Authorship Verification with the Minmax Metric

Mike Kestemont  
University of Antwerp  
mike.kestemont@uantwerp.be

Justin Stover  
University of Oxford  
justin.stover@classics.ox.ac.uk

Moshe Koppel  
Bar-Ilan University  
moishk@gmail.com

Folgert Karsdorp  
Radboud University Nijmegen  
fbkarsdorp@fastmail.nl

Walter Daelemans  
University of Antwerp  
walter.daelemans@uantwerp.be

Authorship studies have long played a central role in stylometry, the popular subfield of DH in which the writing style of a text is studied as a function of its author’s identity. While authorship studies come in many flavors, a remarkable aspect is that the field continues to be dominated by so-called ‘lazy’ approaches, where the authorship of an anonymous document is determined by extrapolating the authorship of a document’s nearest neighbor. For this, researchers use metrics to calculate the distances between vector representations of documents in a higher-dimensional space, such as the well-known Manhattan city block distance. In this paper, we apply the minmax metric to the problem of authorship verification.

We illustrate the broader applicability of authorship verification by reporting a high-profile case study from Classical Antiquity. The ‘War Commentaries’ by Julius Caesar (100-44 BC) are a group of Latin descriptions of the military campaigns of the famous Roman statesman. 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.
To evaluate our verification approach, we use the procedure used in the 2014 track on authorship verification in the PAN competition on uncovering plagiarism, authorship, and social software misuse. This track focused on the “open” task of authorship verification in 6 data sets. Each dataset holds a number of “PROBLEMS”, where given (a) at least one training text by a particular target author, (b) a set of similar mini-oeuvres by other authors, and (c) a new anonymous text, the task is to determine whether or not the anonymous text was written by the target author. A system must output for each of the verification PROBLEMS a real-valued confidence score between 0.0 and 1.0. For each dataset, a fully independent training and test corpus are available (i.e. the PROBLEMS, nor authors and texts in both sets do not overlap). Systems are eventually evaluated using two scoring metrics which were also used at the PAN: the established AUC-score, as well as the so-called C@1, a variation of the traditional ACCURACY-score, which gives more credit to systems that decide to leave some difficult verification problems unanswered.

As common in text classification, we vectorize the datasets under a ‘bag-of-words’ assumption, which is largely ignorant of the original word order in document. We use character tetragrams below (Koppel and Winter 2014) and experiment with a number of different vector space models:

- **plain tf** (where simple relative frequencies are used);
- **tf-std**, where the tf-model is scaled using a feature’s standard deviation in the corpus (cf. Burrows’s Delta: Burrows 2002);
- **tf-idf**, where the tf-model is scaled using a feature’s inverse document-frequency (to increase the weight of rare terms).

In our experiments, we include the ‘minmax’ distance metric, a still fairly novel algorithm in stylometry (Koppel and Winter 2014), which calculates a real-valued distance score between two document vectors \( A \) and \( B \):

$$
\text{minmax}(\bar{A}, \bar{B}) = 1 - \frac{\sum_{i=1}^{n} \min(\text{tf}(A_i), \text{tf}(B_i))}{\sum_{i=1}^{n} \max(\text{tf}(A_i), \text{tf}(B_i))}
$$

**Figure 1** The minmax metric

In our experiments, we make use of the General Imposters Method, a bootstrapped approach to authorship verification. We use Algorithm 1 to determine whether an anonymous text was written by the target author specified in the problem:
Figure 2 The General Imposters Algorithm

Algorithm 1: General Imposters Method

input: $x$, an anonymous document; $T = \{T_1, \ldots, T_n\}$, a set of document by the target author; $I = \{I_1, \ldots, I_n\}$, a set of documents by other authors;

output: $0 \leq score \leq 1$

Set $score = 0.0$;

for $i \leftarrow 1$ to $k$ do

    Randomly select $rate\%$ of the available features;
    Randomly select $m$ imposters from $I$ as $I'$;
    if $\min(dist(T, x)) < \min(dist(I', x))$ then
        $score = score + 1/k$;
    end

end

Return $1 - score$;

During $k$ iterations (default 100), we randomly select a sample (default 50%) of all the available features in the data set. Likewise, we randomly select $m$ ‘imposter’ documents (default 30), which were not written by the target author. Next, we use a $dist()$ function to assess whether the anonymous text is closer to any text by the target author than to any text written by the imposters. Here, $dist()$ represents a regular, distance metric, such as the Manhattan, Cosine or Ruzicka distance metric. The general intuition is that we do not just calculate how different two documents are; rather we test whether the stylistic differences between them are consistent (a) across many different feature sets, and (b) in comparison to other randomly, sampled documents. We compare the Imposters Approach to a strong baseline proposed by Potha and Stamatatos (2014) on a reference corpus of Latin prose from Antiquity. We will demonstrate that the imposter approach produces extremely strong results across most combinations of vector spaces and distance metrics (cf. the precision-recall curves below).
Finally, we report the case study concerning the Corpus Caesarianum, the group of five commentaries on Caesar’s military campaigns: *Bellum Gallicum*, *Bellum civile*, *Bellum Alexandrinum*, *Bellum Africum*, and *Bellum Hispaniense*. The first two commentaries are mainly by Caesar himself, the only exception being the final part of the *Gallic War* (Book 8), which is by Caesar’s general Aulus Hirtius. Suetonius, writing a century and a half later, suggests that either Hirtius or another general,
named Oppius, authored the remaining works. We will report experiments which broadly supports the Hirtius's own claim that he himself compiled and edited the corpus of the non-Caesarian commentaries. Figure 3, for instance, shows a heatmap-like visualisation, in which Hirtius's Book 8 of the Gallic Wars clearly clusters with the bulk of the Alexandrian War (labeled x).

![Figure 3: Heatmap-like visualisation showing Hirtius's Book 8 of the Gallic Wars clustering with the bulk of the Alexandrian War.](image)

Figure 5 Minmax-based clustermap of 1000-word samples of the Corpus Caesarianum.

References

