The Journal of Open Source Software

idiolect: An R package for forensic authorship analysis

Andrea Nini [1¶]

1 University of Manchester, UK ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Samuel Forbes

Reviewers:

- @stefanocoretta
- @cmaimone

Submitted: 29 August 2024
Published: unpublished

Summary

Authorship Analysis is defined as the task of determining the likelihood that a certain individual is the author of a certain set of questioned texts. This determination is done by analysing the language of the questioned texts and the language of samples produced by the candidate author or authors. This kind of analysis is often applied in the context of literary problems (e.g. Robert Galbraith as J.K. Rowling's alias (Juola, 2015), the identity of Elena Ferrante (Tuzzi & Cortelazzo, 2018)), historical problems (e.g. Lincoln's Bixby letter (Grieve et al., 2019), the Jack the Ripper letters (Nini, 2018), the writings of Julius Caesar (Kestemont et al., 2016)) or forensic contexts (e.g. the *devil strip* ransom letter (Leonard, 2005), the Amanda Birks murder (Grant, 2013) or the Ayia Napa rape statements (Donlan & Nini, 2022)).

Especially when dealing with a forensic problem, the best practice is to carry out authorship analysis within the Bayesian Likelihood Ratio Framework for expressing evidence in forensic science, which is logically aligned with the role of the expert witness in a court of law. Rather than expressing a final binary judgement (same author vs. different author, or author A vs. author B), the framework instead leads the analyst to express the strength of the linguistic evidence in favour or against a set of two competing hypotheses, for example:

H_p: The candidate author and the author of the questioned text are the same individual.
H_d: The candidate author and the author of the questioned text are two different individuals.

In this way, the analyst can assist the decision maker, the judge/jury or perhaps an historian, to reach a verdict that often needs to take into account evidence and information that is not just linguistic.

Statement of need

Within this context, `idiolect` is an R package that contains functions to pre-process datasets, run state-of-the-art authorship analysis algorithms, calibrate the results using the Likelihood Ratio Framework, and then explore the results. `idiolect` is fundamentally based on `quanteda` (Benoit et al., 2018) for the Natural Language Processing functions and this allows its objects and outputs to be handled efficiently using `quanteda`'s own functions if needed. By being based on `quanteda`, the functions in `idiolect` can efficiently handle very large matrices and can therefore process data quickly or handle very large datasets. This factor may lead to significant advantages in performance compared to other R packages for authorship analysis such as `stylo` (Eder et al., 2016). In addition to this advantage, `idiolect` also offers recent authorship analysis algorithms that are currently not widely available, especially in R, such as the *Ranking-Based* variant of the *Impostors Method* (Potha & Stamatatos, 2017, 2020), *N-gram Tracing* and several of its variants (Grieve et al., 2019; Nini, 2023), and *LambdaG* (Nini et al., forthcoming).

Most significantly, what sets `idiolect` apart is its use of the Likelihood Ratio Framework. Through a suite of functions, `idiolect` facilitates the calibration of likelihood ratios from the results of any of the authorship analysis functions and then the assessment of the performance

of this likelihood ratio using standard performance metrics, such as the C_{llr} (Ramos et al., 2013).

Another novelty in `idiolect` is that the package also offers functions that aid the *post-hoc* interpretation of the results. Computational authorship analysis techniques are often hard to interpret by the analyst. Although this is true, for example, for algorithms such as the *Impostors Method* that are based on the frequency of short sequences of characters, `idiolect` facilitates interpretation by returning the most important features and allowing the user to see these features in context. For *LambdaG*, a purpose-built function can return a colour-coded heat map of a text highlighting the words or constructions that influenced the results.

Although `idiolect` has been designed for research in authorship analysis, stylometry, digital humanities, and forensic linguistics, it can also be used effectively to run analyses for real-life forensic linguistics casework. The code being open source is particularly important in a forensic context to allow opposing experts to replicate the analysis and scrutinize the procedure in full.

Acknowledgements

I would like to thank Shunichi Ishihara and Marie Bojsen-Møller for helpful comments on the documentation of this package.

References

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). Quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30). https://doi.org/10.21105/joss.00774

Donlan, L., & Nini, A. (2022). *A forensic authorship analysis of the ayia napa rape statement* (I. Picornell, R. Perkins, & M. Coulthard, Eds.; pp. 29–43). Wiley-Blackwell.

Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with r: A package for computational text analysis. *The R Journal*, *8*(1), 1–15.

Grant, T. (2013). TXT 4N6: Method, consistency, and distinctiveness in the analysis of SMS text messages. *Journal of Law and Policy*, *21*, 467–494.

Grieve, J., Chiang, E., Clarke, I., Gideon, H., Heini, A., Nini, A., & Waibel, E. (2019). Attributing the bixby letter using n-gram tracing. *Digital Scholarship in the Humanities*, *34*(3), 493–512.

Juola, P. (2015). The rowling case: A proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanities*, *30*(suppl_1), i100–i113. https://doi.org/10.1093/llc/fqv040

Kestemont, M., Stover, J., Koppel, M., Karsdorp, F., & Daelemans, W. (2016). Authenticating the writings of julius caesar. *Expert Systems With Applications*, *63*, 86–96. https://doi.org/10.1016/j.eswa.2016.06.029

Leonard, R. (2005). Forensic linguistics: Applying the scientific principles of language analysis to issues of the law. *The International Journal of the Humanities*, *3*.

Nini, A. (2018). An authorship analysis of the jack the ripper letters. *Digital Scholarship in the Humanities*, *33*(3), 621–636.

Nini, A. (2023). *A theory of linguistic individuality for authorship analysis*. Cambridge University Press.

Nini, A., Halvani, O., Graner, L., Titze, S., Gherardi, V., & Ishihara, S. (forthcoming). Grammar as a behavioral biometric: Using cognitively motivated grammar models for

authorship verification. *Humanities and Social Sciences Communications*. Forthcoming. https://arxiv.org/abs/2403.08462

Potha, N., & Stamatatos, E. (2017). An Improved Impostors Method for Authorship Verification. In G. J. F. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeuriot, T. Mandl, L. Cappellato, & N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (Vol. 10456, pp. 138–144). Springer, Cham. ISBN: 978-3-319-65813-1

Potha, N., & Stamatatos, E. (2020). Improved algorithms for extrinsic author verification. *Knowledge and Information Systems*, *62*(5), 1903–1921. https://doi.org/10.1007/s10115-019-01408-4

Ramos, D., Gonzalez-Rodriguez, J., Zadora, G., & Aitken, C. (2013). Information-Theoretical Assessment of the Performance of Likelihood Ratio Computation Methods. *Journal of Forensic Sciences*, *58*(6), 1503–1518. https://doi.org/10.1111/1556-4029.12233

Tuzzi, A., & Cortelazzo, M. A. (2018). What is elena ferrante? A comparative analysis of a secretive bestselling italian writer. *Digital Scholarship in the Humanities*, *33*(3), 685–702. https://doi.org/10.1093/llc/fqx066