Authenticating the writings of Julius Caesar

Authenticating the Writings of Julius Caesar

Mike Kestemont

University of Antwerp, Prinsstraat 13, B-2000 Antwerp, Belgium

Corresponding author: mike.kestemont@uantwerp.be; 0032/477.91.86.68

Justin Stover

University of Oxford, All Souls College, Oxford OX1 4AL, United Kingdom

justin.stover@classics.ox.ac.uk

Moshe Koppel

Bar-Ilan University, 52900 Ramat-Gan, Israel

moishk@gmail.com

Folgert Karsdorp

Center for Language Studies, Radboud University, P.O. Box 9103, NL-6500 HD, Nijmegen, The Netherlands

fbkarsdorp@fastmail.nl

Walter Daelemans

University of Antwerp, Belgium, Prinsstraat 13, B-2000 Antwerp, Belgium

walter.daelemans@uantwerp.be

Abstract

In this paper, we shed new light on the authenticity of the *Corpus Caesarianum*, a group of five commentaries describing the campaigns of Julius Caesar (100-44 BC), the founder of the Roman empire. While Caesar himself has authored at least part of these commentaries, the authorship of the rest of the texts remains a puzzle that has persisted for nineteen centuries. In particular, the role of Caesar's general Aulus Hirtius, who has claimed a role in shaping the corpus, has remained in contention. Determining the authorship of documents is an increasingly important authentication problem in information and computer science, with valuable applications, ranging from the domain of art history to counter-terrorism research. We describe two state-of-the-art authorship verification systems and benchmark them on 6 present-day evaluation corpora, as well as a Latin benchmark dataset. Regarding Caesar's writings, our analysis allow us to establish that Hirtius's claims to part of the corpus must be considered legitimate. We thus demonstrate how computational methods constitute a valuable methodological complement to traditional, expert-based approaches to document authentication.

Keywords: Authentication, Authorship Verification, Stylometry, Julius Caesar

1. Introduction

Throughout the twentieth century, influential post-structuralist thinkers, such as Foucault or Barthes have fiercely argued against the importance of the notion of 'authorship' (Barthes, 1968; Foucault, 1969). Across many fields in the Humanities for instance, this famously led to a temporary devaluation of the importance attached to the relationship between texts and their original producers (Love, 2002). However, numerous examples demonstrate that the public interest in authorship currently shows few signs of abating. The highly mediatized discovery of an pseudonymously published novel by the appraised Harry Potter novelist J.K. Rowling is a good example in this respect (Juola, 2015, 2013). In recent years, many other authorship-related research, such as the Shakespeare controversy (Burrows, 2012), has continued to make frequent headlines in the popular media. In academia too, the much debated application of bibliometry (Cronin, 2001) or well-known cases of plagiarism (Maurer et al., 2006) hardly suggest that the notion of authorship would have suffered a major loss of public interest. Unsurprisingly, automated authorship analysis (Juola, 2006; Koppel et al., 2009; Stamatatos, 2009b) currently receives increasing attention in Computer and Information Sciences too, as a form of document authentication with promising practical applications across various domains, such as plagiarism detection (Stein et al., 2011) or even in forensic sciences (Chaski, 2005; Juola, 2015).

Most computational authorship studies in computer science are still restricted to present-day document collections. In this paper, we illustrate the broader applicability of computational authorship verification by reporting a high-profile case study from Classical Antiquity (Koppel & Seidman, 2013; Stover et al., 2016). The 'War Commentaries' by Julius Caesar (*Corpus Caesarianum*) refers to a group of Latin prose commentaries, describing the military campaigns of the world-renowned statesman Julius Caesar (100–44 BC), the founder of the Roman Empire. While Caesar must have authored a significant portion of these commentaries himself, the exact delineation of his contribution to this important corpus remains a controversial matter. Most notably, Aulus Hirtius – one of Caesar's most trusted generals – is sometimes believed to have contributed significantly to the corpus. Thus, the authenticity and authorship of the Caesarian corpus is a philological puzzle that has persisted for nineteen centuries. In this paper, we use computational authorship verification to shed new light on the matter.

Below, we will first situate our work in the field of stylistic authentication studies, focusing on the style versus content debate, as well as the difference between open set and closed set attribution. We go on to discuss our implementation of two verification systems, a first-order and a second-order approach, which represent the state of the art in the field, given the results of the latest relevant competitions on

authorship verification. We first benchmark both systems on 6 present-day data sets, before testing them on an evaluation set of Latin documents from Antiquity. Finally, we analyse the *Corpus Caesarianum*, offering a detailed discussion of the historical implications of our results.

2. Style vs Content

Traditionally, scholars have long employed a pragmatic distinction between the 'style' and 'content' of written documents (Stamatatos et al., 2000), the former encapsulating all aspects of an individual author's language use at the textual level (Hermann et al., 2015). In authorship studies, there is nowadays a general consensus that features related to style are more useful (Juola, 2006; Koppel et al., 2009; Stamatatos, 2009b), since topical, content-related features vary much more strongly across the documents authored by a single individual. Much research nowadays therefore concerns ways to effectively extract stylistic characteristics from documents that are not affected by a text's specific content or genre (Argamon & Levitan, 2005; Kestemont et al., 2012; Efstathios, 2013; Sapkota et al., 2015; Seroussi et al., 2014; Sapkota et al., 2014). This has not always been the case: historical practitioners in earlier centuries, commonly based attributions on a much looser defined set of linguistic criteria, including, for instance, the use of conspicuous, rare words (Love, 2002; Kestemont, 2014). Naturally, an expert reader's subjective intuitions (*Gelehrtenintuition*, *connoisseurship*) would play a much larger role in studies than would nowadays be acceptable. Especially, the focus on striking characteristics would turn out to be problematic. Importantly, low-frequency features are typically tied to fairly specific topics, and thus do not scale well to new texts. More importantly, these whimsical items also appeal to imitators and followers: in the case of malignant forgeries or benigne epigones, the authentication of documents will fail, if it is restricted to easy-to-copy, low-frequency characteristics (Love, 2002).

The pioneering work by Mosteller and Wallace on the pseudonymously published *Federalist papers* has marked a turning point in this respect (Mosteller & Wallace, 1964). Mosteller and Wallace proposed to rigidly restrict analyses to high-frequency characteristics and only considered an author's use of function words, or the small and closed set of grammatical items in a language which – as opposed to content words as nouns or verbs – do not carry a straightforward semantics when used in isolation (e.g. the article 'the' or the preposition 'of') (Aronoff & Fudeman, 2005). For authorship studies, function words are extremely attractive: they are frequent and well-distributed variables across documents, and consequently, they are not specifically linked to a single topic or genre. Importantly, psycholinguistic research suggests that grammatical morphemes are less consciously controlled in human language processing, since they do not actively attract cognitive attention (Stamatatos, 2009b; Binongo, 2003; Argamon & Levitan, 2005; Peng et al., 2003). This suggests that function words are relatively resistant to stylistic imitation or forgery.

With respect to function words, a number of recent developments are relevant. Ever since the *Federalist papers*, research into English-language documents has dominated authorship studies. In English, many functional morphemes are realised as individual words which can be easily identified in running text (Aronoff & Fudeman, 2005). In recent decades, the attention for other, low-resource languages has increased, including languages that display a much higher level of word inflection (e.g. the Finno-Ugric family) (Rybicki & Eder, 2011). Until fairly recently, other types of style markers (e.g. syntactical), rarely outperformed simple, word-level style markers (Holmes, 1994, 1998; Halteren et al., 2005). Later, character n-grams were introduced as a powerful alternative to function words (Kjell, 1994; Daelemans, 2013). This representation from Information Retrieval (originally used for automatic language identification) models texts at the sub-word level and segments them into a series of consecutive, partially overlapping groups of *n* characters; under a third-order trigram model ($n = 3$), for instance, the word 'trigrams' would yield the n-grams {'tri', 'rig', 'gra', 'ram', 'ams'}.

Multiple studies have demonstrated the excellent performance of character n-grams for modelling authorship, especially when it comes to more highly inflected languages such as Latin (Sidorov et al., 2014; Efstathios, 2013). This modelling strategy has the advantage that it can also capture morphemic information at the subword level, and is thus potentially sensitive to functional morphemes that are not realised as individual words (e.g. word endings) (Kestemont, 2014; Sapkota et al., 2015; Stamatatos, 2009b). Similarly, certain n-grams also pick up word stems and research increasingly demonstrates that text representations based on function words can be supplemented with information from lower-frequency strata in languages (Burrows, 2007), such as word stems (Koppel et al., 2009). Naturally, such approaches carefully need to avoid overfitting on the content of a specific document collection. Recent research demonstrated that predominantly functional character n-grams (including punctuation (Grieve, 2007)) are powerful authorship predictors (Sapkota et al., 2015). This helps explain why this family of features proves more robust with respect to cross-genre problems (Efstathios, 2013; Sapkota et al., 2014). Other recent studies have successfully applied Bayesian topic models to automatically separate style from content (Seroussi et al., 2014).

This paper will not dwell on feature selection, although we recognise the substantial efforts and advances which have been made on the topic of feature engineering in authorship studies. We limit the stylistic properties studied below to two commonly used feature types: word unigrams and character n-grams. These feature types have the advantage that they can be easily extracted from corpora, without requiring the application of preprocessing tools, such as part-of-speech taggers or parsers, which might not be available for all languages. Their relevance has moreover clearly motivated in the existing literature (Daelemans, 2013; Kestemont, 2014; Sapkota et al., 2015; Stamatatos, 2009b). While many studies

have indeed reported the successful application of other feature types (Stamatatos, 2009b), it is clear from comparative experiments that word unigrams and character n-grams represent state of the art feature types in authorship studies.

3. Methods

A number of different experimental procedures should be distinguished in present-day authorship studies (Stamatatos, 2009b). A first important distinction is that between authorship attribution and authorship verification (also known as open-set attribution). In the simple attribution scenario, the task is to attribute an anonymous text to a known author, through selecting the correct author from a set of candidate authors. In this closed-set scenario, the algorithm can safely assume that the correct target author is present in the set of available candidate authors, a scenario resembling a police line-up. It has been shown that the difficulty of this task increases as the number of candidate authors grows, and the length and or number of the available texts decreases (Daelemans & Van den Bosch, 2005). While the attribution setup is not incompletely unrealistic, it has been noted that in many real-world applications, it cannot be guaranteed that a text's true author is present among the candidates. This is why the verification scenario was introduced, in which the task is to decide whether *or not* an anonymous text was written by a given candidate author (hence, *verification*). The verification setup is known to be a more generic, yet also more difficult setup. Recent research has explored interested ways of combining both attribution and verification in a single system (Puig et al., 2016), although both setups are usually treated separately. The Caesarian corpus under scrutiny is a textbook example of a problem in authorship verification, since we do not have any guarantees as to the identity of the authors involved. For this paper, we will therefore use generic implementations of two verification methods which represent the state of the art in the field, especially when looking at the results of the latest PAN competitions. Both systems have proven to be successful approaches to authorship verification, and many of the top-performing contestants in competitions have integrated variations of them.

Authorship verification is a problem which has been studied for a number of years in the annual PAN competition. The design and evaluation of our analyses closely adheres to this competition's conventions to increase the comparability of our results (Stamatatos et al., 2014). Each dataset in the PAN competition consists of a set of 'problems', in which at least one, but possible more 'known' documents are available, which were all written by the same target author. Additionally, each problem defined an 'unknown text' for which has to be determined whether or not it has been written by the author of the 'known' texts, through assigning a score between 0 (definitely not the same author) and 1 (definitely the same author),

with a threshold at .5. Systems are allowed to leave a selection of difficult problems unanswered by assigning a score of exactly .5. The problems in each dataset fell apart in two non-overlapping sets: one development set of problems, on which systems could be calibrated, and a roughly equal-sized set of test problems, on which the calibrated systems were evaluated. The performance of the submitted systems is evaluated on the basis of two metrics: the AUC score (*area under the curve*, a well-known scalar evaluation score for binary classifiers) and the more recently proposed c@1 score (Peñas & Rodrigo, 2011). Unlike the AUC score, c@1 extends the traditional accuracy score (i.e. the ratio of correct answers), by rewarding careful systems that choose to leave those problems unanswered which it considers too difficult. The final performance of systems is reported as the product of the AUC and c@1 metric. Following the conventions used at the PAN competition, we statistically compare the accuracy of classifiers using approximate randomisation: this non-parametric test is valuable it does not make assumptions about the (potentially highly complex) distributions of the compared system outputs.

3.1. Verification Systems

The first verification system (termed O1 here) used here was seminally introduced by Kjell et al. (Kešelj et al., 2003) and was subsequently refined (Potha & Stamatatos, 2014; Kestemont et al., 2011; Stamatatos, 2009a). O1 resorts to the direct (or 'first order') calculation of a distance metric between a target author's stylistic profile in a given problem, and the unknown text. Following (Potha & Stamatatos, 2014; Koppel & Seidman, 2013), we define an author's profile here as the mean centroid of the known document vectors for that author (i.e. we average an author's score for a particular term across all training texts). Originally, O1 was introduced with a specific distance metric, called 'common n-grams' (cng). Let A and B be the respective vectors representing an author's centroid and the unknown document respectively; consisting of n character n-gram values in some fixed order. Let a_i and b_i represent the value of the i-th feature in both documents respectively:

$$cng(A, B) = \sum_{i=1}^{n} \left(\frac{2(a_i - b_i)}{a_i + b_i} \right)^2 \tag{1}$$

Studies vary in their exact implementation of this method: the earliest papers would calculate this distance function only for character n-grams which were present in both the profile and the unknown document (hence 'common' n-grams), but subsequent research showed that it is beneficial to apply the distance function only to the items which are present in the unknown document (Stamatatos, 2007), so that we use this implementation. To verify whether the unknown document was written by the target author in the problem, O1 uses thresholding: unknown documents resulting in a distance below this threshold are attributed to the target author, while all others are not. To normalize the resulting distance

score to probability scores in the 0-1 range, they are scaled using the set of all non-zero pairwise scores which can obtained between the known documents in a problem set, before their positive complement is taken (Potha & Stamatatos, 2014). While O1 has so far primarily been used with the cng metric, it can also be used with the other distance metrics introduced below.

The second verification system (termed O2 here) is a generic implementation of the General Imposters (GI) framework (Koppel & Winter, 2014). The general intuition behind the GI, is not to assess whether two documents are simply similar in writing style, given a static feature vocabulary, but rather, it aims to assess whether two documents are significantly more similar to one another than other documents, across a variety of stochastically impaired feature spaces (Stamatatos, 2006; Eder, 2012), and compared to random selections of so-called distractor authors (Juola, 2015), also called 'imposters'. O1 relies on the calculation of a direct, first-order distance measure between two documents to assess whether they are similar enough to be attributed to the same individual. The GI, however, resorts to the calculation of a 'second-order' metric (see Alg. 1, SI). Let x be the vector representing an anonymous document which is compared to $T = \{t_1, \ldots, t_n\}$, a set of documents by the target author. The task is to determine whether the documents in T were or were not written by the same author as x. Additionally, the GI procedure has access to $I = \{i_1, \ldots, i_n\}$, a set of distractor documents by so-called imposter authors. The GI then starts a bootstrapped procedure: during k iterations, it randomly samples a subset of the available features, as well as a random subset of imposters from I as I'. In each iteration, we determine whether x is closer than any of the documents in T than in I', given the impaired feature space and a distance function. Instead of returning a first-order distance, the GI returns a second-order metric, indicating the proportion of iterations in which x was closer to an item in T than in I'. As a proportion, the second-order score produced by O2 will automatically lie between 0 and 1 (higher scores indicate a higher attribution confidence). A similar thresholding procedure is therefore applied as with O1. O2 too can used with a variety of distance metrics, including the cng metric used in O1.

Note that O2 is an example of an 'extrinsic' verification method (Juola & Stamatatos, 2013): as opposed to the 'intrinsic' setup of O1, O2 also uses known documents from other authors in a particular problem set. In this paper, we sample imposter authors from the known documents that are available for other authors in a particular problem set. To ensure the comparability of O1 and O2, we sample author profiles (i.e. mean centroids), instead of individual documents from the imposter pool. Previous studies have automatically crawled the web for useful imposter documents, which yields results that might be difficult to reproduce exactly. Additionally, there is the inherent danger that one might obtain imposter documents that were indeed written by the target author, which would compromise the proper working of O2. Naturally, this problem is even more real in the case of the Latin data sets used below, because of the

relatively sparse online availability of Latin documents from Classical Antiquity.

3.2. Vector space models

In technical terms, a collection of texts in authorship studies is typically represented using a vector space model (VSM), as is common in text classification research (Sebastiani, 2002; Stamatatos et al., 2000). Both O1 and O2 are applied to such a VSM, yielding a matrix-like representation of a text collections, in which each document is assigned an equal-sized vector, which numerically represents a selection of its stylistic and linguistic properties, also called features, such as word unigram frequencies (Salton & Buckley, 1988; Manning et al., 2008). This process of vectorization typically operates under a 'bag-of-words' assumption, which models the occurrence of items in a text, but is in many cases insensitive to their relative order or exact position in a document. A number of different VSMs are currently dominant, the choice for which clearly reflects the style vs content assumptions outlined above.

The simplest vectorization model is the term-frequency model (tf), which records the relative frequency of the individual terms (e.g. words or n-grams) in a document in some fixed order. In authorship studies, it is not uncommon to aggressively truncate such VSMs to the most frequent items in the document collection (sometimes as little as 30 items (Burrows, 2002)). This truncation is a simple yet efficient strategy to combat vector sparsity and automatically causes models to focus on functional morphemes, since grammatical items are typically the most frequent ones in corpora (Stamatatos, 2009b). When allowing larger vectors, the tf-model has the disadvantage that it quickly comes to suffer from sparsity artefacts. Additionally, tf assigns equal weights to stylistic properties across the frequency spectrum in a language; therefore, it does not provide any form of feature weighing.

Another commonly used VSM is the $tf - idf$-model from Information Retrieval (Manning et al., 2008). The $tf - idf$ model extends the plain tf-model by weighing a word with its inverse document frequency (idf) in the collection. Thus, rare words that are present in only a few documents will be attached more importance. In many ways, this model can be contrasted with the assumption that low-frequency items are bad predictors of authorial style (Binongo, 2003). Nevertheless, a few studies suggest that it might be useful (Koppel & Winter, 2014). Arguably, this model captures the intuition that if a highly rare feature is present in two documents, this increases the likelihood that the two documents were authored by the same individual. While the method might therefore be sensitive to overfitting on low-frequency properties, this might be an attractive characteristic in certain (e.g. single-domain) authorship problems.

Thirdly, there is the std-model which weighs the tf-model through scaling term frequencies by their standard deviation across the document in the corpus. The model has initially been suggested by Burrows

(Burrows, 2002) as part of a memory-based learning system for authorship attribution and was later theoretically simplified (Argamon, 2008). A similar approach has been proposed in (Kešelj et al., 2003). This model captures the inverse intuition of the $tf-idf$ model, since it will boost the performance of very common items in a document collection, which will have a relatively low standard deviation in tf. This is highly uncommon in other applications in Information Sciences (e.g. document retrieval), although the model has been shown to work surprisingly well for authorship attribution in many studies (Stamatatos, 2009b).

3.3. Distance metrics

Both O1 and O2 crucially depend on distance metrics which can be applied to two vectors, in this case a vector representing an author's profile and a vector representing an unknown document. In authorship studies, it is a well known fact that the choice for a particular distance metric has a clear effect on the performance of systems (Evert et al., 2015), which is why distance metrics have continued to attract a lot of attention in authorship studies (Kešelj et al., 2003; Hoover, 2004; Stamatatos, 2007; Smith & Aldridge, 2011; Luyckx & Daelemans, 2011; Jockers et al., 2008; Evert et al., 2015). Previous studies have amply shown that specific metrics might behave and perform rather differently in different problem setups, stressing the fundamental *ad hoc* nature of many authorship problems (Juola, 2006; Evert et al., 2015). While many variations have been proposed, only a small set of metrics (or slight variations thereof) seem to have yielded consistent and good performance across studies. The traditional 'Manhattan' city block distance is a popular choice, which defines the difference between two documents as the sum of the absolute differences between all features. The city block distance predominantly works well for small and dense VSMs, with very limited vocabularies, such as small sets of function word frequencies. Cosine-based metrics are known to scale better to larger, sparse vectors, and they are therefore more common in Information Sciences (Manning et al., 2008). The cosine distance, for instance, is a pseudo-distance measure based on the complement (in positive space) of the angular cosine similarity between two document vectors.

In this paper, we will also compare these more established metrics to the still fairly novel $minmax$ measure (Koppel & Winter, 2014), originally introduced in geobotanics by M. Ružička (Ružička, 1958). While the metric has re-emerged a number of times in different disciplines (e.g. as the 'Jaccardized Czekanowski index' (Schubert & Telcs, 2014)), the method is only a recent addition to authorship studies. In mathematical notation, the minmax measure was originally formulated as the following similarity measure (Cha, 2007). Let a and b represent two document vectors, consisting of n features in some fixed order. Let a_i and b_i represent the value of the i-th feature in both documents respectively (e.g. the relative

frequencies of a particular word in both documents, in the case of the simple tf-model):

$$minmax(a,b) = \frac{\sum_{i=1}^{n} \min(a_i, b_i)}{\sum_{i=1}^{n} \max(a_i, b_i)} \tag{2}$$

We turn this similarity metric into a true distance measure by taking its complement in positive space (Schubert & Telcs, 2014): $1 - minmax(a,b)$. So far, $minmax$ has only been studied in the context of larger verification systems (Koppel & Seidman, 2013; Koppel & Winter, 2014; Seidman, 2013; Khonji & Iraqi, 2014), so that its individual contribution has not been clearly studied yet. More importantly, its performance has not rigorously been compared yet to other distance measures, under different experimental setups or in combination with different VSMs. In this paper, we will therefore elucidate the interplay of this distance metric and the VSMs described. In the context of the $tf - idf$ model, for instance, the $minmax$ metric will naturally boost the importance of features with larger values (i.e. those that are highly document-specific), whereas the opposite will happen in the std-model. We will empirically investigate the effect of this additional feature weighing.

4. Benchmark results

4.1. PAN data

To demonstrate the overall validity of our approach, we first benchmark O1 and O2 on 6 publicly available benchmark corpora which have been used in the 2014 edition of the PAN competition on authorship verification (Stamatatos et al., 2014) (`pan.webis.de`). In this yearly competition, teams can participate in a number of challenges involving forensic text analysis, such as plagiarism detection or authorship classification tasks. The organizers release training data that teams can independently develop systems on, before submitting their software. The organizers then run the software on new, unseen test data and rank the submitting teams according to their performance. We focus on the authorship verification track which has been organised since a number of years. The PAN 2014 verification datasets (see SI) only concern present-day writing samples, and vary strongly in both nature, size and difficulty, so that they provide a solid point of reference. The availability of the results reported by competitors on a fixed test set, moreover makes it easy to compare our results to the best performing systems which were entered into the competition. We report our full results in the SI and limit the discussion in the main text to a sample of illustrative examples. First, we calibrate O1 and and O2 on the development problems and then apply both systems to the test problems, reporting the AUC $\cdot$ c@1 for the test problems. In the SI, we report results for each combination of a VSM and distance metric, for the following feature

types: word unigrams, character trigrams, and character tetragrams. For each feature type, we used VSMs that represent full vocabularies. To assess whether O1 and O2 produce significantly different results, we have applied an approximate randomisation test to each pair of scores from O1 and O2. Table 1 gives a representative list of results in terms of AUC · c@1, namely the test results for using word unigrams in each corpus, for O1 and O2. For each problem set, we also list the performance of the best-performing individual system in that task, as well as the meta-classifier trained on all submitted systems (which often, but not always, yields the strongest overall result) (Stamatatos et al., 2014).

Table 1: A representative list of the main verification results on the PAN corpora in terms of AUC · c@1, namely the test results for using word unigrams in each corpus, for O1 and O2. For each problem set, we also list the performance of the best-performing individual system in that task, as well as the meta-classifier trained on all submitted systems (which often, but not always, yields the strongest overall result) (Stamatatos et al., 2014)

Combination	Dutch essays	Dutch reviews	English essays	English novels	Greek articles	Spanish articles
			O1			
cng - tf-std	76.89	31.95	23.94	22.26	28.79	59.54
cng - tf-idf	76.81	32.62	24.38	22.04	28.89	60.50
cng - tf	75.85	31.32	23.40	21.83	28.21	61.95
cosine - tf-std	67.51	27.91	23.61	50.22	42.65	57.20
cosine - tf-idf	61.41	27.50	16.58	33.68	44.33	50.71
cosine - tf	48.11	36.79	**28.77**	27.60	41.13	47.50
minmax - tf-std	71.32	34.66	25.46	45.14	54.98	45.67
minmax - tf-idf	**76.95**	**42.09**	24.22	**52.69**	59.33	45.63
minmax - tf	70.16	40.32	27.12	47.45	**67.47**	**76.98**
manhattan - tf-std	61.93	35.06	23.64	21.31	27.37	40.44
manhattan - tf-idf	71.44	37.08	24.33	43.44	33.52	68.51
manhattan - tf	76.35	34.59	23.92	40.61	42.80	67.83
			O2			
cng - tf-std	83.8	35.91	26.49	34.72	48.26	74.25
cng - tf-idf	81.70	36.55	27.95	35.47	48.83	73.10
cng - tf	80.26	35.99	27.25	35.80	50.90	80.54
cosine - tf-std	87.50	33.58	29.12	37.35	50.20	63.41
cosine - tf-idf	90.96	36.82	18.16	33.80	41.26	64.54
cosine - tf	76.59	36.58	24.95	30.55	48.63	69.52
minmax - tf-std	89.52	38.78	**35.13**	36.48	57.33	71.61
minmax - tf-idf	**93.70**	**48.42**	30.40	**40.66**	67.32	73.03
minmax - tf	87.44	38.90	30.50	36.93	**67.57**	**83.77**
manhattan - tf-std	47.96	33.05	25.31	22.65	27.99	32.94
manhattan - tf-idf	74.22	35.37	27.36	37.61	35.59	50.69
manhattan - tf	84.58	37.38	28.06	37.43	56.68	63.69
2014 Meta-classifier	86.70	42.80	53.10	50.80	72.00	70.90
2014 Best single system	82.30	52.50	51.30	47.60	63.50	69.80

A number of high-level trends emerge. The results immediately illustrate the large differences in overall difficulty which exist between the various data sets, ranging from the good scores which can be obtained for relative easy corpus of Dutch-language essays, to the more difficult corpus of English essays. Overall, O2 typically yields a higher performance than O1, although O1 produce the single highest scores for the English novels, where the length of documents is considerably longer than elsewhere. In two problem sets, the Dutch essays and Spanish articles, O2 and O1 respectively yield surprisingly strong results, even outperforming the meta-classifier and top-performing in the PAN competition. In the Dutch reviews and Greek articles, the performance of O2 can be characterised as very decent, with a performance between between the meta-classifier and that of the best performing individual system. Interestingly, both O1 and O2 perform relatively poorly for the following two data sets: the English essays and English novels (where text length clearly affects performance). With respect to the former corpus, we hypothesise that this loss in performance for O2 is due to the fact that we did not crawl the web for suitable imposters (as other studies have done), but limited our distractor pool to the other known documents in the problem set (because of our focus on Latin documents below). In these particular corpora, the algorithm might suffer from sampling documents that are too similar in content to the unknown document to act as a useful comparand. As to the other feature types, the results show that $manhattan$ only yields acceptable results for the character trigram features, which is an expected outcome, because character trigrams lead to a much denser corpus representation. For sparser representations, the minmax and cosine distance offer a much better fit. Especially in the case of word unigrams – which produce the strongest results across corpora – the novel minmax metric offers surprisingly strong results in comparison to the established metrics (it is part of every winning combination under O2). Interestingly, the effect of VSMs is much less pronounced than distance metrics: the minmax and cosine metric are generally least affected by a change in VSM.

4.2. Latin data

We now proceed to benchmarking our system on a corpus of historic Latin authors. For this study we have collected a representative reference corpus, containing works by some of the main Latin prose authors from Classical Antiquity, such as Cicero, Seneca or Suetonius. They predominantly include historiographical texts (e.g. Livy's *Ab Urbe Condita*) which are sufficiently similar to Caesar's War Commentaries. All original texts were cut up in non-overlapping slices of 1000 words; while this constitutes a challengingly limited document size, this procedure allows us to obtain a sufficiently fine-grained analysis of the Caesarian corpus. For modern documents, promising results are increasingly obtained with small document sizes (Koppel et al., 2013; Koppel & Winter, 2014), such as the PAN data used above.

To create a set of development and test problems, we proceed as follows. We split the available oeuvres at the author-level into two equal-sized sets. For each set we create a balanced set of same-author and different-author problems: for each true document-author pair, we also include a false document-author pair, whereby we randomly assign a different target author to the test document in question. This ensures that there is no overlap between the development and test problems created: therefore we can now parametrize the system on the development set and evaluate it on the test set, in an entirely parallel fashion as with the PAN data.

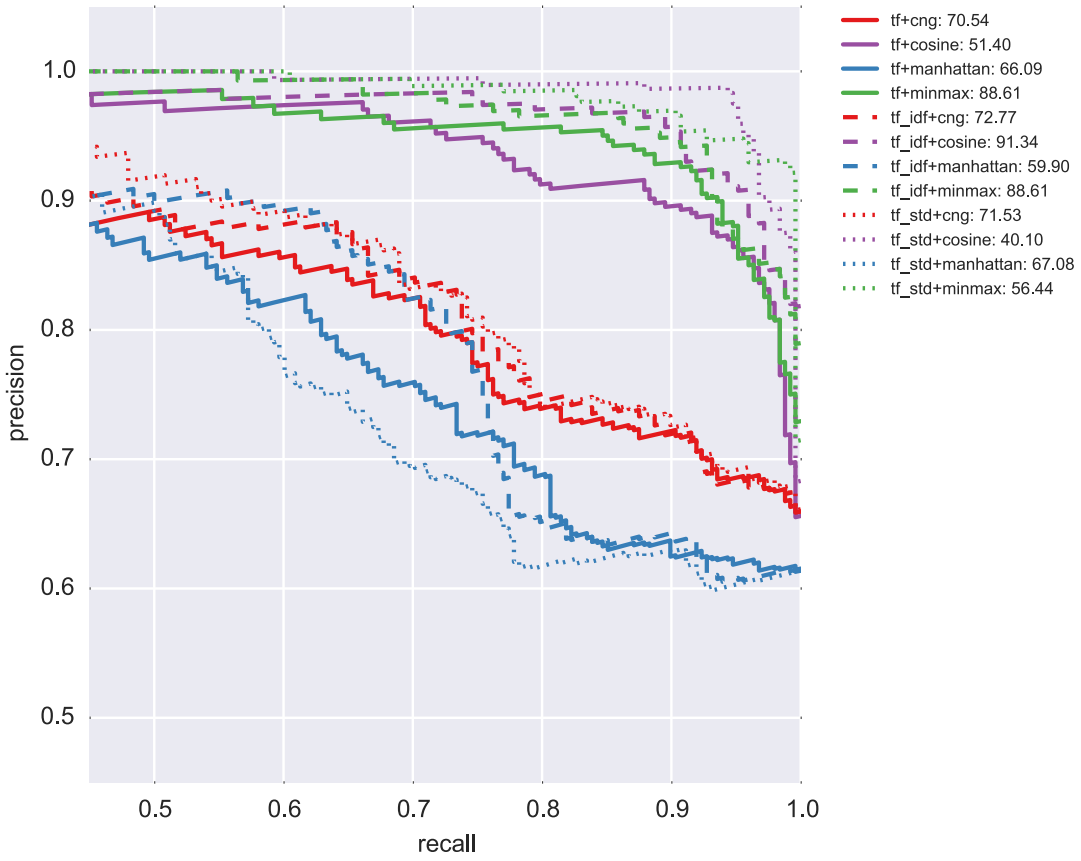

Figure 1: Precision-recall curves for each metric-VSM combination on the Latin benchmark data (test problems), using the O1 'first-order' verification system. The c@1 score is listed in the legend. The $cosine$ and $minmax$ metric consistently yield higher results than cng and $manhattan$.

In Figs. 1 and 2 we graphically show the results for O1 and O2 on the Latin benchmark corpus, again using untruncated vocabularies: for each combination of a VSM a distance metric, we plot a precision-

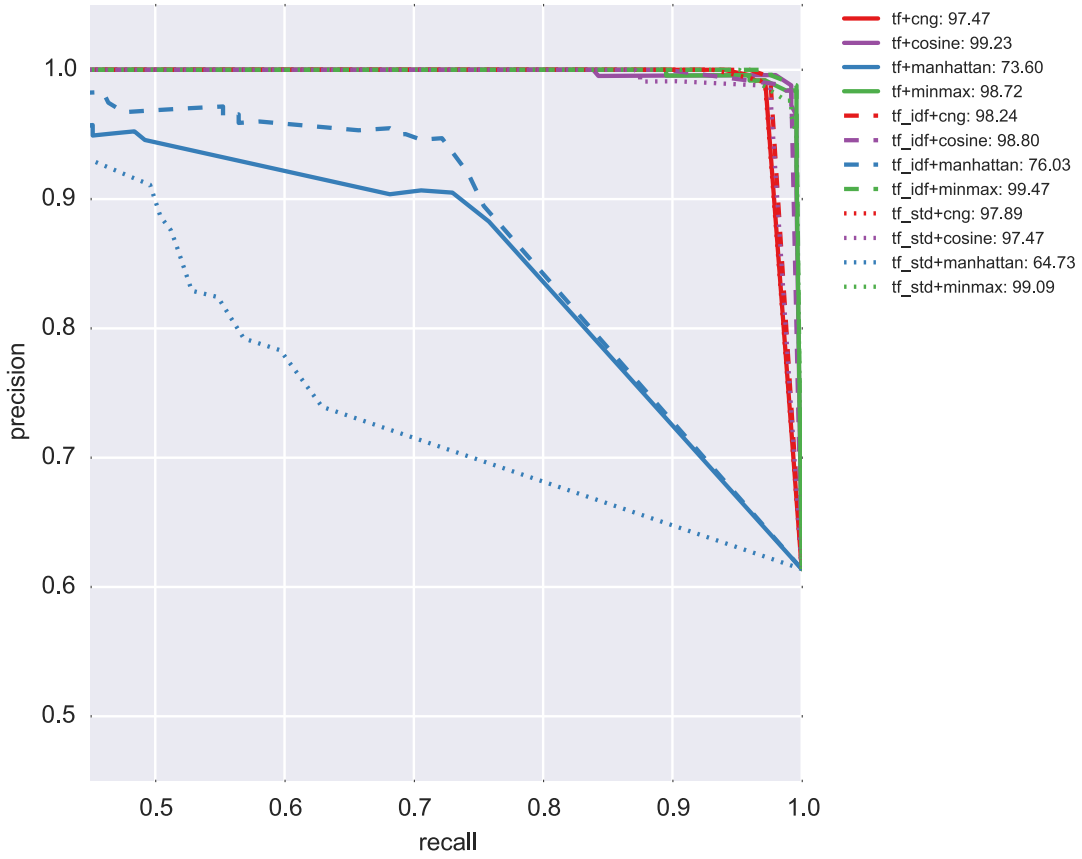

Figure 2: Precision-recall curves for each metric-VSM combination on the Latin benchmark data (test problems), using the O2 'second-order' verification system. The c@1 score is listed in the legend. Only the $manhattan$ distance now yields inferior results: the bootstrapping greatly reduces the variation between the different metric-VSM combinations.

recall curve; the c@1 score is listed in the legend (see SI for detailed results). The following trends clearly emerge: O2 consistently (in most cases significantly) outperforms O1 on the Latin data. O1 shows wildly diverging results, especially across different distance metrics, whereas the effect of VSMs is much less pronounced. In O2, both the $cosine$ distance and $minmax$ distance yield results that are clearly superior to cng and $cityblock$. Overall, O2 yields much stabler results across most combinations and for most combinations the curves can even not be visibly distinguished any longer. Unsurprisingly $cityblock$ is the only metric which yields visibly inferior results for O2. In O2 too, the minmax and cosine distance overall yield the highest c@1, which is invariable in the upper nineties. Our evaluation shows that the recently introduced $minmax$ metric yields a surprisingly good and consistent performance in comparison

to more established metrics. While it is not consistently the best performing metric, it produced highly stable results for the PAN data (and to a lesser extent for the Latin data). Overall, we hypothesize that the formulation of the $minmax$ metric has a regularizing effect in the context of authorship studies. Due to its specific formulation, the minmax metric will automatically produce distances in the 0-1 range, in contrast to the more extreme distances which can be produced by e.g. Manhattan. Perhaps because of this, the minmax metric interacts well with both std and $td - idf$, although these VSMs capture inverse intuitions. Like $cosine$, which also naturally scales distances, $minmax$ is relatively insensitive to the dimensionality of the VSM under which the metric is applied.

5. Caesar's writings

After benchmarking our verification systems, we now proceed to apply them to the Caesarian Corpus (*Corpus Caesarianum*), because it produced more stabler results for the benchmark data set (i.e. on average, it produced the highest results across different metric-vector space combinations). The Caesarian Corpus is composed of five commentaries describing Caesar's military campaigns (Mayer, 2011; Gaertner & Hausburg, 2013):

Gallic War *Bellum Gallicum*, conquest of Gaul, 58–50 BC;

Civil War *Bellum civile*, civil war with Pompey, 49–48 BC;

Alexandrian War *Bellum Alexandrinum*, Middle East campaigns, 48–47 BC;

African War *Bellum Africum*, war in North Africa, 47 to 46 BC

Spanish War *Bellum Hispaniense*, rebellion in Spain, 46–45 BC.

The first two commentaries are mainly by Caesar himself, the only exception being the final part of the *Gallic War* (Book 8), which is commonly attributed to Caesar's general Aulus Hirtius (*c*90 – 43 BC). Caesar's primary authorship of these two works, except for Book 8, is guaranteed by the ancient testimonia of Cicero, Hirtius, Suetonius, and Priscian as well as the unanimous evidence of the manuscript tradition. Caesar's ancient biographer Suetonius, writing a century and a half after his death, suggests that either Hirtius or another general, named Oppius, authored the remaining works: '[Caesar] also left *commentarii* of his deeds during the Gallic War and the Civil War with Pompey. For the author of the *Bellum Alexandrinum*, *Africum*, and *Hispaniense* is uncertain. Some think it is Oppius, others Hirtius, who supplemented the last, incomplete book of the *Bellum Gallicum*' (Appendix I). We also have a letter of Hirtius to Cornelius Balbus, a fellow supporter of Caesar, which is transmitted in the manuscripts

preceding the Hirtian 8th book of the *Gallic War*. In this letter, Hirtius lays out his project: 'I have continued the accounts of out Caesar on his deeds in Gall, since his earlier and later writings did not fit together, and I have also finished the most recent and incomplete account, extending it from the deeds in Alexandria down to the end, not admittedly of civil discord, of which we seen no end, but of Caesar's life' (Gaertner & Hausburg, 2013).

Despite occasional doubts, the most recent analysis has shown that there is no reason at all for doubting the authenticity of the letter (Gaertner & Hausburg, 2013). Hence, a puzzle that has persisted for nineteen centuries: what are the relationships of the different war commentaries to one another, to Hirtius, and to Caesar (Mayer, 2011)? Current scholarship has focused primarily on the authorship of the *Alexandrian War*. J. Gaertner and B. Hausburg (Gaertner & Hausburg, 2013) concluded that Hirtius knit together disparate sources to complete the text, including a Caesarian core of material in chapters 1–21. He also exercised a role in the formation of the whole corpus, though with much less firm editorial hand. Their analysis was based on a painstaking account of all sorts of evidence, including statistical analysis of usage and language. Their account represents the pinnacle of what can possibly by achieved by manual analytical methods, and offers a ripe target for re-analysis with automated computational methods. We are not the first to do so: in 2002 M. Trauth proposed a computer-assisted analysis of the *Corpus* which failed to reach any definitive conclusions on the authorship of the *Bellum Alexandrinum*, based on an automated tabulation of the most frequent words. (Trauth, 2002). More than a decade of advances in computational philology allow us to go beyond his inconclusive analysis.

To shed new light on the authenticity of the Caesarian corpus, we proceed as follows. To obtain documents of a similar size, we have divided all original commentaries in consecutive, non-overlapping slices of 1000 words and treat these slices as individual documents. We label these documents according to the assumption that the Gallic and Civil Wars were written by CAESAR, with the exception of 8th book of the former commentary, which we ascribe to HIRTIUS. To label the disputed authors of the Alexandrian, African and Spanish War, we use the provisional labels X, Y and Z respectively. Fig. 3 offers an initial inspection of the stylistic structure in this corpus, in the spirit of the first-order distance-calculations of O1. We generated a square distance table using the minmax distance metric to every document pair in the Caesarian collection and we scaled the distances to the 0–1 range. Next, we plotted a heat map of the distance matrix, and ran a conventional cluster analysis on top of the rows and columns. For the generating the hierarchical dendrograms next to the heatmap, we used the default agglomerative clustering routine in the *Seaborn* library (`https://web.stanford.edu/~mwaskom/software/seaborn/`), which is based on the pairwise Euclidean distance between entries and the average linkage method. The labels indicate the most plausible authorial provenance of a document (if known), given the annotation labels

we just described.

This rather naive approach demonstrate a clear-cut distinction: a significant portion of the *Bellum Alexandrinum* (X) clusters with Hirtius's contribution to the *Gallic Wars*, under a clade that is clearly separate from Caesar's accepted writings. Thus, Hirtius's writings are distinguished from Caesar's own core contributions; Hirtius's samples are compellingly close in style to X. Samples from the *Alexandrian War* appear to be stylistically close to Hirtius's contribution to the *Gallic Wars* in Book 8 – which itself is surprisingly distinct from the other chapters in it. The more fundamental question now is *how* close these texts should truly be, in order to proceed to an actual attribution. We therefore turn to a more advanced analysis using O2. As with the problems in the benchmark experiments, each sample in the commentary collection was individually paired with the profile of all five Caesarian 'authors' available (including X, Y and Z): using the bootstrapped procedure from O2, we calculate a second-order similarity score by assessing in which proportion of a series of iterations one of these documents would be attributed to a particular Caesarian author's profile, instead of a distractor author in the background corpus. This procedure as such yields, per document, 5 second-order scores, reflecting the probability that the sample must be attributed to a Caesarian's authors profile, rather than an imposter. Following the outcome of the benchmark results, we perform this analysis for the five top-scoring metric-VSM combinations. Afterwards, we average the results over these five simulations and we graphically present the results in Fig. 4 (the full results are included in the SI). Note that in this setup we are especially interested in attribution *leakage* from one potential author to another: the fact that a text is attributed to the profile based on the other samples from its own text is an expected result; the attribution to another Caesarian 'author', however, is not.

Our O2 analyses divide the Caesarian corpus into two branches at the top-level, which might be called 'Caesarian' and 'non-Caesarian'. As we would expect, the Caesarian branch includes both the *Civil War* and the *Gallic War*, books 1–7. However, it also includes the first three samples from the *Alexandrian War*, providing dramatic confirmation of the theory of a Caesarian core in the first 21 chapters of the work. The other branch includes *Gallic War*, book 8, the rest of the *Alexandrian War*, the *African War*, and the *Spanish War*. The first two are closely affiliated with one another, indicating shared authorship. Stylistically there is no good reason for rejecting Hirtius's authorship of the *Alexandrian War*, once we remove the Caesarian chapters 1–21. Gaertner and Hausburg (Gaertner & Hausburg, 2013) argue strongly against Hirtius's authorship of the *Alexandrian War*, instead assigning him an amorphous role as editor of the corpus. It is true that the *Alexandrian War* shows far great heterogeneity that the *Spanish War*, for example, but it clearly clusters with the *Gallic War*, book 8, in a way the other texts do not, and displays no greater stylistic heterogeneity than Caesar's own commentaries.

The *African War* and the *Spanish War* are the most internally consistent of the texts, perhaps an indication of separate authorship. They do, however, cluster with one another and with Hirtius, and the non-Caesarian texts all show a greater similarity with each other than with the Caesarian texts. While they are not stylistically homogenous enough to allow us to positive single-authorship in a naive sense, they display no greater stylistic heterogeneity than is present in the Caesarian texts. On both branches, we find the stylistic range we ought to expect in the genre of war commentaries, where commanders drawing up the official account of their campaigns would draw upon the dispatches of their legates and subordinates, sometimes integrating them into their own style, other times incorporating their texts with few changes. Importantly, Fig. 4 has an additional feature: whereas other x samples could be found scattered across Caesar's authentic writings in the non-bootstrapped verification, O2 adds a distinct clade for these and a small set of other samples. This is a strong indication that the bootstrapped O2 system is not only able to distinguish authentic Caesarian material from non-authentic writings, but that it can even differentiate between a pure Caesarian style from the impure style resulting from collaborative authorship or the use of source texts. Hence, our analyses broadly supports the following conclusions:

1. Caesar himself wrote, in addition to *Gallic Wars*, books 1–7 and the *Civil War*, as well as the first 21 chapters of the *Alexandrian War*.
2. Hirtius wrote Book 8 of the *Gallic Wars* and the remainder of the *Alexandrian War*.
3. At least one other author wrote the *African War* and the *Spanish War*. The *African War* and the *Spanish War* were probably written by two different authors.
4. Our results do not invalidate Hirtius's own claim that he himself compiled and edited the corpus of the non-Caesarian commentaries.
5. The significant stylistic heterogeneity we have detected in parts of the *Gallic War* and the *Civil War* likely represents Caesar's compositional practice of relying on, and sometimes incorporating, the briefs written for him by his legates.

These findings are entirely consistent with a natural interpretation of Hirtius's own words in his letter to Balbus, that he composed *Gallic War*, book 8 as a bridge between the preceding 7 books and the Civil War, that he completed the *Alexandrian War*, and added the two other commentaries to make the whole group a continuous narrative of Caesar's campaigns. Chronologically the corpus thus ends in March, 45 BC with the Battle of Munda in Spain, but since we know that the end of the *Spanish War* is missing, there is no reason why we cannot assume that it originally continued with a brief epilogue bringing the narrative up to conclude with Caesar's assassination in 44 BC.

6. Acknowledgements

The authors would like to thank [anonymized] for their valuable feedback on earlier drafts of this article. Moshe Koppel acknowledges the support of the Intel Collaboration Research Institute for Computational Intelligence. The work of Folgert Karsdorp has been supported by the Computational Humanities Programme of the Royal Netherlands Academy of Arts and Sciences, under the auspices of the Tunes & Tales project.

7. References

Argamon, S. (2008). Interpreting Burrows's Delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, *23*, 131–147.

Argamon, S., & Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. In *Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing (2005)*.

Aronoff, M., & Fudeman, K. (2005). *What is Morphology?*. Blackwell.

Barthes, R. (1968). La mort de l'auteur. *Manteia*, *5*, 12–17.

Binongo, J. (2003). Who Wrote the 15th Book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, (pp. 9–17).

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, *17*, 267–287.

Burrows, J. (2007). All the way through: Testing for authorship in different frequency strata. *Literary and Linguistic Computing*, *22*, 27–47. URL: `http://llc.oxfordjournals.org/content/22/1/27.abstract`. doi:`10.1093/llc/fqi067`. `arXiv:http://llc.oxfordjournals.org/content/22/1/27.full.pdf+html`.

Burrows, J. (2012). A second opinion on Shakespeare and Authorship Studies in the Twenty-First Century. *Shakespeare Quarterly*, *63*, 355–392.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, *1*, 300–307.

Chaski, C. E. (2005). Who's at the keyboard? authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, *4*, 1–13.

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology*, *52*, 558–569. URL: http://dx.doi.org/10.1002/asi.1097. doi:10.1002/asi.1097.

Daelemans, W. (2013). Explanation in computational stylometry. In *Computational Linguistics and Intelligent Text Processing* (pp. 451–462). Springer.

Daelemans, W., & Van den Bosch, A. (2005). *Memory-Based Language Processing*. Studies in Natural Language Processing. Oxford University Press.

Eder, M. (2012). Computational stylistics and biblical translation: how reliable can a dendrogram be? In T. Piotrowski, & Ł. Grabowski (Eds.), *The Translator and the Computer* (pp. 155–170). Wrocław: WSF Press.

Efstathios, S. (2013). On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, *21*, 421–439.

Evert, S., Proisl, T., Schöch, C., Jannidis, F., Pielström, S., & Vitt, T. (2015). Explaining delta, or: How do distance measures for authorship attribution work? URL: http://dx.doi.org/10.5281/zenodo.18308. doi:10.5281/zenodo.18308.

Foucault, M. (1969). Qu'est-ce qu'un auteur? *Bulletin de la Société française de philosophie*, *3*, 73–104.

Gaertner, J., & Hausburg, B. (2013). *Caesar and the Bellum Alexandrinum: An Analysis of Style, Narrative Technique, and the Reception of Greek Historiography*. Vandenhoeck & Ruprecht.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, *22*, 251–270. URL: http://llc.oxfordjournals.org/content/22/3/251.abstract. doi:10.1093/llc/fqm020.

Halteren, H. V., Baayen, H., Tweedie, F., Haverkort, M., & Neijt, A. (2005). New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, *12*, 65–77.

Hermann, J., Oskam K., V. D., & Schöch, C. (2015). Revisiting style, a key concept in literary studies. *Journal of Literary Theory*, *9*, 25–52.

Holmes, D. (1994). Authorship attribution. *Computers and the Humanities*, *28*, 87–106.

Holmes, D. (1998). The evolution of stylometry in Humanities scholarship. *Literary and Linguistic Computing*, *13*, 111–117.

Hoover, D. L. (2004). Testing Burrows's Delta. *Literary and Linguistic Computing*, *19*, 453–475. doi:10.1093/llc/19.4.453.

Jockers, M. L., Witten, D. M., & Criddle, C. S. (2008). Reassessing authorship of the Book of Mormon using delta and nearest shrunken centroid classification. *Literary and Linguistic Computing*, *23*, 465–491.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, *1*, 233–334.

Juola, P. (2013). Rowling and Galbraith: an authorial analysis. URL: http://languagelog.ldc.upenn.edu/nll/?p=5315.

Juola, P. (2015). The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions. *Digital Scholarship in the Humanities*, . doi:10.1093/llc/fqv040.

Juola, P., & Stamatatos, E. (2013). Overview of the author identification task at PAN 2013. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013.*.

Kešelj, V., Peng, F., Cercone, N., & Thomas, C. (2003). N-gram-based author profiles for authorship attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03* (pp. 255–264). Dalhousie University, Halifax, Nova Scotia, Canada.

Kestemont, M. (2014). Function words in authorship attribution. from black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature* (pp. 59–66). Association for Computational Linguistics.

Kestemont, M., Luyckx, K., & Daelemans, W. (2011). Intrinsic plagiarism detection using character trigram distance scores - notebook for PAN at CLEF 2011. In *CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands*. URL: http://ceur-ws.org/Vol-1177/CLEF2011wn-PAN-KestemontEt2011.pdf.

Kestemont, M., Luyckx, K., Daelemans, W., & Crombez, T. (2012). Cross-genre authorship verification using unmasking. *English Studies*, *93*, 340–356.

Khonji, M., & Iraqi, Y. (2014). A Slightly-modified GI-based Author-verifier with Lots of Features (ASGALF). In *Working Notes for CLEF 2014 Conference, Sheffield, UK* (pp. 977–983).

Kjell, B. (1994). Discrimination of authorship using visualization. *Information Processing and Management*, *30*, 141–50.

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, *60*, 9–26.

Koppel, M., Schler, J., & Argamon, S. (2013). Authorship attribution: What's easy and what's hard? *Journal of Law & Policy*, *21*, 317–331.

Koppel, M., & Seidman, S. (2013). Automatically identifying pseudepigraphic texts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1449–1454). Association for Computational Linguistics.

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

Love, H. (2002). *Attributing authorship. An introduction*. Cambridge: Cambridge University Press.

Luyckx, K., & Daelemans, W. (2011). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, *26*, 35–55.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

Maurer, H., Kappe, F., & Zaka, B. (2006). Plagiarism - A Survey. *Journal of Universal Computer Science*, *12*, 1050–1084.

Mayer, M. (2011). Caesar and the corpus caesarianum. In G. Marasco (Ed.), *Political autobiographies and memoirs in antiquity: A Brill companion* (pp. 189–232). Brill.

Mosteller, F., & Wallace, D. (1964). *Inference and disputed authorship: The Federalist*. Addison-Wesley.

Peñas, A., & Rodrigo, A. (2011). A simple measure to assess non-response. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* HLT '11 (pp. 1415–1424). Stroudsburg, PA, USA: Association for Computational Linguistics. URL: `http://dl.acm.org/citation.cfm?id=2002472.2002646`.

Peng, F., Schuurmans, D., Wang, S., & Keselj, V. (2003). Language independent authorship attribution using character level language models. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1* EACL '03 (pp. 267–274). Stroudsburg, PA, USA: Association for Computational Linguistics. URL: http://dx.doi.org/10.3115/1067807.1067843. doi:10.3115/1067807.1067843.

Potha, N., & Stamatatos, E. (2014). A profile-based method for authorship verification. In A. Likas, K. Blekas, & D. Kalles (Eds.), *Artificial Intelligence: Methods and Applications* (pp. 313–326). Springer International Publishing volume 8445 of *Lecture Notes in Computer Science*. URL: http://dx.doi.org/10.1007/978-3-319-07064-3_25. doi:10.1007/978-3-319-07064-3_25.

Puig, X., Font, M., & Ginebra, J. (2016). A unified approach to authorship attribution and verification. *The American Statistician*, *advance access*, x. URL: http://dx.doi.org/10.1080/00031305.2016.1148630. doi:10.1080/00031305.2016.1148630. arXiv:http://dx.doi.org/10.1080/00031305.2016.1148630.

Ružička, M. (1958). Anwendung mathematisch-statistischer methoden in der geobotanik (synthetische bearbeitung von aufnahmen). *Biológia (Bratislava)*, *13*, 647–661.

Rybicki, J., & Eder, M. (2011). Deeper Delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing*, (pp. 315–321).

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*, 513–523.

Sapkota, U., Bethard, S., Montes, M., & Solorio, T. (2015). Not all character n-grams are created equal: A study in authorship attribution. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 93–102). Denver, Colorado: Association for Computational Linguistics. URL: http://www.aclweb.org/anthology/N15-1010.

Sapkota, U., Solorio, T., Montes-y-Gómez, M., Bethard, S., & Rosso, P. (2014). Cross-topic authorship attribution: Will out-of-topic data help? In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland* (pp. 1228–1237). URL: http://aclweb.org/anthology/C/C14/C14-1116.pdf.

Schubert, A., & Telcs, A. (2014). A note on the Jaccardized Czekanowski similarity index. *Scientometrics*, *98*, 1397–1399. doi:10.1007/s11192-013-1044-2.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*, 1–47.

Seidman, S. (2013). Authorship verification using the impostors method. In *CLEF 2013 Evaluation Labs and Workshop-Online Working Notes*.

Seroussi, Y., Zukerman, I., & Bohnert, F. (2014). Authorship attribution with topic models. *Comput. Linguist.*, *40*, 269–310. URL: http://dx.doi.org/10.1162/COLI_a_00173. doi:10.1162/COLI_a_00173.

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernndez, L. (2014). Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, *41*, 853 – 860. URL: http://www.sciencedirect.com/science/article/pii/S0957417413006271. doi:http://dx.doi.org/10.1016/j.eswa.2013.08.015. Methods and Applications of Artificial and Computational Intelligence.

Smith, P. W. H., & Aldridge, W. (2011). Improving authorship attribution: Optimizing Burrows' Delta method. *Journal of Quantitative Linguistics*, *18*, 63–88.

Stamatatos, E. (2006). Authorship attribution based on feature set subspacing ensembles. *International Journal on. Artificial Intelligence tools*, *15*, 823–838.

Stamatatos, E. (2007). Author identification using imbalanced and limited training texts. In *Proceedings of the 18th International Conference on Database and Expert Systems Applications* DEXA '07 (pp. 237–241). Washington, DC, USA: IEEE Computer Society. URL: http://dx.doi.org/10.1109/DEXA.2007.41. doi:10.1109/DEXA.2007.41.

Stamatatos, E. (2009a). Intrinsic plagiarism detection using character n-gram profiles. In *Third PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse* (pp. 38–46).

Stamatatos, E. (2009b). A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology*, *60*, 538–556.

Stamatatos, E., Daelemans, W., Verhoeven, B., Stein, B., Potthast, M., Juola, P., Sánchez-Pérez, M. A., & Barrón-Cedeño, A. (2014). Overview of the author identification task at PAN 2014. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014.* (pp. 877–897).

Stamatatos, E., Kokkinakis, G., & Fakotakis, N. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, *26*, 471–495.

Stein, B., Lipka, N., & Prettenhofer, P. (2011). Intrinsic plagiarism analysis. *Language Resources and Evaluation*, *45*, 63–82.

Stover, J., Winter, Y., Koppel, M., & Kestemont, M. (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*, 239–242.

Trauth, M. (2002). Caesar incertus auctor. Ein quantifizierendes Wort zur Kritik von Verfassersfragen in Lateinischen Texten. In J. Jährling, U. Meves, & E. Timm (Eds.), *Röllwagenbüchlein. Festschrift Walter Röll* (pp. 313–334). Niemeyer.

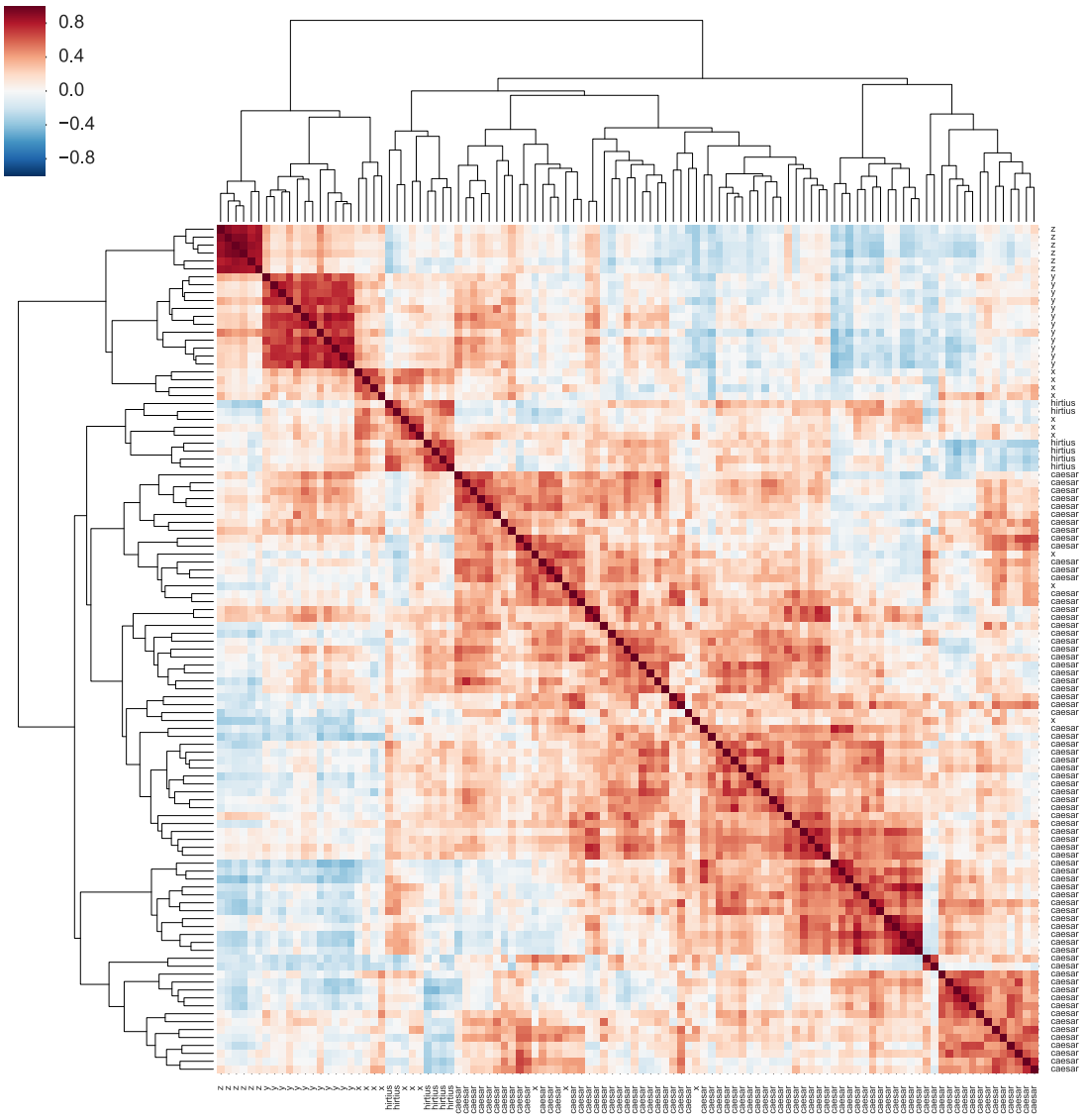

Figure 3: Naive heatmap visualisation of the stylistic structure in the *Corpus Caesarianum*, based on the scaled, pairwise distance matrix on the basis of the first-order $minmax$ distance metric and the tf VSM (full vocabulary). Conventional clustering was ran on top of rows and columns, representing non-overlapping 1000-word samples from the text. A significant portion of the *Bellum Alexandrinum* (labeled x) clusters with Hirtius's contribution to the *Gallic Wars*, under a clade that is separate from Caesar's accepted writings.

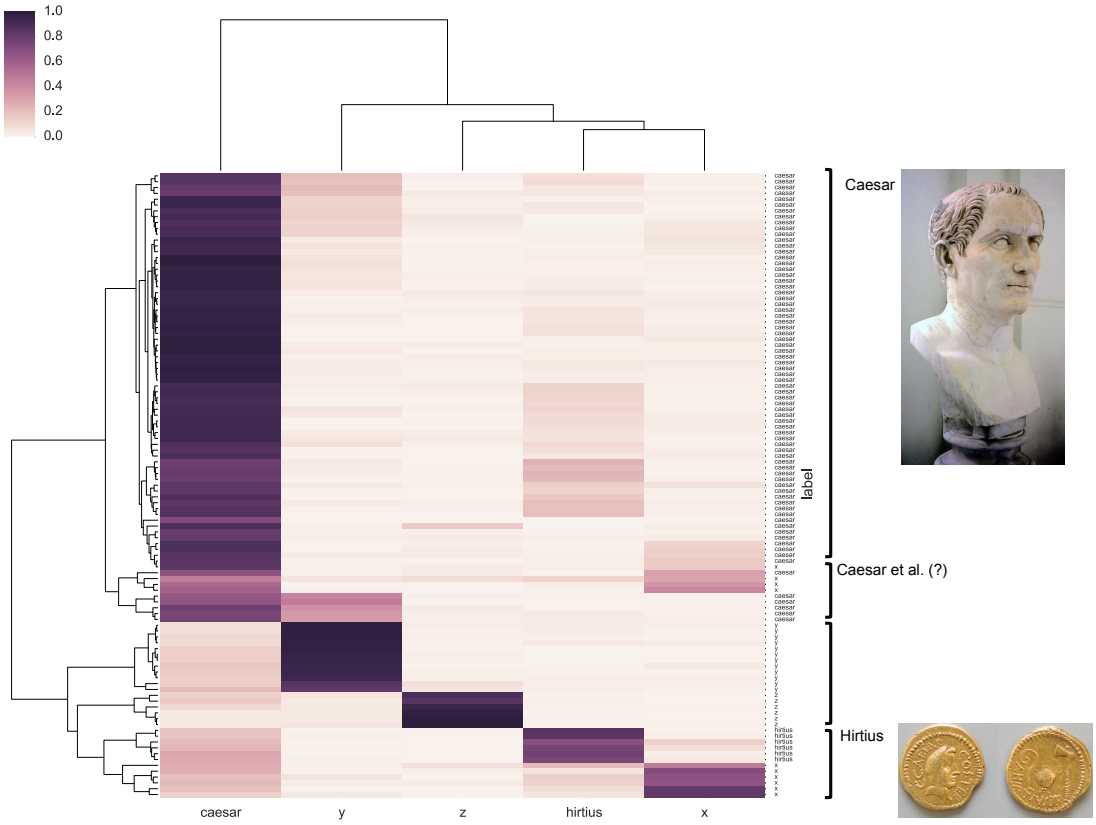

Figure 4: Cluster and heatmap visualisation of the results of the O2 verification procedure on the Caesarian corpus. Cell values represent the average probability of a sample being attributed to one of the five profiles distinguished. Five independent analysis were run with the 5 top-performing metric-VSM combination in the benchmark section. O2 seems not only able to distinguish authentic Caesarian material from non-authentic writings, but arguably also differentiates between a 'pure' Caesarian style and the mixed style resulting from e.g. the general's dependence on pre-existing briefs by legates.

Point-to-Point responses to the reviewers

We would like to thank the reviewers for their very useful comments. Below we respond to their reviews point-by-point.

Reviewer #1: The manuscript is very understandable. It covers most of the literature related to the topic. After testing different methods (combinations of distance metrics and Space Vector Models) the authors apply the chosen method to a real case study. As a result the manuscript becomes an interesting work.
> *Thank you for your kind words.*

Some specific/minor comments:

- In the first paragraph of section 3 (methods), the authors distinguish between authorship attribution and authorship verification problem as it is usually done in literature until now. But in February 2016 an interesting report was published in the American Statistician journal where a unified Bayesian approach to analyze both authorship attribution and verification problem is presented. This novel point of view could open new lines of research. (http://www.tandfonline.com/doi/abs/10.1080/00031305.2016.1148630)
> *We were unaware of this recently published reference and gladly include a mention in the paper.*

- In line 329 the table's number is missing. And in page 14 the table's caption is missing too.
> *We have corrected/added this information.*

- The numbers in the Table on the manuscript (page 14) do not coincide with the numbers of the tables that appear in the supplementary material. For example, why is the first column in the table of the manuscript different from the last column of the table 1 in section 3.1 of SI? Shouldn't they be the same?
> *Excellent remark. The results in the main paper were indeed performed using a different random initialization to show the stability of the results, but we agree that this is confusing and we have removed the discrepancy. Now, the tables perfectly match.*

- The results showed in section 4.1 and 4.2 suggest an interesting fact. It is that analyzing Latin data the minimax distance runs quite satisfactory and when the Spanish articles are analyzed the minimax distance runs quite satisfactory too. Observe that the Spanish language descends from Latin, therefore it could be a future line to evaluate if the good performance of this distance also applies other Latin-based languages, such as Italian, French, Catalan or Portuguese.
> *Interesting point, although the same could be said for the Greek articles. What is especially remarkable is that the effect of the feature representation seems less important than the effect of the distance metric used, although one expect character n-grams to work relatively better for more highly-inflected languages.*

Reviewer #2:

Review

The topic of this paper is interessting and suitable for the journal, despite the fact that
there are many papers circulating in the field of Authorship Verification.
However, there a number of issues I really would like to comment on.
> *Thank you for your kind words. We agree that there are a lot of (questionable) papers circulating in this subfield.*

Abstract:

The paper deals with the authenticity of the "Caesarianum" corpus,
which coprises of five commentaries describing the campaigns of Julius Caesar.
To achieve this goal the authors use two state-of-the-art authorship verification (AV) methods,
which were evaluated on the publicly available corpora "PAN-2014" as well as the "Latin benchmark dataset".
In order to run their AV analysis the authors use word 1-grams and character n-grams.

The authors come to the conclusion that regarding Caesar's writings, Hirtius's claims to be part of the corpus "must" be considered legitimate.

---
2 Style vs. Content:

The authors discuss (in more than a paragraph) the usefulness of function words.
In particular, they say that function words are not linked the topic or genre, which is also my opinion. However, I absolutley cannot understand why the authors stick to the usage of features (namely word unigrams and character n-grams), which are known to be affected by topic and genre. Yes it's true, in many studies these features have been used (more or less) successfully, but using these features just to avoid preprocessing techniques such as pos-taggers is not an argument. Function words occur in any language and can be extracted (and also looked up) from any text easily and efficiently. So why the authors don't use them?
*> Thank you for raising this issue. We should probably stress that function words *are* included in this study, but of course as a part of the whole ngram-vocabulary. Note that in our discussion we stress how some the metric-space combinations will automatically boost the importance of function words, because of their higher frequency, according to which VSM is being used. More importantly, the use of function words have been primarily useful in the study of longer texts (which we don't work with, except for the English novels corpus). Function words, in our understanding will primarily be useful when we have larger portions of text from which we can extract reliable frequency statistics. For the kind of short texts we work with, we deem it highly unlikely that restricting the analyses to function words only would yield an increase in the verification accuracy.*

---
3 Methods:

The authors claim that both AV-methods (O1 = Keselj-method & O2 = Imposters-method) are state-of-the-art. It might be true, but however --> It would be very helpful for the reader to understand WHY both AV-methods are considered state-of-the-art (i.e. due to the high citation count). I would recommend to make this clear...
> We base this observation primarily on the results of the latest PAN competitions and we have added a sentence to clarify this.

One very important question which remains unanswered for me (after going trough the paper seveal times) regarding both AV-methods is which parameters/configurations/settings were used and why these are not given in the paper? These are necessary in order to reproduce the results/experiments of the authors on the used corpora collections. It seems for me that both AV-methods were used in a black-box manner without training them on the same training set. With training I mean finding promising parameters such as the underlying distance functions, the [n] parameter (regarding the character n-grams), the number of extraxted features, etc.
*> As we explicitly note in the paper, previous papers by Koppel et al. have shown that the O2 method is largely insensitive to its exact parametrization (e.g. the number of imposter does not play a significant role as long as the imposter pool is large enough). Therefore, we limit our parameters to the ones recommended in previous research and we clearly report these. We agree that vocabulary size plays a minor role: in our initial experiments we have amply experimented with the vocabulary size but the results were not interesting enough to report because they confirmed what we already noted in the presentation of the methods (i.e. Manhattan distance works better for denser vectors). We therefore decided to leave out an explicit discussion about this, also because the sampling process in O2 makes it more difficult to compare O1 to O2 when varying the vocabulary size. Finally, note that both systems *were* trained on the same data, because otherwise the comparison would be false…*

By the way, a question for the interested reader: What exactly is the "PAN competition"?
(of course I as a reviewer know it..) The authors should at least mention a url* to this competition in a footnote within the paper. *(I would recommend: http://pan.webis.de).
I believe that not everyone is familiar with this competition.
> We have included the URL and added some more information.

A couple of questions regarding the following statements:

"Authorship verification is a problem which has been studied for a number of years in the annual PAN competition"
--> And even more years outside of the PAN competition...
> *Evidently, but this paragraph deals with the PAN data, hence the focus.*

"the AUC score (area under the curve, a well-known scalar evaluation score for binary classifiers) and a competition-specific c@1 score."
--> Actually, the c@1 score has nothing to do with PAN, it was "developed" in 2011 by Peñas & Rodrigo:
http://www.aclweb.org/anthology/P11-1142.pdfion.html
> *Correct, we changed the wording and added the reference.*

"we define an author's profile here as the mean centroid of the known document vectors for that author" --> Can you give a formal representation of this "mean centroid" ? I would also like to know if (and how) you normalize the features inside the feature vector (i.e. which normalization technique do you use?).
> *To clarify this, we changed the wording to "we define an author's profile here as the mean centroid of the known document vectors for that author (i.e. we average an author's score for a particular term across all training texts)". As to the normalization, we clearly state in the appendix that "all models are scaled to unit norm" (on top, of course, of the actual vectorization model).*

"Originally, O1 was introduced with a specific distance metric, called 'common n-grams' (cng)."
--> According to "Encyclopedia of Distances" (http://www.springer.com/us/book/9783642002342) there is no distance metric with the name "common n-grams". I think you mean the name of the specific AV-method rather then the distance metric?
> *It is true that this is not an actual distance metric (the name indeed refers to the features used, instead of the metric), but this is how the metric was originally defined (and is subsequently referred to in the literature), so we would rather stick to this name for clarity.*

"Let ai and bi represent the value of the i-th feature in both documents respectively:"
--> What do you mean by "value"...? The frequency? If yes which one, absolut or relative frequency ?
> *Here, it is important to state that this depends on the vector space model used: in case of tf, it is a 'relative frequency', but in case of tf-idf it is a 'tf-idf score'. Also, the models have been normalized, so we would really prefer the term 'value' here, since it is more generic and is true for all models, instead of just tf.*

"The original paper would calculate this distance function only for character n-grams
which were present in both the profile and the unknown document"
--> Actually, Stamatatos and Potha say in "A Profile-Based Method for Authorship Verification" that it is a symmetrical dissimilarity function, rather than a distance function.
> *Yes, the implementation differs between papers (which can be confusing) and our wording was ambiguous. We clarified which implementation we use (and why).*

"Subsequent research showed that it is beneficial to apply the distance function
only to the items which are present in the unknown document"
---> The original source for this claim is not stated in "A Profile-Based Method for Authorship Verification"
but in "Author Identification Using Imbalanced and Limited Training Texts" (Stamatatos).
> *See previous remark. We corrected the reference (thanks for spotting this!).*

"The general intuition behind the GI, is not to assess whether two documents are simply similar,"
--> Similar in terms of what? Style or content?
> *In writing style. We clarified this.*

Last paragraph of 3.1 --> So, how exactly do you define the imposter pool? You say:
"Previous studies have automatically crawled the web for useful imposter documents,
which yields results that might be difficult to reproduce exactly"
--> Do you also crawl random texts from the internet that match in terms of genre/topic?
> *This is clearly stated: 'In this paper, we sample imposter authors from the known documents that are available for other authors in a particular problem set. To ensure the comparability of O1 and O2, we sample author profiles (i.e. mean centroids), instead of individual documents from the imposter pool. Previous studies have automatically crawled the web for useful imposter*

documents, which yields results that might be difficult to reproduce exactly. Additionally, there is the inherent danger that one might obtain imposter documents that were indeed written by the target author, which would compromise the proper working of O2. Naturally, this problem is even more real in the case of the Latin data sets used below, because of the relatively sparse online availability of Latin documents from Classical Antiquity.'.

3.3 --> You describe that there are different distance functions where the choice of one has a specific influence regarding the performance. OK, but why? What causes a distance function d1 to perform better than a distance function d2? What happens excatly in the feature space when we "play" with different distance functions?
> As to the minmax function (which is the main focus), a number of reasons are hinted at in the discussion. Also see the paragraph mentioned in the next point.

---
5 Caesar's Writings:

Just a couple of questions regarding some statements:

"because it produced more stabler results (...)"
--> Do you have any explanition why the results where more stabler?
> We added a paragraph describing our intuitions.

"The first two commentaries are mainly by Caesar himself (...)"
--> Who is saying this? What is the ground truth?
> This claim is based on traditional scholarship and we have found no reason whatsoever to doubt this claim. True ground truth is of course non-existent when it comes to historical data… We added: ""Caesar's authorship of these two works is guaranteed by the ancient testimonia of Cicero, Hirtius, Suetonius, and Priscian as well as the unanimous evidence of the manuscript tradition."

"Trauth proposed a computer-assisted analysis of the Corpus which failed to reach any definitive conclusions on the authorship of the Bellum Alexandrinum"
---> I don't understand, is his approach manual, semi-automated or fully-automated? Clarify this in a short sentence.
> We added a clarification.

"To obtain documents of a similar size, we have divided all original commentaries in consecutive, non-overlapping slices of 1000 words and treat these slices as individual documents."
--> So you split documents into fragments that share the same genre as well as the same topic ? I don't think this contribute much to the generalisability of both AV-methods that you use. I would even say that the performed verification does not take the writing style but instead the topic influence into account.
*> We beg to differ in this respect: first of all, our method obtains state-of-the-art results on the (cross-topic) PAN datasets, showing that we are indeed modelling writing style, instead of *just* content. Secondly, the samples are all about the same topic (they are all are highly similar war commentaries), which makes it even more likely that we are modelling style instead of topic. Most importantly, if we were indeed modelling topic instead of style, this could never explain why the final book of the Gallic war clusters with the x-samples, which are from a completely different commentary. In this respect, shared authorship is clearly the most likely explanation given our knowledge about these texts.*

"We label these documents according to the assumption that the Gallic and Civil Wars were written by CAESAR"
--> So your data (more precisely the "ground truth") is labeled based on an assumption regarding the true authorship? Just think about this...
> As mentioned above, we have no reason to doubt this attribution based on external evidence.

"Hirtius's samples are compellingly close in style to X."
--> I would say "close in terms of topical content", rather than style.
Reason: Again, the features you are using in your analysis ar highly inflected by topic.
> See above. This is highly unlikely.

"and ran a conventional cluster analysis on top of the rows and columns."

--> Could you elaborate more on this "conventional cluster analysis" that you have performed? Was it k-means, k-means++, DB-Scan, or what else? Which parameters have been used? Which features? Important details are missing here...
> Fair point, we added all necessary details.

Highlights:

2.) I would replace "must have" with "might have" as I don't believe that
any AV-method is able to guarantee true authorships.
> That would be true, if our claim would based on the verification results alone. In the final paragraphs, we demonstrate how these arguments can be paired with philological arguments which allow us to arrive at a stronger conclusion.

3.) In my own opinion this is definitively not a "highlight", but instead a fundamental element
of any IR/ML/NLP-paper that deals with the comparison of several methods/algorithms.
When applying algorithms (in the context of this paper --> AV-methods) one should ALWAYS consider using publicly available corpora in order to make results (or more precisely experiments) reproduceable.
> We agree, but this typically very difficult in the context in literary studies, because of the difficult copyright that rests on this kind of materials.

Highlights

1. We shed new light on the authenticity of the writings of Julius Caesar.
2. Hirtius, one of Caesar's generals, must have contributed to Caesar's writings.
3. We benchmark two authorship verification systems on publicly available data sets.
4. We test on both modern data sets, and Latin texts from Antiquity.
5. We show how computational methods inform traditional authentication studies.