Science).* Springer, Trento, IT, 148–158. https://doi.org/10.1007/978-3-030-30754-7_15
[41] Patrick Juola. 2006. Authorship attribution. *Foundations and Trends in Information Retrieval* 1, 3 (2006), 233–334. https://doi.org/10.1561/1500000005
[42] 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
[43] 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
[44] 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

[45] 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

[46] 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

[47] 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

[48] Samuel Larner. 2014. *Forensic Authorship Analysis and the World Wide Web*. Springer, Heidelberg, DE.

[49] 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.

[50] 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

[51] 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

[52] 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

[53] 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

[54] 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

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

[56] 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

[57] 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.

[58] 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