Can Correspondence Analysis Challenge Transformers in Authorship Attribution Tasks?

Andrea Sciandra and Arjuna Tuzzi

Abstract: With reference to a large corpus of 76 Italian contemporary popular mystery novels by 16 different authors, this study aims to assess the performance of large language models in an authorship attribution test. The results obtained through both transformers and correspondence analysis vector representations are compared and contrast in machine learning classification tasks. Although in previous works transformers have been shown to perform better than other alternatives, in this case, correspondence analysis wins the challenge. Results support the hypothesis that specialized large corpora require tailor-made representations.

Riassunto: Attraverso l'analisi di un corpus di 76 gialli contemporanei italiani di 16 diversi autori, questo studio intende valutare le *performance* dei *large language models* in una prova di attribuzione d'autore. Vengono confrontati in diverse classificazioni basate su *machine learning* i risultati ottenuti adottando rappresentazioni vettoriali con i *transformers* e con l'analisi delle corrispondenze. Sebbene in lavori precedenti i *transformers* si siano dimostrati migliori di altre alternative, in questo caso è l'analisi delle corrispondenze a vincere la sfida. I risultati confermano l'ipotesi che grandi corpora specializzati richiedano soluzioni su misura.

Key words: authorship attribution, machine learning, correspondence analysis, transformers, popular Italian mystery novels

[1] Andrea Sciandra, Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata (FISPPA); email: andrea.sciandra@unipd.it

Arjuna Tuzzi, Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata (FISPPA); email: arjuna.tuzzi@unipd.it

Introduction

The emergence of large language models (LLM) and the availability of transformers (Wolf et al., 2020) for vector representation of words and texts have significantly changed the landscape of text mining. They have proven particularly innovative in numerous applications related to text classification tasks. However, the question to determine whether LLMs are the ultimate solution to all problems persists, particularly because the black-box nature of this technology significantly constrains the ability to achieve explainable models. This study aims to assess the performance of LLMs in an authorship attribution task, comparing and contrasting the results with those obtained through correspondence analysis (CA, Greenacre, 1984; Murtagh, 2005) as suggested by Lebart (1997).

The corpus includes 76 Italian contemporary popular mystery novels by 16 different authors. Each author contributes a minimum of three to a maximum of six novels. It is a very homogeneous corpus by text-genre, created with the support of the publishing house, which lively contributed to the choice of the most significant authors and works. On the basis of simple tokenization that takes into account only alphabetic sequences (all other signs were considered separators, and numbers were excluded) and simple normalization that only transformed the words into lowercase letters, the corpus has a total size of just under five million words (4,845,966 word tokens) and expresses a vocabulary of just over 100 thousand different words (109,643 word types). The length of the novels ranges from a minimum of 22,402 to a maximum of 123,457 words, thus the length of the works is quite variable, as is the contribution to the corpus of each author (Fig. 1). Given that the type-token ratio is equal to 2.2% (average frequency of word types 44.2) and the percentage of *hapax legomena* is 38.1% of the vocabulary, the corpus shows a desirable high degree of redundancy.

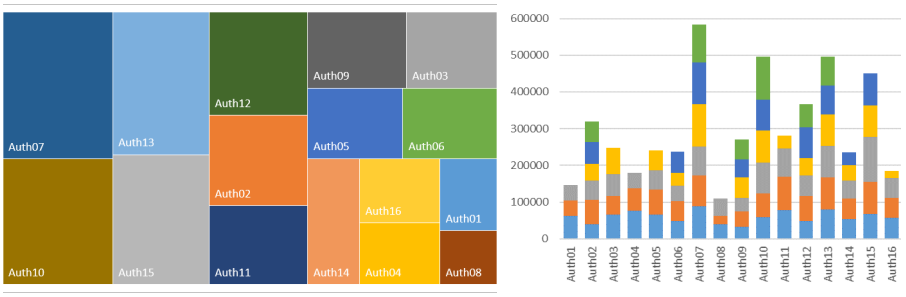

Figure 1: Dimensions in word tokens by author (left) and by author and novel (right).

As it is homogeneous in terms of text genre and includes novels of undisputed authorship, our corpus represents a good basis for testing the ability of a classification pipeline for an authorship attribution task. This study aims at assessing the performance of the same classification methods based on machine learning but with different ways of representing the features: we compete between vectorization based

on modern transformers and that provided by traditional correspondence analysis. All texts are partitioned into 9526 equal-size text chunks (500 words, since we used a LLM model that can process texts of the maximal length of 512 tokens), and performance is measured based on the ability to attribute each chunk of the test set to the right author.

Methods

We selected word types with number of occurrences >=76 (one occurrence per novel on average) and started from the (5061 x 9526) Term-Document Matrix (TDM) with (relative) frequencies as weights. After calculating the chi-square distance and Singular Value Decomposition (SVD), each text (word) becomes a column (row) vector in the 5060-dimensional space of CA orthogonal axes. We can take the first dimensions of these vectors as representations of texts (words) to submit to ML classification models. We chose to select the first 76 axes, equal to the number of texts available. The explained variance for the first 76 axes is about 13%.
Next, we obtained a representation of texts using transformers, specifically by exploiting the BERT Base Italian Cased LLM pre-trained vectors (https://huggingface.co/dbmdz/bert-base-italian-cased). BERT (Bidirectional Encoder Representations from Transformers; Devlin et al. 2019) provides contextualised embeddings that consider the context in which a word is used. Therefore, they dynamically produce word representations informed by the surrounding words. In this study, we concatenated the 12 BERT layers that represent the same token and used the mean to aggregate the embeddings from different tokens to represent a text. As the BERT-base mapping involves 768 vectors, to reduce dimensionality we applied a principal component analysis, extracting the first 76 components. These components account for approximately 82% of the cumulative proportion of variance of embeddings obtained via LLM.
Regarding ML-based classification methods, we chose two models: Random Forests (RF; Breiman, 2001), and Extreme Gradient Boosting (XGBoost; Chen et al., 2019) to provide two informative validations through bagging and boosting techniques. The training set was generated by randomly selecting 80% of the chunks of each author, while the remaining chunks were used for the test set. Each model involved a 5-fold cross-validation, repeated 30 times. The evaluation of the results included the analysis of accuracy, unweighted Kappa statistic, and a one-sided test to verify if the accuracy is greater than the "no information rate" (the largest class percentage in the data: 12.07%; Kuhn, 2008).

Discussion and conclusions

The results of the models for the test set are presented in Table 1.

The models using CA consistently outperformed those using LLM. The best model overall is XGBoost with features derived from CA, achieving an accuracy of approximately 99%. The Kappa statistic confirms that the XGBoost model with CA coordinates is the most accurate. All models are significant according to the one-sided test, as the accuracy is higher than the no information rate (NIR). These results indicate that CA features demonstrate the highest classification capacity in this particular corpus.

The heatmaps (confusion matrices) in Figure 2 compare the classification results of the XGBoost-CA and XGBoost-LLM models in terms of the probability of attributing chunks of the test set to the authors of the corpus. The red-coloured chunks represent correct author attributions with probabilities close to 100%, highlighting the superior performance of the CA model.

The ability of the CA to discriminate the authors of the novels is evident by projecting the position of the chunks belonging to the test set onto a Cartesian plane (Dim. 1 – Dim. 2, and Dim. 4 – Dim. 5 in this example). It can be observed that the first dimension isolates author number 2, while the fifth dimension isolates authors number 9 and 14 (Fig. 3).

Although in previous works (Jones et al., 2022; Ai et al., 2022; Rodella et al., forthcoming), transformers have been shown to perform better than other alternatives, in this case, CA wins the challenge. Our hypothesis is that the performance of transformers depends on the specific application context. In particular, we are convinced that transformers, born from the training of LLMs on general, multilingual, multifaced, multigenre databases, offer a very good representation for a text classification in general, but perform worse when the language is not English, when texts are large and complex, when the corpus represents a special language and a specific genre. In the latter cases, methods that arise from the analysis of the textual data of the corpus itself could be more powerful, as they exploit endogenous information from the training set. Our hypothesis requires better testing on different corpora and with different ML methods.

It is important to note that some LLMs do not explicitly state the material on which they were trained, which can pose a challenge for this type of research. Given the vast array of open-source tools, resources, and code available online, we can observe that there is not a commensurate availability of discussion, documentation, and scientific publications. It is very challenging to choose among so many opportunities without any criteria to evaluate which ones are the best, for what purposes, and for what contexts of use.

Table 1: Comparison of author classification performance: LLM and CA (Diagnostics: Accuracy, unweighted Kappa statistic, and a one-sided test *p*-value to verify if the accuracy is greater than NIR).

ML Model	***Accuracy***	***Kappa***	***P-Value (Acc > NIR)***
LLM - RF	0.7608	0.7395	< 2.2e-16
LLM - XGBoost	0.8667	0.8557	< 2.2e-16
CA - RF	0.9831	0.9812	< 2.2e-16
CA - XGBoost	0.9863	0.9852	< 2.2e-16

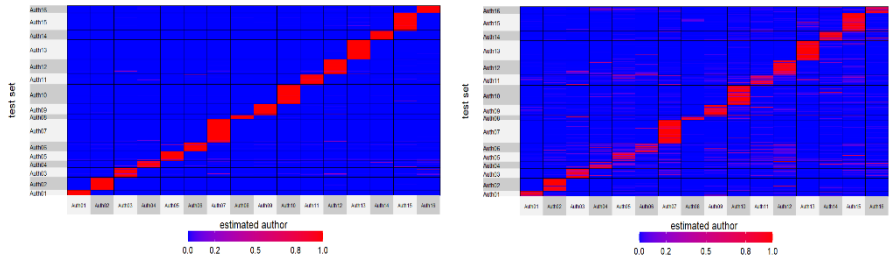

Figure 2: Heatmaps (confusion matrices) for authors' attribution (probabilities): CA-XGBoost model (left) and LLM-XGBoost (right).

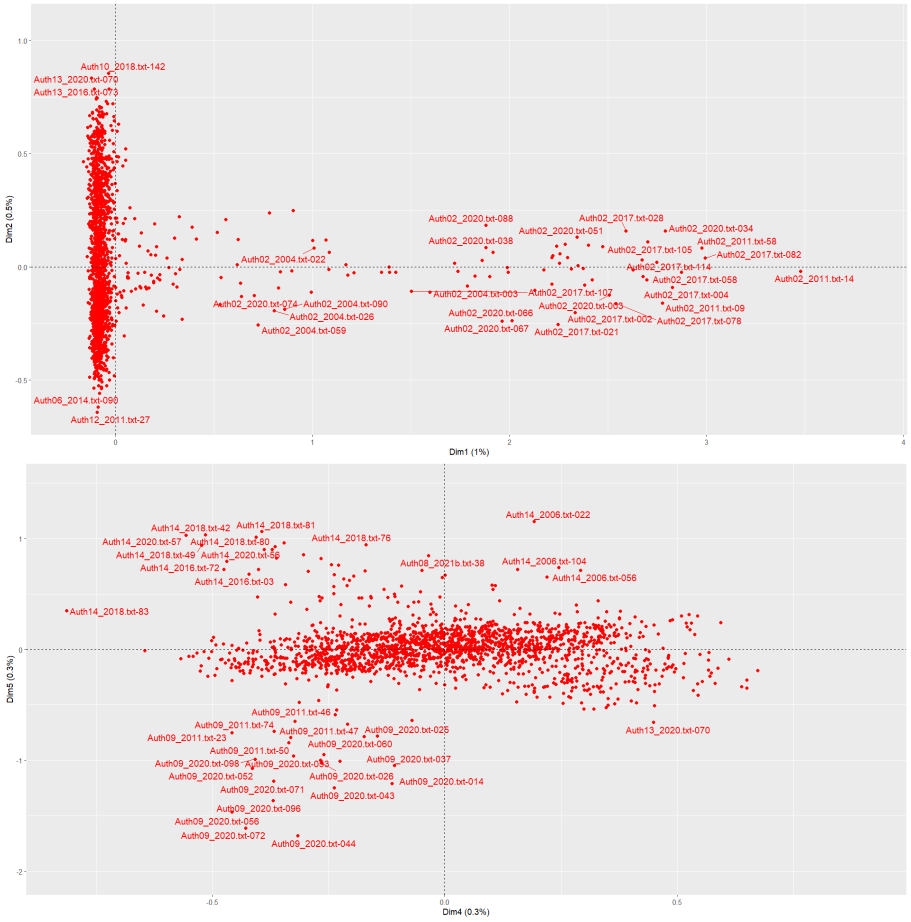

Figure 3: Projection of the chunks belonging to the test set onto the Cartesian planes formed by dimensions 1-2 (above) and dimensions 4-5 (below) of the CA.

References

1. Ai, B., Wang, Y., Tan, Y., & Tan, S. (2022). *Whodunit? Learning to Contrast for Authorship Attribution* (arXiv:2209.11887). arXiv. http://arxiv.org/abs/2209.11887
2. Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.
3. Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2019). *xgboost: Extreme gradient boosting*. R package version 0.81.0.1, 1–4
4. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423
5. Greenacre, M. J. (1984). Theory and applications of correspondence analysis. London: Academic Press.
6. Jones, K., Nurse, J. R. C., & Li, S. (2022). Are You Robert or RoBERTa? Deceiving Online Authorship Attribution Models Using Neural Text Generators. *Proceedings of the International AAAI Conference on Web and Social Media*, *16*, 429–440. https://doi.org/10.1609/icwsm.v16i1.19304
7. Kuhn, M. (2008), "Building predictive models in R using the caret package" *J. of Statistical Softw.*, (doi:10.18637/jss.v028.i05).
8. Lebart, L. (1997). Correspondence analysis, discrimination, and neural networks. In *Data Science, Classification, and Related Methods*. Hayashi C., Ohsumi N., Yajima K., Tanaka Y., Bock H.- H. and Baba Y. (eds), Springer, Berlin, 423-430.
9. Murtagh, F. (2005). *Correspondence analysis and data coding with Java and R*. CRC Press.
10. Rodella, I., Sciandra, A., & Tuzzi, A. (forthcoming). Analysis of Marie Skłodowska-Curie Actions (MSCA) evaluations and models for predicting the success of proposals*, JADT - 17th International Conference on Statistical Analysis of Textual Data*.
11. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* , 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6