Unsupervised Authorship Attribution for Medieval Latin using Transformer-Based Embeddings

Loic De Langhe, Orphée De Clercq, Veronique Hoste
LT3, Language and Translation Technology Team, Ghent University, Belgium
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
firstname.lastname@ugent.be

Abstract

We explore the potential of employing transformer-based embeddings in an unsupervised authorship attribution task for medieval Latin. The development of Large Language Models (LLMs) and recent advances in transfer learning alleviate many of the traditional issues associated with authorship attribution in lower-resourced (ancient) languages. Despite this, these methods remain heavily understudied within this domain. Concretely, we generate strong contextual embeddings using a variety of mono -and multilingual transformer models and use these as input for two unsupervised clustering methods: a standard agglomerative clustering algorithm and a self-organizing map. We show that these transformer-based embeddings can be used to generate high-quality and interpretable clusterings, resulting in an attractive alternative to the traditional feature-based methods.

Keywords: Authorship Attribution, Medieval Latin, Unsupervised Learning

1. Introduction

Throughout modern history, scholars have always been greatly interested in the authenticity and authorship of important historical documents. In the fifteenth century, Renaissance scholar Lorenzo Valla exposed the purported 4th century imperial decree *Donatii Constantini* as an 8th century forgery by comparing the language in the document with actual 4th century Latin sources. A little over 500 years later, Mosteller and Wallace (1963) showed through statistical analysis that James Madison, rather than Alexander Hamilton, was the author of 12 disputed documents in the (in)famous *Federalist papers*. In short, the methods may have changed, but the question has remained the same.

In a computational setting, authorship analysis is often analogous to stylometry i.e. the use of quantifiable and statistical methods to unmask an author's stylistic DNA or signature (Holmes, 1998). At the forefront of this field lies the idea that individual authors have a marked and highly specific writing style that can be used to separate them from others (Stamatatos, 2009). Modern stylometric studies typically focus on the attribution of essays, emails and forum posts to distinct online users or groups of users (Kestemont et al., 2018). Naturally, the field ties in to modern-day applications such as plagiarism detection, identity deception on social media platforms and multi-modal authentication on mobile devices (Neal et al., 2017). While there is an emphasis on applying stylometric methods in modern settings, the stylistic analysis of work from the antiquity and medieval periods also remains a highly studied topic. The emergence and distribution of large electronic document collections containing heaps of anonymous or (seemingly) miss-attributed texts has lead to many researchers continuing directly in Lorenzo Valla's footsteps, more than 500 years after his passing.

Research on antique and medieval texts is hampered by a general lack of spelling and language standardization as well as transcription errors (Kestemont, 2012). This naturally poses an additional layer of difficulty, as it is hard to determine whether or not the spelling of a word is due to the original author's stylistic signature, or was introduced by those transcribing the work. Nonetheless, computational stylometric analysis of antique and medieval texts has led to the identification of previously anonymous authors, or the rectification of the authorship of misattributed work (Stover et al., 2016; Kabala, 2020). It is to be noted that, unlike in most NLP domains, the use of neural approaches remains limited, mostly due to the lack of large amounts of training data, which these deep neural architectures typically require to function optimally (Corbara et al., 2023). Nonetheless, recent advancements in the field of Natural Language Processing (NLP) have given rise to large-scale transformer architectures which circumvent the need for large task-specific corpora through transfer learning. Despite their ability to capture accurate representations of longer documents and encode implicit textual structures, transfer learning methods remain understudied in the context of medieval stylometry.

In this paper, we explore the potential of using a variety of transformer-based models for unsupervised authorship attribution in Medieval Latin. Concretely, we generate powerful vectorial representations of Medieval Latin texts and use these as a basis for two unsupervised clustering methods:

LT4HALA 2024@LREC-COLING-2024, pages 57–64

a standard agglomerative clustering algorithm and a self-organizing map (SOM). The former serves as our primary method for intrinsic and extrinsic evaluation of the generated clusters, while the latter aims to create highly interpretable visualisations of the data. We show that, without relying on a series of highly specialized manually crafted features, we can accurately cluster a large number of 13th-14th century Latin texts by author, illustrating the potential of using transfer-learning methods in future stylometric studies.

2. Related Work

Work on computational methods for authorship attribution goes back to the very beginning of the field of Computational Linguistics (CL) as a whole (Holmes, 1998). Earlier work often focused on well-known contested English texts, with the disputed *Federalist Papers* being a notable example that has been studied multiple times throughout the years (Mosteller and Wallace, 1963; Tweedie et al., 1996). More recently however, there has been a growing interest in performing computational stylistic analysis on a wider range of languages such as Dutch (Kestemont, 2012; Morante et al., 2022), Ancient Greek (Gorman and Gorman, 2016), Spanish (López-Escobedo et al., 2013) and many others (Savoy, 2020).

For Latin specifically, there have been, among others, stylometric studies regarding the works of Hildegard of Bingen (Kestemont et al., 2015), Dante Alighieri (Corbara et al., 2019) and the attribution of a newly discovered manuscript to the writer Apuleius (Stover et al., 2016). Additionally, specific authorship attribution tools such as *Medievalla* have been developed and made available to the wider research community (Corbara et al., 2022). Note that most of these studies largely follow the same approach: the combination of rigorously handcrafted stylistic features combined with traditional machine learning algorithms (Muldoon et al., 2021). While there have been recent studies that combine well-known stylistic markers such syllabic patterns with deep neural networks (Corbara et al., 2023), more modern neural methods such as transformer-based architectures remain a largely unexplored approach.

All of the methods earlier described made use of the standard supervised learning paradigm in which the ground truth (or gold-standard labeling) is known and used to evaluate the performance of a given algorithm. Nonetheless, unsupervised approaches are often being applied to (historical) NLP tasks to automatically find underlying patterns without the need for human intervention (Kehler and Stolcke, 1999; Bharadiya, 2023). For authorship attribution specifically clustering algorithms are often applied to uncover implicit similarity between the works of known writers and anonymous documents or to determine outliers (i.e. possibly misattributed works) in their bibliography (Martín-del Campo-Rodríguez et al., 2022). Research on unsupervised methods for stylometry often makes use of the popular agglomerative clustering algorithm (Layton et al., 2013; Panicheva et al., 2019), but other methods such as c-means (Demir, 2013) and self-organizing maps (Ranatunga et al., 2011; Neme et al., 2015) have also been applied. Note also that most studies involving unsupervised learning forgo the use of hand-crafted feature sets and instead focus on more easily extractable textual information such as character n-grams (Kapočiūtė-Dzikienė et al., 2015), punctuation (Tanguy et al., 2012) or rudimental similarity functions between texts (Qian et al., 2015).

3. Experiments

3.1. Data

Our data consists of the Medlatin1 and Medlatin2 corpora, which are composed of 13-14th century Latin epistles (MedLatin1) and literary analyses (Medlatin2) by a variety of authors (Corbara et al., 2022). As was done in Corbara et al. (2023), we merge the two corpora resulting in one dataset encompassing 324 medieval Latin texts. We then remove a total of 31 epistels for which no specific author is known, resulting in a final collection of 293 documents.

3.2. Experimental Setup

3.2.1. Agglomerative Clustering

First, we apply an agglomerative clustering algorithm which uses an average linkage criterion i.e. two clusters are merged based on the average of distances between all pairs of both objects. For two clusters A and B the distance between them is defined as:

$$d_{AB} = \frac{1}{kl} \sum_{i=1}^{k} \sum_{j=1}^{l} d(X_i, Y_j)$$

Where X_i and Y_j are objects within clusters A and B respectively and $d(.)$ is the distance (cosine) function. The results of this algorithm will serve as our prime (numerical) evaluation of cluster quality. Note that unsupervised methods are typically evaluated both intrinsically (unsupervised, cluster quality and how well the clusters are separated) and extrinsically (supervised, based on the gold-standard labels). For our analysis we will take both evaluation strategies into account.

3.2.2. Self-Organizing Map

In addition to the standard clustering algorithm, we train a self-organizing map (SOM) neural network, which will allow a more interpretable analysis of the obtained clusters. The self-organizing map (Oja and Kaski, 1999) is a 2-dimensional representation of a series of data points which respects the topological structure of the dataset. We follow the standard SOM algorithm as it was presented in Oja and Kaski (1999). First, a document x is sampled randomly from the collection and based on the randomly initialized weights w of the neurons in the lattice the best matching unit (BMU) is determined:

$$i(x) = argmin_j \|w - w_j\|$$

The weights in the lattice are then updated through a Hebbian learning rule where η is the learning rate and $h(j, i(x))$ is the (Gaussian) neighborhood function which allows incremental updates to neurons surrounding the BMU:

$$w_j \leftarrow w_j + \eta h(j, i(x)))(x - w_j)$$

3.2.3. Textual representation

For both methods we present each individual document in the dataset as a transformer-generated representation of said document. Each text is passed through a transformer encoder to create a high-dimensional vector representation (embedding). Following earlier studies on the effectiveness of using transformer-based embeddings (Devlin et al., 2018), we generate document embeddings based on several encoder layers, rather then only using the last layer as an instance's representation. We concatenate the transformers' last four encoder layers (each a vector of length 768) to a 3072-dimensional feature representation for each document. We compare four distinct models in order to broadly gauge their capabilities w.r.t medieval Latin. First, a monolingual Latin RoBERTa model[1] which was trained on the Latin part of the cc-100 corpus (Conneau et al., 2019). Second, a multilingual encoder model which was trained on a total of 104 languages (including Latin) (Devlin et al., 2018). Third, a multilingual model using the DeBERTaV3 architecture (He et al., 2021), which has been shown to outperform most monolingual models in a large variety of languages. The final model tested in our experiments is a longformer model. Most BERT-based encoders suffer from processing longer texts as the token limit of an input is restricted to 512. Longformer-inspired models however use a linearly scaling attention mechanism which poses significantly less strain on computational resources and allows processing of sequences of up to 4096 tokens (Beltagy et al., 2020). Given the fact that many texts of the Medlatin1 and Medlatin2 corpora are quite lengthy, long-document transformers may be more suited. The multilingual longformer model used in the experiments was trained on 103 languages (including Latin) of the cc-100 corpus [2].

3.3. Hardware and Software Implementation

All experiments were trained and evaluated on a single Tesla V100-SXM2-16GB GPU. For the implementation of the agglomerative clustering algorithm we relied on the use of Python's Scikit-Learn module (Pedregosa et al., 2011). The training and visualisation of the SOM algorithm was performed through the MiniSom package [3]. Specific training parameters can be found in Appendix A.

4. Results

4.1. Agglomerative Clustering

Most unsupervised clustering algorithms are evaluated through intrinsic methods, which evaluate the quality of a clustering by how well the clusters are separated. For this paper, we evaluate the generated clusterings through two intrinsic measures, which both measure how similar an object is to its own cluster compared to other clusters: the silhouette coefficient (SC), which ranges from -1 to +1 with higher values indicating better clusterings and the Calinski-Harabasz Index (CHI), the value of which is unrestricted and for which higher values indicate higher quality clusters. In addition to these intrinsic measures we also include the Rand Index (RI) as an evaluation metric, which computes the degree of similarity between two data partitions (the predictions and the ground truth). This metric ranges from 0 to 1, with higher values indicating larger similarity between the generated clusters and the gold standard. Table 4.1 contains the results for each of the clusters generated through the embeddings of the different encoder models.

Model	SC	CHI	RI
Longformer	0.3975	**197.00**	**0.7382**
mBERT	**0.4796**	34.77	0.6672
RoBERTa Latin	0.2526	25.01	0.6708
mDeBERTaV3	0.3036	26.41	0.6155

Table 1: Silhouette Coefficient (SC), Calinski-Harabasz Index (CHI) and Rand Index (RI) scores for the generated clusterings.

[1] https://huggingface.co/pstroe/roberta-base-latin-cased

[2] https://huggingface.co/markussagen/xlm-roberta-longformer-base-4096

[3] https://github.com/JustGlowing/minisom

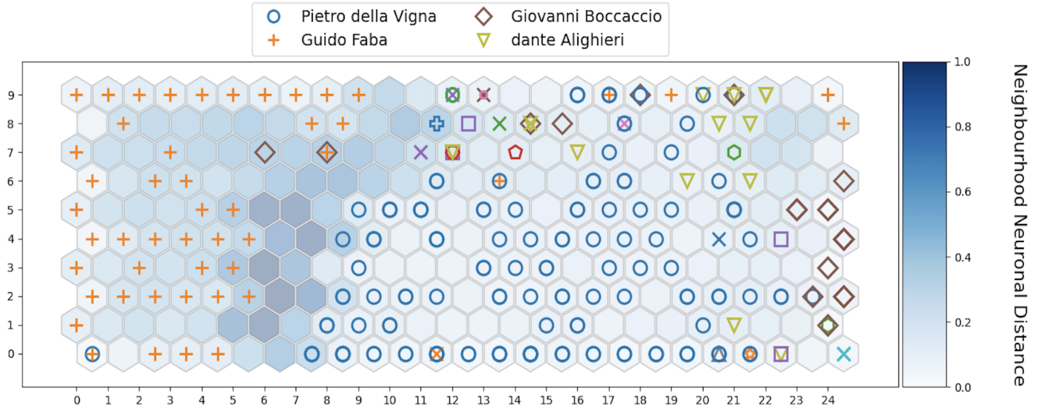

Figure 1: Visualisation of the trained self-organizing map using the Longformer embeddings as document representations.

Overall, we find that clusters generated through the embeddings of the longformer-xlmr model performed best on average, both by means of intrinsic and extrinsic evaluation. We hypothesize here that the signficantly larger context length of 4096 tokens (as opposed of 512 for the other models) ultimately plays a significant role in capturing an author's stylistic signature. We also note that the obtained RI scores for each of the models can be interpreted as moderate-to-high overlap between the generated clusters and the ground truth, indicating that unsupervised clustering through transfer learning may be a viable method for large scale analysis in the future. Interestingly, while the monolingual Latin model shows comparatively good results for the extrinsic evaluation, the intrinsic evaluation is significantly worse than the other models. This can indicate that the generated clusters, while distinguishable to a degree, are highly similar to one another. In the context of this task, this means that the authors' stylistic signatures are captured comparatively less by the monolingual Latin model.

4.2. Self-Organizing Map

We obtain a detailed topological map of the data by initializing the SOM with a 10-by-25 lattice and training the algorithm using the learning rules described in Section 3.2.2. The resulting topological map using the best model embeddings (longformer) can be seen in Figure 1. For readability's sake the legend in Figure 1 only includes the 4 most represented authors of the dataset which are (in order): Pietro Della Vigna (n = 146), Guido Faba (n = 78), Giovanni Boccaccio (n = 27) and Dante Alighieri (n = 14). A detailed legend of all 22 authors as well as the topological representations generated with the other three encoder models can be found in Appendix B.

We do not rely on quantitative metrics for the evaluation of the generated lattice, but rather on visual analysis. We observe that the SOM presents a qualitative clustering of the various authors, with the four most prominent authors clearly occupying four distinct spaces on the map. Note that the works of Guido Faba are seen as highly distinct from the other works in the dataset. Interestingly, one particular letter by Pietro Della Vigna is significantly closer to the letters of Guido Faba than to della Vigna's other works. In the end, only close reading and study can ultimately provide clarity regarding the authorship of unattributed or dubious manuscripts. Nonetheless, the identification of outliers, such as the one mentioned, through unsupervised computational analysis can serve as an early diagnostic step in this process as well as narrowing the scope of this complex task.

Finally, we also observe that for some authors with only one work in the dataset, the neuronal distance to neighboring positions is remarkably high. This indicates that the SOM neural network can segment individual authors' stylistic signatures even if there is only a limited amount of their work available. In this way, the SOM algorithm can be an effective way to detect outliers within larger document collections. This is a notable advantage of applying a SOM compared to more traditional clustering methods, which often continuously merge clusters until an arbitrary threshold is reached and thus concentrate less on the uniqueness of individual data points.

5. Conclusion

We show for the first time that transformer-generated contextual embeddings can be used to render qualitative unsupervised clusterings of author attributions in medieval Latin. We examined the embeddings of four distinct transformer models and found, through both intrinsic and extrinsic evaluation, that long-document transformer models lead to the best available clusterings. While close-reading and traditional feature-based methods are still needed to conclusively determine the authenticity or attribution of (ancient) manuscripts, we believe that transfer learning methods can be used as an early diagnostic tool for both outlier detection and narrowing the search space within large medieval document collections.

6. Acknowledgements

This work was supported by the Research Foundation–Flanders under project grant number FWO.OPR.2020.0014.01.

7. Bibliographical References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Jasmin Bharadiya. 2023. A comprehensive survey of deep learning techniques natural language processing. *European Journal of Technology*, 7(1):58–66.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Silvia Corbara, Alejandro Moreo, and Fabrizio Sebastiani. 2023. Syllabic quantity patterns as rhythmic features for latin authorship attribution. *Journal of the Association for Information Science and Technology*, 74(1):128–141.

Silvia Corbara, Alejandro Moreo, Fabrizio Sebastiani, and Mirko Tavoni. 2019. The epistle to cangrande through the lens of computational authorship verification. In *New Trends in Image Analysis and Processing–ICIAP 2019: ICIAP International Workshops, BioFor, PatReCH, e-BADLE, DeepRetail, and Industrial Session, Trento, Italy, September 9–10, 2019, Revised Selected Papers 20*, pages 148–158. Springer.

Silvia Corbara, Alejandro Moreo, Fabrizio Sebastiani, and Mirko Tavoni. 2022. Medlatinepi and medlatinlit: Two datasets for the computational authorship analysis of medieval latin texts. *Journal on Computing and Cultural Heritage (JOCCH)*, 15(3):1–15.

Nesibe Merve Demir. 2013. Artificial neural network techniques in authorship attribution. *Southeast Europe Journal of Soft Computing*, 2(2).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Vanessa B Gorman and Robert J Gorman. 2016. Approaching questions of text reuse in ancient greek using computational syntactic stylometry. *Open Linguistics*, 2(1).

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

David I Holmes. 1998. The evolution of stylometry in humanities scholarship. *Literary and linguistic computing*, 13(3):111–117.

Jakub Kabala. 2020. Computational authorship attribution in medieval latin corpora: the case of the monk of lido (ca. 1101–08) and gallus anonymous (ca. 1113–17). *Language Resources and Evaluation*, 54(1):25–56.

Jurgita Kapočiūtė-Dzikienė, Andrius Utka, and Ligita Šarkutė. 2015. Authorship attribution of internet comments with thousand candidate authors. In *Information and Software Technologies: 21st International Conference, ICIST 2015, Druskininkai, Lithuania, October 15-16, 2015, Proceedings 21*, pages 433–448. Springer.

Andrew Kehler and Andreas Stolcke. 1999. Unsupervised learning in natural language processing. In *Association for Computational Linguistics. Proceedings of the workshop. In Preface A. Kehler and A. Stolcke, editors*.

Mike Kestemont. 2012. Stylometry for medieval authorship studies: an application to rhyme words. *Digital Philology: A Journal of Medieval Cultures*, 1(1):42–72.

Mike Kestemont, Sara Moens, and Jeroen Deploige. 2015. Collaborative authorship in the twelfth century: A stylometric study of hildegard of bingen and guibert of gembloux. *Digital Scholarship in the Humanities*, 30(2):199–224.

Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. 2018. Overview of the author identification task at pan-2018: cross-domain authorship attribution and style change detection. In *Working Notes Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018/Cappellato, Linda [edit.]; et al.*, pages 1–25.

Robert Layton, Paul Watters, and Richard Dazeley. 2013. Automated unsupervised authorship analysis using evidence accumulation clustering. *Natural Language Engineering*, 19(1):95–120.

Fernanda López-Escobedo, Carlos-Francisco Méndez-Cruz, Gerardo Sierra, and Julián Solórzano-Soto. 2013. Analysis of stylometric variables in long and short texts. *Procedia-Social and Behavioral Sciences*, 95:604–611.

Carolina Martín-del Campo-Rodríguez, Grigori Sidorov, and Ildar Batyrshin. 2022. Unsupervised authorship attribution using feature selection and weighted cosine similarity. *Journal of Intelligent & Fuzzy Systems*, 42(5):4357–4367.

Roser Morante, Eleanor LT Smith, Lianne Wilhelmus, Alie Lassche, and Erika Kuijpers. 2022. Identifying copied fragments in a 18th century dutch chronicle. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5865–5878.

Frederick Mosteller and David L Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309.

Connagh Muldoon, Ahsan Ikram, and Qublai Ali Khan Mirza. 2021. Modern stylometry: A review & experimentation with machine learning. In *2021 8th International Conference on Future Internet of Things and Cloud (FiCloud)*, pages 293–298. IEEE.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSuR)*, 50(6):1–36.

Antonio Neme, JRG Pulido, Abril Muñoz, Sergio Hernández, and Teresa Dey. 2015. Stylistics analysis and authorship attribution algorithms based on self-organizing maps. *Neurocomputing*, 147:147–159.

Erkki Oja and Samuel Kaski. 1999. *Kohonen maps*. Elsevier.

Polina Panicheva, Olga Litvinova, and Tatiana Litvinova. 2019. Author clustering with and without topical features. In *International Conference on Speech and Computer*, pages 348–358. Springer.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Tie-Yun Qian, Bing Liu, Qing Li, and Jianfeng Si. 2015. Review authorship attribution in a similarity space. *Journal of Computer Science and Technology*, 30(1):200–213.

RVSPK Ranatunga, AS Atukorale, and KP Hewagamage. 2011. Intrinsic plagiarism detection with kohonen self organizing maps. In *U The International Conference on Advances in ICT for Emerging Regions-ICTer2011*, volume 125.

Jacques Savoy. 2020. Machine learning methods for stylometry. *Cham: Springer*.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Justin Anthony Stover, Yaron Winter, Moshe Koppel, and Mike Kestemont. 2016. Computational authorship verification method attributes a new work to a major 2nd century a frican author. *Journal of the Association for Information Science and Technology*, 67(1):239–242.

Ludovic Tanguy, Franck Sajous, Basilio Calderone, and Nabil Hathout. 2012. Authorship attribution: Using rich linguistic features when training data is scarce. In *PAN Lab at CLEF*.

Fiona J Tweedie, Sameer Singh, and David I Holmes. 1996. Neural network applications in stylometry: The federalist papers. *Computers and the Humanities*, 30:1–10.

A. Appendix A

Parameter	Value
Lattice Dimension	10x25
Learning Rate	0.7
Neighborhood Function	Gaussian
Distance Metric	Cosine Distance
Topology Configuration	Hexagonal
Neighborhood Radius	6
Training Iterations	1000

Table 2: Training configuration for the SOM algorithms. All SOM representations were trained using identical parameters.

B. Appendix B

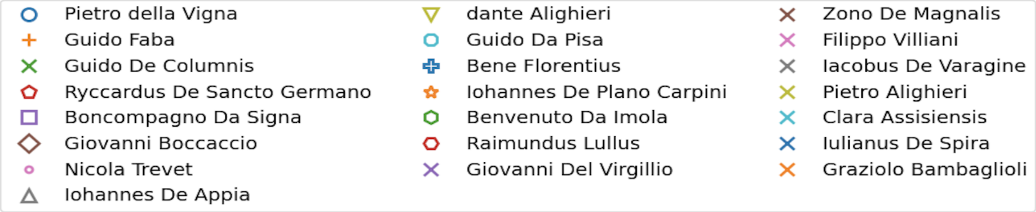

Figure 2: Complete legend of all 22 authors for the SOM visualisations.

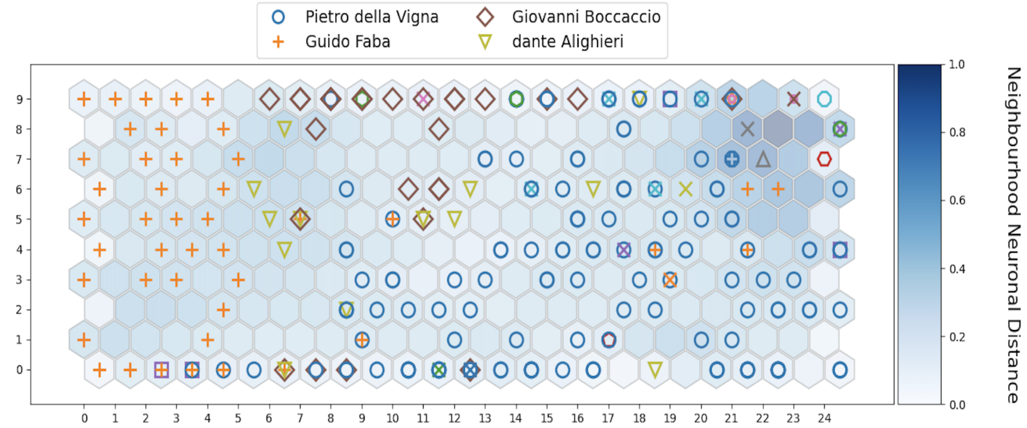

Figure 3: Visualisation of the trained self-organizing map using the mDeBERTaV3 embeddings as document representations.

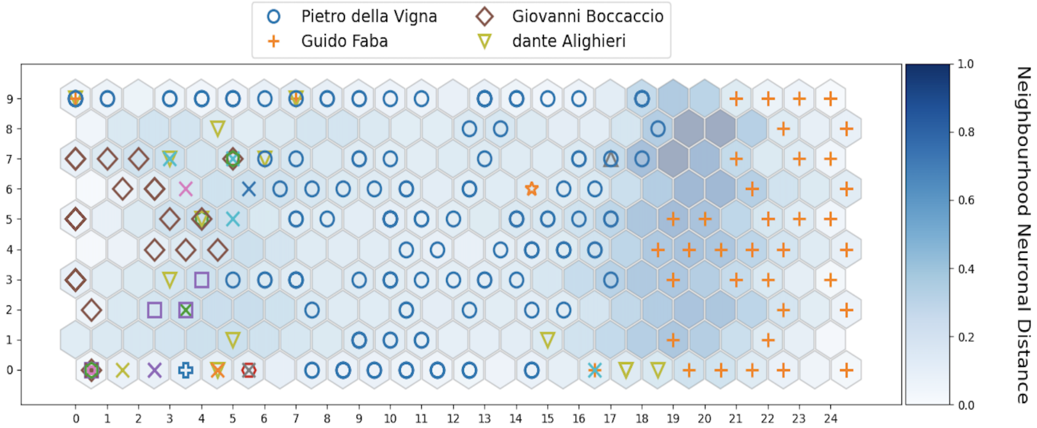

Figure 4: Visualisation of the trained self-organizing map using the Latin RoBERTa embeddings as document representations.

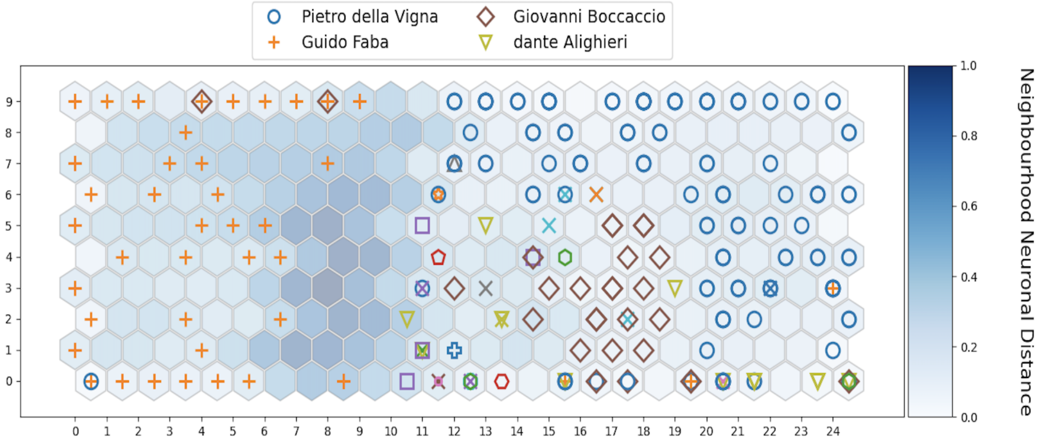

Figure 5: Visualisation of the trained self-organizing map using the mBERT embeddings as document representations.