International Journal of Digital Humanities (2023) 5:1–3
https://doi.org/10.1007/s42803-022-00060-9

EDITORIAL

Introduction

George Mikros[1]

Published online: 6 December 2022

Stylometric authorship attribution research aims to show that a text of unknown authorship can be ascribed to a specific author using a set of quantifiable text features as indicators of the author's style. It is one of the oldest applications of quantitative methods in linguistic data, with relevant studies based on the manual counting of linguistic features dating back to the 15th century.

Since the late 1990s, authorship attribution has been receiving a new impetus brought by developments in several key research areas, such as Information Retrieval, Machine Learning, and Natural Language Processing. Furthermore, machine-readable text is now massively available. With the advent of Web 2.0, new forms of online expression, such as blogs, tweets, and instant messaging, appeared alongside the internet genres that had already become standard (email, web pages, and online forum messages). Moreover, since 2014, the NLP community has been revolutionized by the rise of word embeddings and deep neural document representation models like BERT, which offer holistic textual representation models for all text mining tasks.

Authorship analysis research is now concerned not only with problems in the broad field of the Humanities (Literature, History, Theology) but also with applications in various law-enforcement tasks such as Intelligence, Forensics, etc. At the same time, a number of different research questions have been raised, including issues of author profile (gender, age, personality, etc.) that formed a broader research agenda and fostered the development of the broader field of computational stylistics.

In this rapidly changing research environment, authorship attribution methodology closely follows the evolution of text mining methods. This special issue aims to capture the state-of-the-art in the field and broadly cover the most active research areas. We were happy to receive some very innovative papers on the broader topic of authorship identification that cover many interesting aspects of this field.

✉ George Mikros
gmikros@hbku.edu.qa

[1] Department of Middle Eastern Studies, College of Humanities and Social Sciences, Hamad Bin Khalifa University, Doha, Qatar

Our first paper in this volume is written by Erwan Moreau and Carl Vogel and titled "CLG Authorship Analytics: a library for authorship verification." This paper describes the CLG Authorship Analytics software focusing on authorship verification task, i.e., the detection of whether the same person has written two texts. Authorship verification remains one of the most challenging tasks since it is an open classification problem. In the relevant literature, it has been tackled both in a supervised and unsupervised learning context. CLG Authorship Analytics is a versatile software that can be used in both the abovementioned contexts (although the version described here is mainly based on the supervised version), and it is designed so that new functions (ranging from stylometric features to distances and verification algorithms) can be added easily. The general design of the system is based on ensemble learning methods combining the strengths of multiple heterogeneous individual approaches to authorship verification. These methods are trained on various stylometric feature categories (observation families, as they are called). The methods employed in the system (called strategies) are three well-known algorithms for authorship verification: (a) Basic strategy based on simple distance measure (cosine and min-max), (b) General Impostor, and (c) Universum Inference. Each strategy returns a set of features for every verification problem and gives an answer indicating whether the two groups of documents are from the same author or not. This set of labeled verification problems is used to train a supervised model (a meta-model) that can compute confidence levels in the final verification answer using regression modeling. Moreover, since each strategy comes with its own hyper-parameters space, the software offers a hyper-parameter tuning based on a genetic algorithm to optimize the strategies' performance automatically. A series of experiments were conducted to determine the effect of various standard factors on the performance of the authorship verification task. The results indicate that the size of the documents and the number of occurrences in the training set contribute to the overfitting reduction while increasing the known author bias. In contrast, increasing the number of papers per group has little or no effect on overfitting but tends to reduce the author bias that is already established. The General Impostor approach performed the best overall. In addition, the results indicated that the meta-model generally results in less overfitting and more steady performance overall. However, the reported experiments did not conclusively prove the superiority of the meta-model. The best individual technique, GI, performs as well or slightly better in most cases. An interesting conclusion from this research is that despite the variety of hyper-parameters for each strategy and the various learners merged in the meta-model, the final ensemble used might benefit from a greater selection of model types.

Tunç Yılmaz and Tatjana Scheffler submitted a paper titled "Song Authorship Attribution: A Lyrics and Rhyme Based Approach". Their research focuses on applying authorship attribution to a large corpus of music lyrics. This contribution is very interesting since song lyrics, as a subgenre of poetry, incorporate cultural components and stylistic characteristics that are absent from prose. The authors used a variety of stylometric features not usually employed in prose-based authorship research, such as recurring sound patterns and rhyme-based structures in lyrics. They constructed a new, large-scale, balanced data set consisting of 12,000 song lyrics from 120 different artists. Moreover, they proposed CNN models for authorship attribution on this song

lyric data set to utilize structural information included within the lyrics, a technique that is also used in image classification. Several experiments were conducted exploiting a multi-modal approach that integrated character, phoneme, and sub-word level embeddings. Furthermore, the authors implemented variations of CNN architectures that have been efficient in previous text classification tasks in the past, achieving overall accuracy scores of around 48% in genre and 30% in author classification tasks. Finally, using occlusion maps, the authors were able to examine the impact of lyrics-specific phoneme features. This study extends the authorship attribution methodology to the genre of the lyrics and contributes to the relevant research community a large pre-processed, public, and ready-to-use corpus for similar research purposes.

Our third paper, titled "Computational Authorship Analysis of the Homeric Poems," is written by John Pavlopoulos and Maria Konstantinidou. This paper contributes to the authorship attribution research by tackling the authorship of the Homeric Poems. This long-standing philological problem is considered highly complex due to the lack of direct historiographic sources, the intervention of many authors, and subsequent developments of the text at both oral and written levels. Authorship analysis of the Iliad and Odyssey was performed using character-level statistical language models. To verify that these models identify the Homeric poems as distinct works of literature, the Iliad and Odyssey were examined and evaluated both internally (examining the stylometric similarity between the rhapsodies inside each of the Homeric Poems) and against other similar texts. The results showed that the language models could adequately differentiate between each of the Homeric poems and two of the works most similar to them: Hesiod's *Theogony* and *Works and Days*. In a different experiment, it was established that the models could be used to classify excerpts as belonging to the Iliad or the Odyssey. Students and scholars of Greek literature were asked to complete a questionnaire. Their responses were compared against the developed language model-based classifier, showing that the latter performs better than most annotators. These language models were also employed for analyzing subparts of the Homeric poems and using authorship verification methods to detect text passages in both poems that exhibit low linguistic proximity to the other textual fragments of the poems, implying that they might have been written by a different author.

The final version of this special issue might include papers whose revisions were submitted after the deadline and thus had not been accepted by the time this introduction was written.

The publication of this issue would not have been possible without the invaluable insights and feedback from the reviewers of the submitted papers. I would also like to immensely thank Dr. Péter Tamás for continuously supporting me and coordinating the authors and the Publishing House. A final note of gratitude to the editor-in-chief of the International Journal of Digital Humanities, Dr. Palkó Gábor (Centre for Digital Humanities, Eötvös Loránd University, Hungary), and the rest of the core editorial team, Dr. Thorsten Ries (Department of Germanic Studies, University of Texas, USA) and Dr. Kees Teszelszky (Koninklijke Bibliotheek - National Library of the Netherlands, the Netherlands).