The *Epistle to Cangrande* Through the Lens of Computational Authorship Verification

Silvia Corbara[1(✉)], Alejandro Moreo[1], Fabrizio Sebastiani[1], and Mirko Tavoni[2]

[1] Istituto di Scienza e Tecnologie dell'Informazione Consiglio Nazionale delle Ricerche, 56124 Pisa, Italy
{silvia.corbara,alejandro.moreo,fabrizio.sebastiani}@isti.cnr.it
[2] Dipartimento di Filologia, Letteratura e Linguistica Università di Pisa, 56126 Pisa, Italy
mirko.tavoni@unipi.it

Abstract. The *Epistle to Cangrande* is one of the most controversial among the works of Italian poet Dante Alighieri. For more than a hundred years now, scholars have been debating over its real paternity, i.e., whether it should be considered a true work by Dante or a forgery by an unnamed author. In this work we address this philological problem through the methodologies of (supervised) *Computational Authorship Verification*, by training a classifier that predicts whether a given work is by Dante Alighieri or not. We discuss the system we have set up for this endeavour, the training set we have assembled, the experimental results we have obtained, and some issues that this work leaves open.

Keywords: Machine learning · Authorship Verification · Digital humanities · Dante Alighieri · Medieval Latin

1 Introduction

The *Epistle to Cangrande*, from now on "EpXIII", is the thirteenth of the letters from Dante Alighieri's epistolary corpus that have survived until our times. Written in Latin, it is addressed to Can Francesco della Scala, known as Cangrande I, the ruler of the Italian cities of Verona and Vicenza at the beginning of the 14th century. Scholars traditionally divide it into 90 paragraphs and into 2 thematic portions: the first portion (paragraphs 1–13 – hereafter EpXIII(I)) is the dedicatory section, with proper epistolary characteristics, while the second portion (paragraphs 14–90 – hereafter EpXIII(II)) contains an exegesis of Alighieri's *Divine Comedy*, and in particular a commentary of the first few lines of its third part, the *Paradiso*. EpXIII became renowned over the centuries, especially because it would be the only analysis we have received from Dante Alighieri of his own masterpiece. However, since the start of the 19th century the authenticity of EpXIII has been questioned, and the issue has remained

unsolved. The academic community is split between those who consider EpXIII authentic, those who consider it a forgery, and those who consider authentic the first portion but not the second.

To support the forgery thesis, scholars (e.g., [5]) point out numerous passages in the composition where the logical sequence of discourse is cumbersome, or even incoherent, with itself or with other writings by Alighieri. Moreover, many have noticed that there is a profound dissimilarity between EpXIII(I) and EpXIII(II), in their themes, style, and rhythm [10]. Even figuring out a timeframe when the letter could have been written is not a trivial problem.

Among those who believe EpXIII to be authentic, some (e.g., [2, pp. 280–1]) claim that there is a lexical coherence that cuts through the entire EpXIII, and an inner cohesive logic. Additionally, [1] observes that a forger would have followed more closely Alighieri's prose, and thus, paradoxically, the style dissimilar from Alighieri's should be seen as a further proof of authenticity. Many also note that the author of EpXIII offers some non-traditional and potentially controversial explanations for some exegetical and linguistic aspects of the *Divina Commedia*, and this could indicate a prominent author, since a lesser personality would have probably trodden on more ordinary, "safer" grounds. For a more comprehensive discussion of this controversy, see the analysis in [2,17].

Given this debated and yet unsolved problem, and in order to gain a fresh perspective over it and thus offer scholars yet another useful tool for investigation, in this work we have applied to the "whodunnit" of EpXIII the methodologies of (supervised) *Computational Authorship Verification* (AV), a task concerned with using training data to generate a binary classifier that predicts whether a given text of unknown or disputed paternity was written by a given candidate author or not.

This article is structured as follows. In Sect. 2 we give a brief introduction to AV, also hinting at related works that have applied AV to Latin texts. In Sect. 3 we discuss the methods we have employed to tackle the EpXIII mystery, and the features we have used for generating the vectorial representations of texts that will be fed to the learning algorithm (and to the classifier, once trained). In Sect. 4 we present the results of our experiments, in which we also assess the accuracy of our classifier over the entire dataset we have used, and establish the relative contribution of the various features. Finally, Sect. 5 discusses issues that this work leaves open, and possible avenues towards their solution.

2   Computational Authorship Verification

Authorship Analysis (AA) can be defined broadly as "any attempt to infer the characteristics of the creator of a piece of linguistic data" [11, p. 238], which includes the author's biographical information (age, gender, mother tongue, etc.), as well as their identity. The core of this practice, also known as "stylometry", relies on the idea to identify the author not from the artistic value of the text, or from the meaning of the concepts proposed within it, but from a quantitative analysis of the document's style. Here, "style" is intended as a

summary statistics emerging from one or more numerical features that describe linguistic traits present in written texts, which are believed to remain more or less constant in an author's production and, conversely, to vary in noticeable fashion across different authors [11, p. 241]. These unique stylistic features are also known as "style markers".

This definition allows for every kind of textual trait, as long as it can be counted (hopefully: easily counted). It is the researcher's task to identify and extract the features that they deem most discriminative, i.e., most helpful for determining authorship. In particular, scholars started experimenting with this practice (well before the age of computers) by employing a single set of features comparable to the linguistic elements studied in classical philology, such as the frequencies of word lengths, sentence lengths, *hapax legomena*, and other specific terms. However, in the late 20th century, starting from the work of Mosteller and Wallace [16] on the *Federalist Papers*, the practice veered towards employing several sets of high-frequency features in parallel. Even though this approach captures textual traits of apparently minimal significance, this practice has proven effective in a variety of authorship analysis tasks, since the phenomena involved tend to be out of the conscious control of the author, and hence hard to modify or imitate. The noted historian Carlo Ginzburg describes this approach (as applied not only to text authorship issues, but to many other types of investigation as well) in his essay *Clues*, calling it the *Evidential Paradigm* [8].

The values of these stylistic features are collectively used as a simplified representation of the text, and employed for analyzing its authorship. This may be done via various methods, which are usually classified into similarity-based or machine learning -based. In the former class, specific algorithms are implemented to compute the similarity between different texts based upon a chosen similarity measure. In the latter class, a classifier is trained from a number of labelled training examples, using vectors of the chosen features (the style markers) as representations of the texts of interest; this enables the machine to leverage the values of the features in the training examples in order to classify new unlabelled documents. In the machine learning approach, AA is seen as an instance of (supervised) text classification, a task which generically deals with learning to classify text into a set of predefined classes, where the classes may represent topics, sentiments, literary genres, languages, and so on, depending on the application requirements.

In machine learning -based AA, the most popular methods still make use of "classical" machine learning algorithms, such as support vector machines (SVMs) or logistic regression (LR), even if deep learning algorithms have sometimes proved more accurate. This trend has also been confirmed in the PAN 2018 Author Identification shared task [13], where most of the systems presented were based on SVMs. This is due to two different reasons. On the one hand, in some application domains there is a systematic scarcity of annotated data, which clashes with the fact that deep learning methods typically require very large training sets. On the other hand, deep learning methodologies notoriously lack on the explainability side, which is undesirable when the investigation

concerns a case of genuine controversy and it is indeed advisable that the factors supporting the conclusion drawn by the system can be properly exhibited [11, p. 307]. Both issues are especially relevant in the humanities, where the documents available are usually rather limited in number (as in the case of medieval Latin) and the main objective of computational studies is certainly not to replace the philologist, but to support their research with supplementary evidence and tools, which then need to be as explicit as possible.

The problem of EpXIII is an instance of *Authorship Verification* (AV), a subtask of AA that consists in determining whether a document of unknown or disputed paternity has been written by a given candidate author or by someone else. It is thus different from *Closed-set Authorship Attribution*, where the goal is to infer, for a document of unknown or disputed paternity, the most likely author among a finite set of candidate authors [14]. AV is thus a binary classification task, where the positive training examples are texts known to be by the candidate author, and the negative training examples are texts known to be by other "similar" authors writing in the same language.

In the humanities, AV is not a frequently tackled task: usually, scholars have more than one possible candidate author for a given document, and thus approach their research in a closed-set authorship attribution setting. Some examples of such works are [3] for the 15th Book of Oz, [12] on the works by Monk of Lido and Gallus Anonymous, and the many works presented at a recent workshop about the true identity of pseudonymous novelist Elena Ferrante [19]. An exception to this pattern is [18] on Pliny the Younger's "Letter on the Christians" to Trajan, which is indeed framed as an AV problem.

3   AV Methods Applied to EpXIII

We have approached the problem of the authorship of EpXIII as a supervised binary classification task implemented via a linear classifier. After a few initial test with Logistic Regression (LR) and SVMs, we finally decided to stay with the former, since (a) preliminary experiments on our data had indicated that the two had a similar level of accuracy, and (b) unlike for SVMs, the output of LR admits a probabilistic interpretation, i.e., it can be interpreted as the ("posterior") probability that the document belongs to the class. See [4, pp. 205-6] for a more complete description of LR.

As discussed in Sect. 2, computational methods for AV map a textual document into a vector of features, each representing some linguistic phenomenon that is deemed related to authorship. To this aim, we have selected a combination of different feature types, since this approach usually yields better performance than just using a single type of features [9]. Each feature type we have used has been shown effective to some extent in other authorship-related tasks. The set of features we ended up using 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).

Note that we ignore punctuation marks, since they were not inserted by the authors (punctuation was not used in medieval Latin, and such marks have been introduced into texts for editorial purposes).

In order to deal with the high dimensionality of the feature space we subject the feature types 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.,

$$\chi^2(t_k, d_j) = \frac{[\Pr(t_k, c_i)\Pr(\bar{t}_k, \bar{c}_i) - \Pr(t_k, \bar{c}_i)\Pr(\bar{t}_k, c_i)]^2}{\Pr(t_k)\Pr(\bar{t}_k)\Pr(c_i)\Pr(\bar{c}_i)} \tag{1}$$

where probabilities are interpreted on the event space of documents; in other words, $\Pr(t_k, c_i)$ represents the probability that, for a random document that belongs to class c_i, 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.,

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

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

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

where $\#(t_k, d_j)$ is the number of occurrences of feature t_k in document d_j.

3.1   The Training Set

As already explained in Sect. 1, EpXIII consists of two sections, EpXIII(I) and EpXIII(II), distinct from each other for purpose and style. More importantly, not all scholars agree that the two sections are from the same author. We have thus decided to split the AV problem into two different AV sub-problems, and thus to train two different classifiers, one for EpXIII(I) and another for EpXIII(II). In order to train and evaluate these classifiers, we have created two datasets of medieval Latin texts, by including in each such dataset documents which can be considered, linguistically and stylistically speaking, similar to the document (EpXIII(I) or EpXIII(II)) we are dealing with. The positive class (Dante) is represented by all known works in Latin that are unquestionably by Alighieri: his other

12 epistles for the EpXIII(I) dataset, and *De Vulgari Eloquentia* and *Monarchia* for the EpXIII(II) dataset.[1] Conversely, for the negative class (NotDante) we have assembled two sets of Latin texts by coeval authors: a set of 282 epistles from more than 20 different authors (for the EpXIII(I) dataset) and a set of 28 miscellaneous texts, mostly literary commentaries and treatises, from 19 different authors (for the EpXIII(II) dataset); all the documents date between the 13th and the 15th century (see [6] for more details on these two datasets). In the end, as described in Column 1 of Table 1, the EpXIII(I) and EpXIII(II) datasets consist of 294 and 30 texts, respectively.

Table 1. The two datasets and the results obtained on them.

	1	2	3	4	5	6	7
	# full docs	# training docs	Prediction	Pr(Dante)	F_1(Dante)	Macro-F_1	Micro-F_1
EpXIII(I)	294	1310	NotDante	0.24	0.957	0.886	0.981
EpXIII(II)	30	12312	NotDante	0.39	0.400	0.688	0.775

We have preprocessed all the documents by

– Lower-casing the entire text.
– Removing any symbol 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 the citations in Latin with asterisks, and the citations in languages different from Latin (mostly Florentine vernacular) with curly brackets; this is both to ignore them in the computation (since they are the production of someone different than the author of the text) and to mark a potential authorial-related feature for future development, i.e., the usage of citations in different languages.
– Replacing every occurrence of character "v" with 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.[2]

Additionally, in order to increase the number of training samples, we subject all documents in both datasets to a segmentation policy, i.e., we split each document into segments and consider each resulting segment as a separate, additional training document. This approach is a common practice in ML-based AA when only few labelled texts are available [15, p. 514]. More in detail, by employing

---

[1] Other works by Alighieri, including the *Divine Comedy*, are not included in these two datasets since Alighieri wrote them not in Latin but in the Florentine vernacular (*volgare*), which was to form the basis of what is nowadays the Italian language.
[2] In the medieval writing 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".

the Natural Language Toolkit (NLTK) sentence tokenizer module[3], we split each document into sentences; if a sentence is too short (fewer than 8 words) we join it with the subsequent one, unless it is the last sentence in the text, in which case we join it with the previous one. We then join the sentences thus derived into segments of n consecutive sentences each, without overlapping, and we consider each segment as a single labelled example; the current value of n we use is 3. The final result of this procedure is shown in the first 2 columns of Table 1.

4   Experiments

We train our two "Dante vs. NotDante" classifiers 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}.

The predictions by the optimized classifiers are shown in Column 3 of Table 1, which shows that the classifiers consider both EpXIII(I) and EpXIII(II) the production of someone else than Alighieri. Column 4 lists the posterior probabilities returned by the LR classifier, and indicate that the classifier attributes a probability of 0.24 (resp., 0.39) to the fact that EpXIII(I) (resp., EpXIII(II)) was written by Alighieri; in other words, the hypothesis that Alighieri might be the true author is rejected with more strength for EpXIII(I) than for EpXIII(II).

In order to determine the degree of reliability of our two classifiers, and hence establish how trustworthy the above predictions about EpXIII(I) and EpXIII(II) are, we subject the algorithm to a "leave-one-out" validation test, which consists in predicting, for each dataset D, for each author a represented in the dataset, and for each document d in the dataset, whether a is the author of d or not, where the prediction is issued by a "a vs. (NOT a)" binary classifier trained on all documents in D/d. We exclude from this analysis authors which have only 1 text in the dataset, which means that we train binary classifiers for 5 authors for EpXIII(I) and 6 authors for EpXIII(II); this leads to $5 \times 294 = 1470$ predictions for EpXIII(I) and $6 \times 30 = 180$ predictions for EpXIII(I). Note that, in order to recreate the conditions of the actual classification of EpXIII(I) and EpXIII(II), and in order to avoid any overlap between test and training samples, (i) we test only on the original entire documents (thus ignoring the segments), and, (ii) when document d is used as a test document, we exclude from the training set all the segments derived from d. As evaluation measures we use the well-known *macroaveraged* F_1 and *microaveraged* F_1.

The results of these experiments are shown in Columns 6 and 7 of Table 1. As it can be seen, the classifiers, and especially the one for EpXIII(I), obtain a good level of accuracy (notwithstanding the small size of many training sets), in line with other state-of-the-art methods.

Column 5 of the same table reports the F_1 results for the two "Dante vs. NotDante" classifiers. In this case the F_1 values for EpXIII(II) are much lower

---

[3] https://www.nltk.org/api/nltk.tokenize.html.

Table 2. Results of the feature ablation study; (-) indicates the feature type omitted.

		All features	(-) char n-grams	(-) word n-grams	(-) function words	(-) verbal endings	(-) word lengths	(-) sentence lengths
EpXIII(I)	Macro-F_1	0.886	0.727 (-17.95%)	0.784 (-11.51%)	0.795 (-10.27%)	0.908 (+2.48%)	0.886 (0.00%)	0.886 (0.00%)
	Micro-F_1	0.981	0.947 (-3.47%)	0.979 (-0.20%)	0.985 (+0.41%)	0.983 (+0.20%)	0.981 (0.00%)	0.981 (0.00%)
EpXIII(II)	Macro-F_1	0.688	0.486 (-29.33%)	0.463 (-32.66%)	0.671 (-2.44%)	0.690 (+0.29%)	0.688 (0.00%)	0.679 (-1.29%)
	Micro-F_1	0.775	0.590 (-23.88%)	0.528 (-31.88%)	0.750 (-3.25%)	0.775 (0.00%)	0.775 (0.00%)	0.741 (-4.38%)

than for EpXIII(I); here, the classifier is penalized for not attributing *Monarchia* to Alighieri (since *Monarchia* is one of the two texts by him in the dataset, this mistake alone makes recall equal 0.50, and prevents the value of F_1 – which is the harmonic mean of precision and recall – to be high enough), while it correctly classifies 26 out of the 28 negative examples.

4.1   Feature Ablation and Feature addition

In order to further analyze the behaviour of our classifiers, we have conducted a study of individual feature types via either feature ablation or feature addition. In the feature ablation study the single feature type t is omitted, and the resulting Macro-F_1 and Micro-F_1 values are compared with the analogous values resulting from the "all features" study reported in Columns 6 and 7 of Table 1. In the feature addition study only the single feature type t is used, and the resulting Macro-F_1 and Micro-F_1 values are compared with the analogous values obtained by a hypothetical classifier that uses the empty set of features, i.e., by the random classifier.[4] The results of the feature ablation study are shown in Table 2, while the results of the feature addition study are shown in Table 3.

The feature addition study indicates that all feature types used are in principle informative, as shown by the (often dramatic) improvements in accuracy that each feature type brings about with respect to the random classifier. However, the feature ablation study shows that the very same feature types, when removed from an "all features" classifier, bring about a much less dramatic deterioration (if any deterioration at all – see e.g., word lengths). We think this has two possible explanations. First, some feature types are, when other feature types are already present, redundant; this may be the case, e.g., for verbal endings, since the same character string that forms a verbal ending may already be present

---

[4] "The" random classifier is indeed an abstraction; by the accuracy of the random classifier we mean the average accuracy of all possible classifiers, i.e., of all possible ways the test set might be classified. It is easy to show that this is equivalent to the accuracy of a 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.

Table 3. Results of the feature addition study; (+) indicates the feature type inserted.

		Random classifier	(+) char n-grams	(+) word n-grams	(+) function words	(+) verbal endings	(+) word lengths	(+) sentence lengths
EpXIII(I)	Macro-F_1	0.217	0.818 (+276.96%)	0.630 (+190.32%)	0.559 (+157.60%)	0.523 (+141.01%)	0.429 (+97.70%)	0.368 (+69.59%)
	Micro-F_1	0.264	0.958 (+262.88%)	0.907 (+243.56%)	0.799 (+202.65%)	0.732 (+177.27%)	0.558 (+111.36%)	0.527 (+99.62%)
EpXIII(II)	Macro-F_1	0.149	0.390 (+161.74%)	0.367 (+146.31%)	0.532 (+257.05%)	0.449 (+201.34%)	0.363 (+143.62%)	0.230 (+54.36%)
	Micro-F_1	0.151	0.426 (+182.12%)	0.476 (+215.23%)	0.625 (+313.91%)	0.462 (+205.96%)	0.310 (+105.30%)	0.204 (+35.10%)

in the feature set as a character n-gram. Second, when the dimensionality of the vector space is already high, adding other dimensions may bring about (or increase) overfitting; this is especially true in our case, in which the amount of training data is small.

Aside from these considerations, both experiments seem to show that the feature types that contribute most to AV accuracy are word n-grams and character n-grams, followed by function words and verbal endings; conversely, the contribution of word lengths and sentence lengths seems to be comparatively smaller.

5   Conclusion and Future developments

The predictions output by our classifier seem to align with the theory that the entire EpXIII was the work of a malicious forger. Nevertheless, the conclusions presented here should not be considered definitive. As stated before, the methods displayed here are only the current stage of a project which we consider to be far from completion. The ideas we want to pursue in the near future in order to improve the system can be divided into 3 areas: (a) the datasets, (b) genre and topic bias, and (c) the feature set.

First of all, we intend to expand the datasets with additional documents. Working with medieval Latin makes this task more difficult than when working with modern languages. One possibility we are exploring is the addition of the texts made available by Kabala in [12] to our datasets; since these are one to two centuries older than the ones in our datasets, it remains to be seen whether this addition would be beneficial or detrimental.

As already mentioned in Sect. 1, EpXIII, if proved authentic, would be the only commentary that we have received by Alighieri on his own *Divine Comedy*. Unfortunately, this also means that, while one of the documents we want to classify (i.e., EpXIII(II)) is a commentary on the *Divine Comedy*, no other training document from class Dante is. On the contrary, some documents that are contained in the corpus of EpXIII(II) are, since at the time the *Divine Comedy* had attracted the attention of learned people. It is thus possible that the EpXIII(II) classifier is (at least partially) recognizing the topic of a document, and not its

author[5]. This suspicion is reinforced by the fact that, as feature types, we use both character and word n-grams, which are effective features in classification by topic.

In order to understand whether the dataset is biased by topic and/or by genre, we have run two additional experiments. The first one consisted in labelling each EpXIII(II) text according to whether it consists or not of a commentary on a literary work, and running a leave-one-out validation test obtained by repeatedly training "Commentary vs. NotCommentary" classifiers *on the very same feature set* used for the "Dante vs. NotDante" experiment. The second experiment was analogous, aside from the fact that the EpXIII(II) texts were now labelled according to whether they discussed Alighieri's *Divine Comedy* or not, thus implementing a "Comedy vs. NotComedy" distinction. Note that the former experiment is about *classification by genre*, while the second is about *classification by topic*, which are two dimensions conceptually orthogonal to the one we are interested in, i.e., *classification by author* (or: authorship verification). Both experiments returned F_1 values of 1.00. This suggests that the results presented in Sect. 4 are likely influenced by both genre bias and topic bias present in the dataset, i.e., that those results are not entirely due to the classifier's ability to recognize authorship, as instead one would hope; on the other hand, refraining from inserting commentaries on literary works, and even on the *Divine Comedy* itself, in the NotDante dataset, would prevent us from comparing EpXIII with all the writers who might conceivably be its authors. This is an important open issue, that we plan to address by using methods devised in the field of "fair machine learning" (see, e.g., [7]).

Finally, we intend to further improve the feature set by experimenting with different feature types: for example, employing a POS-tagger for Latin could result in the detection of authorial traits related to the syntactic habits of the author.

References

1. Ascoli, A.R.: Access to authority: dante in the epistle to cangrande. In: Baranski, Z.G. (ed.) Seminario Dantesco Internazionale/International Dante Seminar 1, pp. 309–352. Le Lettere, Firenze (1997)
2. Azzetta, L.: Epistola XIII. In: Baglio, M., Azzetta, L., Petoletti, M., Rinaldi, M. (eds.) Nuova edizione commentata delle opere di Dante, Epistole. Egloge. Questio de aqua et terra, vol. 5, pp. 271–487. Salerno Editrice, Roma (2016)
3. Binongo, J.N.G.: Who wrote the 15th Book of Oz? An application of multivariate analysis to authorship attribution. Chance **16**(2), 9–17 (2003)
4. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
5. Casadei, A.: Sempre contro l'autenticità dell'Epistola a Cangrande. Studi Danteschi **LXXXI**, 215–246 (2016)

---

[5] Note that this specific problem is, as stated, confined to the classifier of EpXIII(II), and does not affect the one for EpXIII(I). Still, in the EpXIII(I) dataset there might be other types of topic bias that we have not detected yet.

6. Corbara, S., Moreo, A., Sebastiani, F., Tavoni, M.: L'Epistola a Cangrande al vaglio della computational authorship verification: Risultati preliminari (con una postilla sulla cosiddetta "XIV Epistola di Dante Alighieri"). In: Casadei, A. (ed.) Atti del Seminario "Nuove Inchieste sull'Epistola a Cangrande". Pisa University Press, Pisa (2019, forthcoming)
7. Dwork, C., Immorlica, N., Kalai, A.T., Leiserson, M.D.: Decoupled classifiers for group-fair and efficient machine learning. In: Proceedings of the 1st ACM Conference on Fairness, Accountability and Transparency (FAT 2018), pp. 119–133, New York(2018)
8. Ginzburg, C.: Clues, myths and the historical method. In: Clues: Roots of an Evidential Paradigm, pp. 96–125. Johns Hopkins University Press, Baltimore (1989)
9. Grieve, J.: Quantitative authorship attribution: an evaluation of techniques. Lit. Linguist. Comput. **22**(3), 251–270 (2007)
10. Hall, R.G., Sowell, M.U.: Cursus in the can grande epistle: a forger shows his hand? Lectura Dantis **5**, 89–104 (1989)
11. Juola, P.: Authorship attribution. Found. Trends Inf. Retrieval **1**(3), 233–334 (2006)
12. Kabala, J.: 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, pp. 1–32 (2019). https://doi.org/10.1007/s10579-018-9424-0
13. Kestemont, M., et al.: Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection. In: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2018), pp. 1–25. Avignon (2018)
14. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. J. Am. Soc. Inform. Sci. Technol. **60**(1), 9–26 (2009)
15. Luyckx, K., Daelemans, W.: Authorship attribution and verification with many authors and limited data. In: Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008), pp. 513–520. Manchester (2008)
16. Mosteller, F., Wallace, D.L.: Inference and Disputed Authorship: The Federalist. Addison-Wesley, Reading (1964)
17. Sasso, G.: Sull'Epistola a Cangrande. La Cultura **3**, 359–446 (2013). https://doi.org/10.1403/75324
18. Tuccinardi, E.: An application of a profile-based method for authorship verification: investigating the authenticity of Pliny the Younger's letter to Trajan concerning the Christians. Digit. Scholarsh. Humanit. **32**(2), 435–447 (2016)
19. Tuzzi, A., Cortelazzo, M.A. (eds.): Drawing Elena Ferrante's profile. Workshop Proceedings. Padova University Press, Padova (2017)