

Neurobiber: Fast and Interpretable Stylistic Feature Extraction

Kenan Alkiek
University of Michigan
kalkiek@umich.edu

Anna Wegmann
Utrecht University
a.m.wegmann@uu.nl

Jian Zhu
University of British Columbia
jian.zhu@ubc.ca

David Jurgens
University of Michigan
jurgens@umich.edu

Abstract

Linguistic style is central to how texts convey meaning and to tasks from register analysis to authorship verification, yet extracting detailed stylistic features at scale is hard: rule-based taggers are interpretable but slow, while neural style embeddings are fast but opaque. We present NEUROBIBER, a RoBERTa tagger that predicts 96 interpretable features from Biber’s Multidimensional Analysis in a single forward pass, bringing this analysis to corpus scale: a one-billion-token corpus tags on a single GPU in about 9.5 hours, over an order of magnitude faster than single-thread taggers. NEUROBIBER is trained on BIBERPLUS, a new open-source rule tagger that computes the same 96 features deterministically from explicit rules and ships built-in analytics such as PCA and factor analysis; users tag at scale with NEUROBIBER and fall back to BIBERPLUS for exact, auditable counts. NEUROBIBER reproduces BIBERPLUS at 0.97 macro-F1; its features recover Biber’s primary dimension of variation, and on PAN 2020 authorship verification they contribute signal complementary to a strong neural encoder. Both tools are released openly, with a live demo.

1 Introduction

Style shapes how a text is understood, situating it socially through choices of register and form (Coupland, 2007; Eckert, 2012). Representing and comparing style across texts underpins applications from style transfer (Shen et al., 2017; Fu et al., 2018; Dai et al., 2019) and machine translation (Niu et al., 2018; Rabinovich et al., 2017) to forensic authorship verification (Boenninghoff et al., 2019; Rivera-Soto et al., 2021; Wegmann et al., 2022; Manolache et al., 2022).

Biber’s Multidimensional Analysis (MDA) provides a validated, interpretable feature set for these purposes (Biber, 1988), used most recently to identify systematic grammatical differences between

LLM-generated and human writing (Reinhart et al., 2025), but the tools that extract it are rule- or parser-based and do not scale to large corpora (Nini, 2019; Wilson et al., 2010); neural style representations scale well but typically are not interpretable (Fincke and Boschee, 2024; Patel et al., 2025). We close this gap with a single toolkit that is interpretable and fast, for corpus linguists, stylometry and forensic-linguistics researchers, and NLP practitioners who need style features at scale.

NEUROBIBER is a RoBERTa tagger that predicts 96 Biber features in a single forward pass, reproducing the rule-based labels at macro-F1 0.97 and micro-F1 0.98 while running over an order of magnitude faster, which makes corpus-scale MDA practical (§4.2). It is trained on BIBERPLUS, a new open-source library that computes those 96 features deterministically through explicit rules over a syntactic parse and ships integrated analytics such as PCA and factor analysis; BIBERPLUS installs in one line and provides the reference labels.

We make three contributions: (i) BIBERPLUS, a new open-source, tested rule tagger that implements 96 features, Biber’s original set together with later social-media extensions, and computes them deterministically from explicit rules, with integrated analytics (PCA, factor analysis); (ii) NEUROBIBER, a neural tagger that reproduces Biber’s established feature set and brings Multidimensional Analysis to corpus scale; and (iii) evidence that the neural surrogate preserves the interpretable feature space, recovering Biber’s primary dimension of variation and contributing authorship-verification signal complementary to a strong neural encoder. The library and model are publicly available¹.

¹Library: <https://github.com/davidjurgens/biberplus>; model: <https://huggingface.co/Blablablab/neurobiber/>.

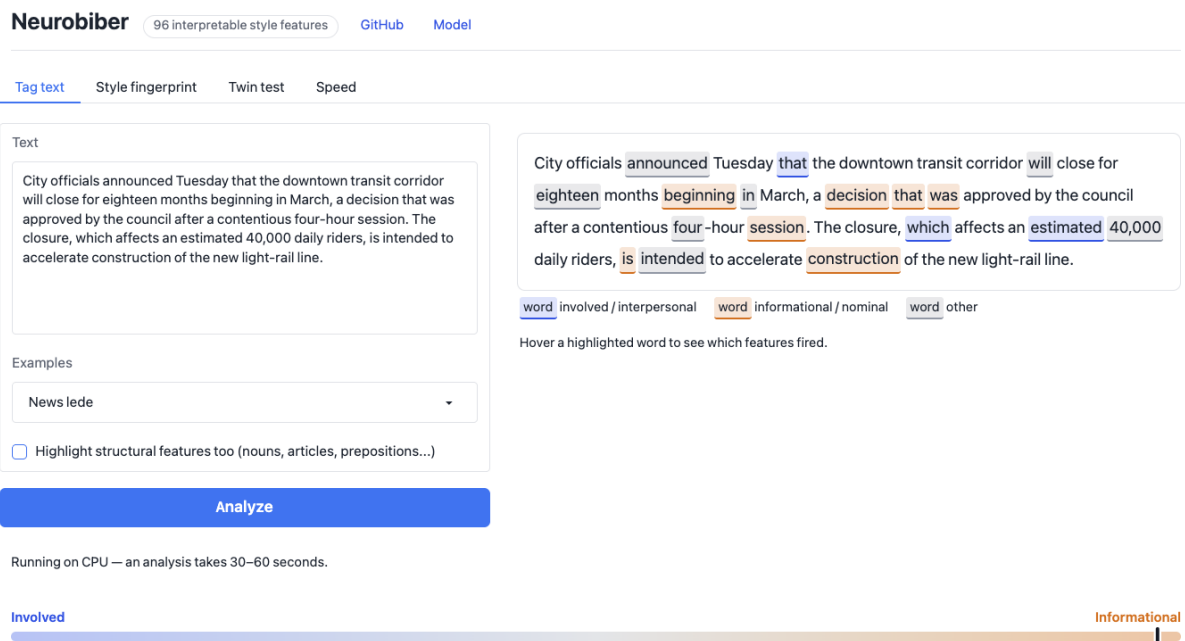


Figure 1: **The hosted NEUROBIBER demo tagging text in real time.** Pasting a passage highlights every word that triggers a Biber feature. Each highlight is a named feature the user can inspect; the tabs expose the style fingerprint, an authorship “twin test,” and a speed benchmark.

2 Background

Linguistic style is the patterned use of lexical, syntactic, and discourse features that varies across authors, groups, and communicative situations (Biber and Conrad, 2019; Crystal and Davy, 1969). Biber’s Multidimensional Analysis operationalizes this variation as counts of interpretable features (e.g., pronouns, passives, stance markers, subordination) that combine into empirically derived dimensions such as *Involved* versus *Informational Production* (Biber, 1988; Biber and Conrad, 2009). These features are transparent and linguistically grounded, but Biber’s original tagger is not publicly available (Nini, 2019), and MDA studies vary widely in the taggers they use instead (Gray, 2019; Goulart and Wood, 2021). Several open tools extract Biber-style features: the Multidimensional Analysis Tagger (MAT) (Nini, 2019), biberpy (Sharoff, 2021), profiling-UD (Brunato et al., 2020), the Multi-Feature Tagger of English (Le Foll and Shakir, 2024), and pseudo-bibeR with its Python reimplement pybiber (Brown and Reinhart, 2024; Brown, 2025), alongside lexicon counters such as LIWC (Tausczik and Pennebaker, 2010). Later work extends this framework to computer-mediated communication: Clarke and Grieve (2017) and Clarke and Grieve (2019) add social-media features (emoji, hash-

tags, @-mentions, laughter acronyms) and, because tweets are far too short for reliable frequency estimates, record feature presence rather than relative frequency. A parallel line of toolkits computes broad batteries of handcrafted linguistic or stylistic features—for example LFTK (Lee and Lee, 2023), StyloMetrix (Okulska et al., 2023), gram2vec (Zeng et al., 2024), and the efficiency-oriented elfen (Maurer, 2026)—but these target general-purpose feature inventories rather than Biber’s smaller, functionally motivated set that combines into interpretable dimensions of variation. Crucially, all of them, Biber-specific or general, are rule- or parser-based, so their per-document parsing cost makes corpus-scale analysis expensive. BIBERPLUS (§3.1) stays in this tradition but is built for scale and reuse: it expands the feature set to 96 and ships as a tested, pip-installable Python library with a scriptable API, a command-line interface, and built-in analytics. It also computes 96 features to pybiber’s 67 and parallelizes natively across cores (§4.2).

Neural methods take the opposite trade-off. Authorship-representation models learn style embeddings by contrastive training over large author populations (Andrews and Bishop, 2019; Rivera-Soto et al., 2021; Wegmann et al., 2022): they are fast and accurate, but their dimensions are opaque,

and Wang et al. (2023) find it unclear how much stylistic, as opposed to topical, information such embeddings encode. LISA (Patel et al., 2023) and StyleDistance (Patel et al., 2025) restore some interpretability by learning named style attributes, but these are elicited from large language models or synthetic paraphrases and sit at a higher level than Biber’s empirically grounded grammatical features. These lines leave a gap: representations that are both interpretable and fast. NEUROBIBER scales an interpretable tagger to corpus size with weak supervision: BIBERPLUS automatically labels a large corpus, and NEUROBIBER learns to reproduce those *silver labels* in one pass (a hard-label form of model distillation (Hinton et al., 2015; Kim and Rush, 2016)), staying grounded in an auditable rule set. NEUROBIBER therefore offers the same interpretable feature vocabulary as BIBERPLUS, though per-instance rule audibility remains a property of BIBERPLUS itself.

3 System Overview

We release a single toolkit for interpretable stylistic analysis with two complementary taggers. BIBERPLUS is a rule-based tagger that applies explicit patterns to a syntactic parse, producing exact counts for 96 Biber-style features and providing the reference labels. NEUROBIBER is a neural surrogate trained on those labels: it predicts the same 96 features in one pass, trading exact counts for the throughput needed to tag corpora at scale.

3.1 BiberPlus: the Rule-based Reference Tagger

BIBERPLUS extends the framework of Biber (1988) with an expanded feature set and integrated analytics. It parses text with spaCy (Honnibal et al., 2020) and applies hand-crafted rules to the resulting trees, tracking 96 features: Biber’s original 67, plus 29 from later work on computer-mediated communication (Clarke and Grieve, 2017, 2019) that Biber’s 1988 set predates, such as emoji, hash-tags, @-mentions, URLs, and laughter acronyms (full list in Appendix A). Each rule encodes an explicit construction (e.g., agentless passives, subordination, stance markers), so every count is faithful to a stated definition. To keep the rules correct, BIBERPLUS ships 288 passing tests.

Counting modes. BIBERPLUS offers two counting modes so the same features apply to a tweet or a novel. *Frequency* mode reports how often each

feature occurs per window (aggregating mean, minimum, maximum, and standard deviation across windows) and suits long texts. *Binary* mode records only presence within a window (a feature that occurs several times still contributes a single 1) and reports the fraction of windows containing it, which is more stable for short texts such as social-media posts, following the presence-based Multidimensional Analysis that Clarke and Grieve (2019) develop for texts too short for reliable frequency estimates. We use 100-token windows throughout. BIBERPLUS installs with a single `pip install`.²

3.2 Neurobiber: a Neural Tagger for Scale

NEUROBIBER makes Biber tagging fast enough for corpus-scale study. We fine-tune RoBERTa (Liu et al., 2019) as a 96-way multi-label classifier trained to reproduce BIBERPLUS’s labels on a large multi-domain corpus (Table 1): BIBERPLUS labels every document, and NEUROBIBER learns to predict those labels in one forward pass. Because each output corresponds to a named Biber feature, NEUROBIBER shares BIBERPLUS’s interpretable feature vocabulary, though only BIBERPLUS is auditable rule-by-rule for an individual document. It reproduces BIBERPLUS closely and runs over an order of magnitude faster; we quantify both fidelity and speed in §4. NEUROBIBER is released on HuggingFace under the same license.³

Labels and inference. Feature presence is defined on 100-token windows: a document-level label is positive when the feature appears in any window. NEUROBIBER consumes each document as a ≤ 512 -token input during training; at inference, long documents are scored in consecutive 512-token spans whose predictions are OR-ed together. NEUROBIBER predicts presence rather than exact counts; when exact frequencies are needed, BIBERPLUS remains the tool of choice.

3.3 Using the Toolkit

Users choose a tagger by scale and need: NEUROBIBER for large corpora or short-to-medium texts where speed matters, and BIBERPLUS when exact counts or long-form parsing are required. Both are driven from the same Python API or a command-line interface, and either returns the same 96-feature representation, which feeds BIBERPLUS’s built-in PCA and factor analysis directly, taking a

²<https://github.com/davidjurgens/biberplus>

³<https://huggingface.co/Blablablab/neurobiber/>

Source	Description	Authors	Texts	Avg. Texts per Author	Avg. Text Tokens
Wikipedia editing histories	Page edit histories	506,992	3,134,068	6.18	232
Wikipedia discussions	Article-editing discussions	188,663	4,204,118	22.28	116
Book3Corpus	Published novels	66,385	1,276,515	19.22	700
Amazon reviews	Product reviews	128,495	1,771,985	13.79	143
Gmane mailing list	Mailing-list archive (since 2002)	693,135	4,820,910	6.95	123
Realnews article collection	News articles (Common Crawl)	212,970	15,874,737	74.53	609
Reddit	User comments (15 years)	2,459,016	12,290,058	4.99	117

Table 1: **Multi-domain training corpus for NEUROBIBER.** Seven sources totaling ~ 43.4 M source texts; we train on the ~ 39.6 M non-empty texts that BIBERPLUS tags with at least one feature.

stylistic study over millions of documents from raw text to an analyzable feature matrix.

3.4 Interactive Demo

A hosted, interactive demo accompanies the release.⁴ Its *Tag text* tab tags pasted text in the browser, highlighting each word that fires a Biber feature and placing the passage on Biber’s Involved–Informational dimension (Figure 1). Three further tabs run common analyses over the same 96 features: *Style fingerprint* profiles a text and overlays a second for comparison, *Twin test* applies the authorship-verification pipeline of §5.2 to a pair of texts, and *Speed* races NEUROBIBER against the parser-based baselines on the visitor’s own machine. Because every prediction resolves to a named, inspectable feature, the demo doubles as a way to audit NEUROBIBER’s outputs.

4 Evaluation

We evaluate the toolkit on the two properties a corpus-scale tagger must deliver: how faithfully NEUROBIBER reproduces the rule-based BIBERPLUS labels (§4.1), and how fast each tagger runs end to end (§4.2).

4.1 Fidelity to BIBERPLUS

Because NEUROBIBER is trained to reproduce BIBERPLUS, we first ask how faithfully it does so. On the held-out test set it reaches **macro-F1 0.97** and **micro-F1 0.98**, with well-calibrated predictions (expected calibration error 0.007). Agreement is high per feature as well (Figure 2): 93 of the 96 features exceed 0.90 F1. Fidelity holds across all seven training domains (Table 2), with per-domain macro-F1 from 0.925 (books) to 0.985 (Reddit). The three that fall below—emoji, discourse particles, and past-participial clauses—are the rarest

Source	Docs	Macro-F1	Micro-F1
Wikipedia edits	313,348	0.947	0.983
Wikipedia discussions	409,383	0.975	0.993
Book3Corpus	127,720	0.925	0.960
Amazon reviews	177,185	0.973	0.993
Gmane	560,071	0.968	0.986
Realnews	1,144,278	0.958	0.981
Reddit	1,228,292	0.985	0.993

Table 2: **Fidelity by training domain.** NEUROBIBER’s macro-F1 and micro-F1 against BIBERPLUS on the held-out test set, by source corpus. Books are the weakest domain (0.925) and Reddit the strongest (0.985).

or most syntactically complex tags; emoji, for example, have only 0.01M positive test instances yet reach 0.99 F1 on Reddit, where they are common; these tags should improve with more in-domain data. Per-feature statistics are in Appendix B.

4.2 Speed and Scale

Corpus-scale Multidimensional Analysis is impractical with existing tools: parser-based taggers such as MAT (Nini, 2019), biberpy (Sharoff, 2021), and profiling-UD (Brunato et al., 2020) run on CPU and tag one document at a time. On one core, BIBERPLUS runs at about $2.2\times$ the throughput of MAT and pybiber at about $1.6\times$ BIBERPLUS (Figure 3; protocol in Appendix C). Because NEUROBIBER produces the *same* 96 features as BIBERPLUS, we take BIBERPLUS as the apples-to-apples reference (external taggers compute different feature sets) with MAT and pybiber as external points of comparison, on the same 2.4M-token, seven-domain sample and machine (Intel Xeon Gold 5317, 24 cores; NVIDIA RTX A6000). To make the CPU baseline as strong as possible we parallelize BIBERPLUS across cores: its throughput peaks at 14,076 tokens/s with 16 workers, $5.4\times$ the single-core rate and about $3\times$ pybiber’s; parallelism is native to the library (an `n_processes` argument). Figure 3 reports end-to-end throughput, timing every method on its full pipeline (tokenization or parsing plus

⁴<https://huggingface.co/spaces/Blablalab/neurobiber-demo>

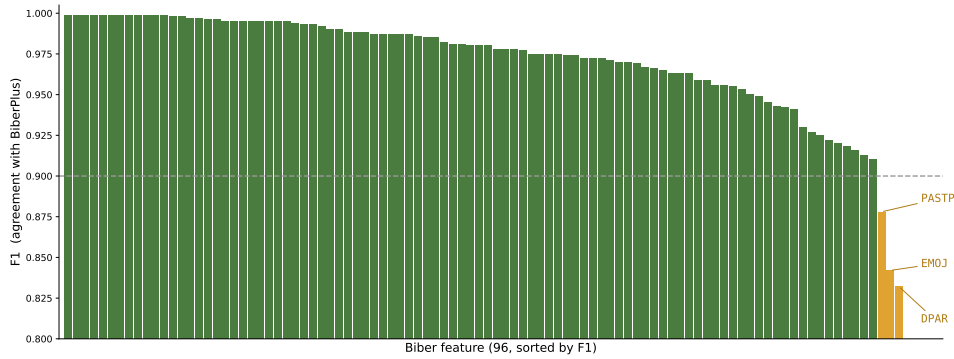


Figure 2: **High per-feature agreement with BIBERPLUS.** Per-feature F1 between NEUROBIBER and the rule-based BIBERPLUS labels on the held-out test set, sorted; the dashed line marks 0.90. 93 of 96 features exceed it; the three below (PASTP, EMOJ, DPAR) are the rarest or most syntactically complex tags. Residual error is dominated by misses: precision exceeds recall for 70 of the 96 features. Per-feature statistics are in Table 4.

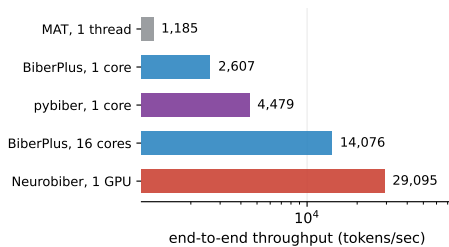


Figure 3: **NEUROBIBER tags faster than parser-based taggers.** End-to-end throughput (tokens/s, log scale) on the same 2.4M-token sample and machine, each method timed on its full pipeline (protocol in Appendix C). On one GPU NEUROBIBER (29,095 tok/s) is $2.1\times$ a fully parallelized 16-core BIBERPLUS and about $25\times$ the single-thread MAT (Nini, 2019).

tagging). End to end, NEUROBIBER tags at 29,095 tokens/s: $2.1\times$ the throughput of a fully parallelized 16-core BIBERPLUS (14,076 tokens/s) and about $25\times$ the single-thread MAT baseline (1,185 tokens/s).

When to use which tagger. The two taggers are complementary. Below roughly 50M tokens, 16-core BIBERPLUS finishes within an hour and returns exact counts, so the rule-based tagger is the convenient choice; above that scale NEUROBIBER becomes the practical option. Extrapolating from measured throughput, a one-billion-token corpus (a large social-media or news study) is taggable on a single GPU in about 9.5 hours, versus about 20 hours for a parallelized 16-core BIBERPLUS. NEUROBIBER thus brings corpus-scale Multidimensional Analysis within reach of a single machine, without giving up the interpretable feature set.

5 Use Cases

We put the toolkit to work on two standard applications of stylistic features: recovering Biber’s dimensions of variation from its features (§5.1) and authorship verification (§5.2). These are use cases that show what the tagger makes easy at scale, not attempts to top task-specific leaderboards. Both run on the neural tagger and so scale to large corpora.

5.1 Recovering Biber’s Dimensions

A principal component analysis of NEUROBIBER features recovers Biber’s dimensions of variation. We tag 33,310 documents from the Corpus of Online Registers of English (CORE), a register-annotated web corpus whose documents are hand-labeled with situational registers such as news reports, personal blogs, and how-to pages (Egbert et al., 2015; Biber et al., 2015), and apply PCA to the resulting NEUROBIBER vectors; if the neural features carry the same information as Biber’s, their principal components should recover his dimensions (Biber, 1988). Figure 4 plots the registers on the first and third components. PC1 (16.9% of variance) orders registers by overall feature density. PC3 (3.0%) separates involved, interactive registers (discussion forums, Q&A) from informational ones (research and encyclopedia articles): its involved end loads on first- and second-person pronouns, interjections, and exclamation marks, its informational end on passives and participial clauses. This is Biber’s Dimension 1, Involved versus Informational Production, recovered from features the neural model predicted in a single pass. The interpretable structure survives the neural surrogate: NEUROBIBER’s features are not only fast

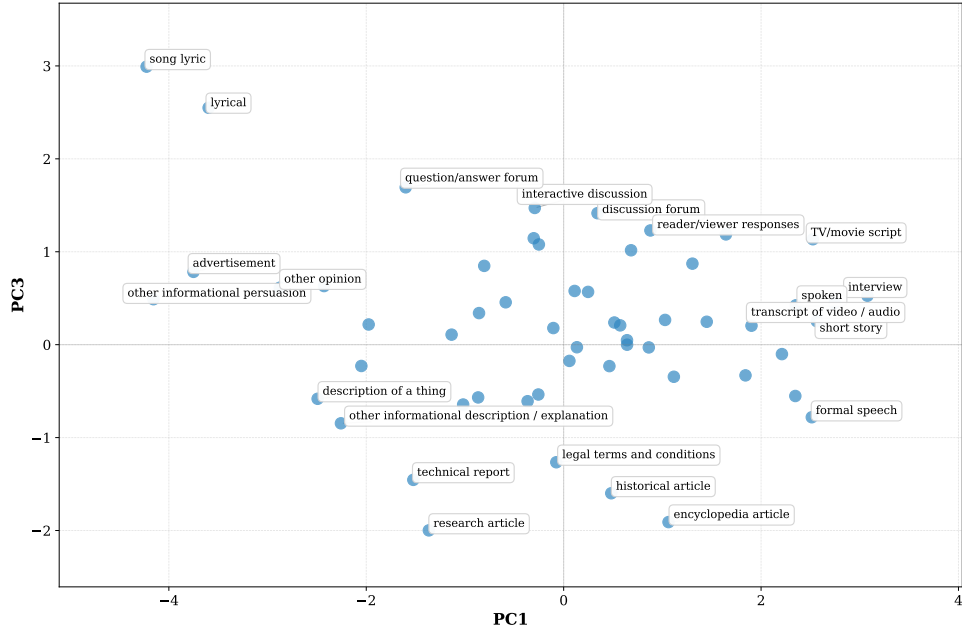


Figure 4: **NEUROBIBER features recover Biber’s Dimension 1.** PCA of NEUROBIBER feature vectors for 33,310 CORE documents; each point is one of 56 registers, placed by its mean score on the two components. PC3 separates involved registers at the top (forums, Q&A, lyrics) from informational ones at the bottom (research, encyclopedia, legal text), reproducing Biber’s Involved versus Informational contrast; PC1 reflects overall feature density.

to compute but remain analyzable in Biber’s own dimensions.

5.2 Authorship Verification

As a second use case, we apply NEUROBIBER features to authorship verification on PAN 2020 (small) (Kestemont et al., 2020), which asks whether two fanfiction texts share an author. Stylistic features are a long-standing tool for this task; our aim is not to top the leaderboard but to show that NEUROBIBER makes a large, interpretable, corpus-scale feature set readily available for the task. Fandom shifts topic and register within and across authors, so a reliable signal must come from stable style rather than content, which makes the task a direct test of interpretable style features. We represent each text by its 96 NEUROBIBER features and classify a pair with a random forest over $[A \mid B \mid |A - B|]$ (the two feature vectors and their absolute difference). We compare against a random baseline, a contrastive RoBERTa bi-encoder (the original baseline), and three style or authorship representations under the same protocol: BIBERPLUS features, LUAR (Rivera-Soto et al., 2021), and StyleDistance (Patel et al., 2025). Table 3 reports weighted F1 and ROC-AUC over five seeds.

On its own, NEUROBIBER reaches 0.795 weighted F1, within half a point of BIBERPLUS

Model	F1 _w	ROC-AUC
Random	0.505	0.504
StyleDistance	0.727	0.800
RoBERTa bi-encoder	0.786	0.867
LUAR	0.798	0.883
BIBERPLUS	0.799	0.879
NEUROBIBER	0.795	0.878
NEUROBIBER + LUAR	0.835	0.916
BIBERPLUS + LUAR	0.839	0.918
RoBERTa bi-encoder + LUAR	0.804	0.889
NEUROBIBER + StyleDistance	0.775	0.855

Table 3: **Biber features complement a strong authorship encoder.** Authorship verification on PAN 2020 (small; test): weighted F1 (F1_w) and ROC-AUC, means over five seeds (all standard deviations ≤ 0.007); **bold** marks the best. Concatenating LUAR with NEUROBIBER (0.835) or BIBERPLUS (0.839) lifts weighted F1 by about four points over LUAR alone (0.798); concatenating a second neural encoder does not (0.804). Dimensionalities and per-seed statistics are in Table 5.

(0.799) and LUAR (0.798) and above the bi-encoder (0.786), using the fewest dimensions of any method. The main result is complementarity: concatenating NEUROBIBER with the black-box LUAR raises weighted F1 to 0.835 and ROC-AUC from 0.883 to 0.916, a gain that holds on every seed (paired $t(4)=42.7$, $p<10^{-4}$; Table 5). The gain comes from the Biber feature space rather than

from combining encoders: BIBERPLUS + LUAR gains as much (0.839), while a second neural encoder adds little (bi-encoder + LUAR, 0.804). NEUROBIBER recovers about 90% of BIBERPLUS’s added value at a fraction of the tagging cost. The effect also requires a strong encoder: pairing NEUROBIBER with the weaker StyleDistance does not help (0.775).

Because every feature has an explicit name and rule, a predicted match is auditable feature by feature. For one correctly classified same-author pair, the random forest’s strongest shared drivers are interjections (UH), exclamation marks (EXCL), time adverbials (TIME), and the pronoun *it* (PIT), each a rule-defined tag a user can inspect. The black-box baselines offer no comparable account of why two texts match.

6 Conclusion

We presented a toolkit for interpretable stylistic analysis: BIBERPLUS, a rule-based tagger with integrated analytics (PCA, factor analysis), and NEUROBIBER, a neural tagger trained on its labels that predicts 96 Biber features in a single forward pass. NEUROBIBER reproduces BIBERPLUS with 0.97 macro-F1 at throughput that makes corpus-scale Multidimensional Analysis practical (§4.2). Its features recover Biber’s primary dimension of variation, Involved versus Informational Production (§5.1), and add predictive value when concatenated with a strong authorship encoder on PAN 2020 (§5.2). Because these features are interpretable, cheap to extract, and complementary to neural embeddings, they suit corpus-scale stylistic monitoring and can augment authorship pipelines with transparent signals. Both tools are openly released.

Limitations

NEUROBIBER predicts feature *presence*, not exact counts; applications that need frequencies should use BIBERPLUS. Its fidelity is measured as agreement with BIBERPLUS’s automatically produced (silver) labels, not against human judgments, so NEUROBIBER inherits the rules’ errors and cannot exceed them except where it generalizes; the rules’ correctness rests on our unit tests rather than on human validation of every feature. Both tools are English-only, whereas some existing taggers are multilingual. Finally, a few rare or syntactically complex features reach lower agreement with

BIBERPLUS (Appendix B) and would benefit from additional training data.

Acknowledgments

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200006. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Ethical Considerations

When applied to authorship identification, stylometric tools carry privacy risks: by exposing distinctive linguistic patterns they can help de-anonymize authors, which is harmful in settings such as whistleblowing or private communication. We therefore urge caution in forensic use of NEUROBIBER: its accuracy on PAN 2020 verification (weighted F1 around 0.80) is well short of reliable identification, and it should not serve as sole evidence of authorship. Interpretability mitigates but does not remove this risk: NEUROBIBER’s features are named, but an individual prediction is not rule-audited and may reflect spurious correlations in the training data, and Biber’s feature set was developed largely on edited English and may encode cultural or register biases. Used for large-scale, aggregate stylistic analysis (rather than identifying specific individuals), these tools support transparent research; used to attribute authorship to a person, they demand the safeguards above. All data were publicly available and processed under applicable terms of service, and authors were anonymized where possible.

References

- Nicholas Andrews and Marcus Bishop. 2019. Learning invariant representations of social media users. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1684–1695.
- Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press.

- Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge Textbooks in Linguistics. Cambridge University Press.
- Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*, 2 edition. Cambridge Textbooks in Linguistics. Cambridge University Press.
- Douglas Biber, Jesse Egbert, and Mark Davies. 2015. Exploring the composition of the searchable web: a corpus-based taxonomy of web registers. *Corpora*, 10(1):11–45.
- Benedikt Boenninghoff, Robert M. Nickel, Steffen Zeiler, and Dorothea Kolossa. 2019. *Similarity Learning for Authorship Verification in Social Media*. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2457–2461. ArXiv:1908.07844 [cs, stat].
- David West Brown. 2025. *pybiber: Aggregate counts of linguistic features*. Python package, version 0.3.1 (2026).
- David West Brown and Alex Reinhart. 2024. *pseudo-bibeR: Aggregate counts of linguistic features*. R package, version 1.2.
- Dominique Brunato, Andrea Cimino, Felice Dell’Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-ud: a tool for linguistic profiling of texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7145–7151.
- Isabelle Clarke and Jack Grieve. 2017. *Dimensions of Abusive Language on Twitter*. In *Proceedings of the First Workshop on Abusive Language Online*, pages 1–10, Vancouver, BC, Canada. Association for Computational Linguistics.
- Isabelle Clarke and Jack Grieve. 2019. *Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018*. *PLOS ONE*, 14(9):e0222062.
- Nikolas Coupland. 2007. *Style: Language Variation and Identity*. Cambridge University Press.
- David Crystal and Derek Davy. 1969. *Investigating English Style*. Longman, London.
- Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. *Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007, Florence, Italy. Association for Computational Linguistics.
- Penelope Eckert. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, 41(1):87–100.
- Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.
- Steven Fincke and Elizabeth Boschee. 2024. Separating style from substance: Enhancing cross-genre authorship attribution through data selection and presentation. *arXiv preprint arXiv:2408.05192*.
- Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 663–670.
- Larissa Goulart and Margaret Wood. 2021. *Methodological synthesis of research using multi-dimensional analysis*. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 6(2):107–137.
- Bethany Gray. 2019. Tagging and counting linguistic features for multi-dimensional analysis. In Tony Berber Sardinha and Marcia Veirano Pinto, editors, *Multi-Dimensional Analysis: Research Methods and Current Issues*. Bloomsbury Academic, London.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. 2020. *spacy: Industrial-strength natural language processing in python*.
- Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2020. Overview of the cross-domain authorship verification task at PAN 2020. In *CLEF 2020 Working Notes*.
- Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.
- Elen Le Foll and Muhammad Shakir. 2024. *The Multi-Feature tagger of English (MFTE): Rationale, description and evaluation*. *Research in Corpus Linguistics*, 13(2).
- Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *Roberta: A robustly optimized bert pretraining approach*.

- Andrei Manolache, Florin Brad, Antonio Barbalau, Radu Tudor Ionescu, and Marius Popescu. 2022. VeriDark: A large-scale benchmark for authorship verification on the dark web. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022) Datasets and Benchmarks Track*.
- Maximilian Maurer. 2026. elfen: A python package for efficient linguistic feature extraction for natural language datasets. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 61–74, Rabat, Morocco. Association for Computational Linguistics.
- Andrea Nini. 2019. The multi-dimensional analysis tagger. *Multi-dimensional analysis: Research methods and current issues*, pages 67–94.
- Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 1008–1021, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Inez Okulska, Daria Stetsenko, Anna Kołos, Agnieszka Karlińska, Kinga Głabińska, and Adam Nowakowski. 2023. StyloMetrix: An open-source multilingual tool for representing stylometric vectors. *arXiv preprint arXiv:2309.12810*.
- Ajay Patel, Delip Rao, Ansh Kothary, Kathleen McKeown, and Chris Callison-Burch. 2023. Learning interpretable style embeddings via prompting LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15270–15290, Singapore. Association for Computational Linguistics.
- Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2025. StyleDistance: Stronger content-independent style embeddings with synthetic parallel examples. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8662–8685, Albuquerque, New Mexico.
- Ella Rabinovich, Shachar Mirkin, Raj Nath Patel, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 1074–1084, Valencia, Spain. Association for Computational Linguistics.
- Alex Reinhart, Ben Markey, Michael Laudenbach, Kachata Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122.
- Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. [Learning universal authorship representations](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Serge Sharoff. 2021. Genre annotation for the web: text-external and text-internal perspectives. *Register studies*, 3(1):1–32.
- Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.
- Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.
- Andrew Wang, Cristina Aggazzotti, Rebecca Kotula, Rafael Rivera Soto, Marcus Bishop, and Nicholas Andrews. 2023. Can authorship representation learning capture stylistic features? *Transactions of the Association for Computational Linguistics*, 11:1416–1431.
- Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. [Same Author or Just Same Topic? Towards Content-Independent Style Representations](#). In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.
- James Wilson, Anthony Hartley, Serge Sharoff, and Paul Stephenson. 2010. Advanced corpus solutions for humanities researchers. In *Proc PACLIC 24*, pages 36–43, Tohoku University.
- Peter Zeng, Hannah Stortz, Eric Sclafani, Alina Shabaeva, Maria Elizabeth Garza, Daniel Greeson, and Owen Rambow. 2024. Gram2Vec: An interpretable document vectorizer. *ArXiv:2406.12131*.

A Biberplus

Feat.	Description	Feat.	Description	Feat.	Description
VBD	Past tense POS	PEAS	Perfect aspect using patterns like HAVE + (ADV) + (ADV) + VBD/VBN	VPRT	Present tense: VBP or VBZ tag
PLACE	Any item in the place adverbials list that is not a proper noun (NNP)	TIME	Time adverbials excluding 'soon' when followed by 'as'	FPP1	First person pronouns
SPP2	Second person pronouns	TPP3	Third person pronouns	PIT	Pronoun 'it' and its related forms
INPR	Indefinite pronouns list	DEMP	Demonstrative pronouns under certain conditions	PROD	Pro-verb 'do' excluding specific patterns
WHQU	Direct WH-questions	NOMZ	Noun endings like -tion, -ment, -ness	GER	Gerunds of length greater than or equal to 10 ending in ing or ings
NN	Total other nouns (not nominalisations or gerunds)	PASS	Agentless passives	BYPA	By-passives following 'PASS' tag
BEMA	Be as the main verb	EX	Existential 'there'	THVC	That-verb complements with specific preceding or succeeding patterns
PASTP	Past participial clauses: punctuation + VBN	WZPAST	Past participial WHIZ-deletion relatives	WZPRES	Present participial WHIZ-deletion relatives: VBG preceded by an NN
TSUB	That-rel. clauses on subject position	TOBJ	That-rel. clauses on object position	WHSUB	WH-rel. clauses on subject position
WHOBJ	WH rel. clauses on object position	PIRE	Pied-piping relatives (prep + whom/who/whose/which)	SERE	Sentence relatives: punctuation + 'which'
CAUS	Any occurrence of the word "because"	CONC	Any occurrence of "although", "though", "tho"	COND	Any occurrence of "if" or "unless"
OSUB	Other adverbial subordinators (multi-word)	PIN	Total prepositional phrases	JJ	Attributive adjectives
PRED	Predicative adjectives	RB	Adverbs (POS tags RB, RBS, RBR, WRB)	CONJ	Conjuncts with preceding punctuation or specific words
DWNT	Specific downtoners	AMP	Specific amplifiers	DPAR	Discourse particle: pattern preceded by punctuation
HDG	Hedge words	EMPH	Emphatics	DEMO	Demonstratives
POMD	Possibility modals	NEMD	Necessity modals	PRMD	Predictive modals
PUBV	Public verbs	PRIV	Private verbs	SUAV	Suasive verbs
SMP	Seem/Appeal verbs	CONT	Reduced forms (contractions)	THATD	Subordinator "that" deletion
STPR	Stranded preposition	SPIN	Split infinitives	SPAU	Split auxiliaries
PHC	Phrasal coordination	ANDC	'and' in coordination of independent clauses	XX0	Words indicating analytic negation ('not')
SYNE	Synthetic negations before adj./noun	QUAN	Quantifier words	QUPR	Quantifier pronouns
ART	Articles ('the', 'a', 'an')	AUXB	Auxiliary verbs that are forms of 'be'	CAP	Words starting with a capital letter
SCONJ	Subordinating conjunctions	CCONJ	Coordinating conjunctions	DET	Determiners
EMOJ	Emojis	EMOT	Common emoticons	EXCL	Exclamation mark '¡'
HASH	Words starting with #	INF	Infinitive verb forms	UH	Interjections
NUM	Numerals	LAUGH	Laughter acronyms like 'lol'	PRP	Possessive pronouns
PREP	Prepositions	NNP	Proper nouns	QUES	Question mark '¿'
QUOT	Quotation marks	AT	Words starting with '@'	SBJP	Subject pronouns
URL	URLs	WH	WH words ('who', 'what', 'where')	INDA	Indefinite articles ('a', 'an')
ACCU	Accusative case pronouns	PGAS	Progressive aspect verb forms	CMADJ	Comparative adjectives

SPADJ	Superlative adjectives	X	Words not fitting other POS categories	AWL	Mean word length
TTR	Type-token ratio	TRB	Total adverbs		

B Neurobiber

B.1 Training Details

We trained a RoBERTa-base model in a multi-label classification setup (96 labels) using `adamw_torch` with a learning rate of $2e-5$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, an ϵ of $1e-8$, and a weight decay of 0.01 . Each mini-batch contained 256 sequences, and training was performed for 8 epochs on a dataset of approximately 31.7 million examples, with a dev and test size of about 4 million examples each. We used gradient checkpointing to reduce memory usage and bf16 precision throughout training; the random seed was fixed at 42. A linear scheduler with no warmup steps was applied, and the best checkpoint was selected via validation.

Data splits and labels. The concatenated corpus is split 80/10/10 within successive 50,000-document chunks (shuffled with a fixed seed), and there is no author-level deduplication across splits. Because the target is fidelity to a deterministic rule set rather than a human-labeled task, author overlap between splits inflates agreement far less than it would inflate a task score. BIBERPLUS labels are computed on the full document (a feature is positive if it fires in any 100-token window) before the model input is truncated to 512 tokens.

Optimization, selection, and calibration. The loss is binary cross-entropy with logits over the 96 labels; predictions threshold the sigmoid at 0.5. The best checkpoint is chosen by dev macro-F1, evaluated every 1% of training on a fixed 1% subsample of the dev split (~40k documents), with early stopping at patience 5. Expected calibration error uses 15 equal-width bins over the positive-class probability, aggregated over all label–instance pairs, on a 39,603-document sample of the test split.

B.2 Performance by Tag

Feat	Highest F1				Feat	Lowest F1			
	P	R	F1	Supp		P	R	F1	Supp
CAP	1.000	1.000	1.000	3.95M	DPAR	0.764	0.913	0.832	0.28M
NN	1.000	1.000	1.000	3.96M	EMOJ	0.980	0.738	0.842	0.01M
PIN	1.000	1.000	1.000	3.95M	PASTP	0.877	0.880	0.878	0.53M
PREP	1.000	1.000	1.000	3.95M	GER	0.928	0.892	0.910	0.74M
ART	1.000	1.000	1.000	3.93M	UH	0.918	0.908	0.913	1.04M
DET	1.000	0.999	1.000	3.95M	AT	0.994	0.850	0.916	0.10M
AUXB	0.999	0.999	0.999	3.82M	WHOBJ	0.921	0.916	0.918	1.09M
CONJ	0.999	0.999	0.999	3.50M	THAC	0.945	0.897	0.920	0.34M
PRP	1.000	0.998	0.999	3.87M	PRESP	0.892	0.953	0.922	1.08M
INDA	0.999	0.999	0.999	3.64M	EMOT	0.943	0.908	0.925	0.09M

Table 4: **Per-feature fidelity to BIBERPLUS: the ten highest- and ten lowest-F1 features.** Precision, recall, F1, and positive test instances (Supp) between NEUROBIBER and the rule-based BIBERPLUS labels on the held-out test set. Precision exceeds recall for 70 of the 96 features. The full 96-feature table is in the BIBERPLUS repository.

B.3 Authorship-Verification Details

Table 5 reports the dimensionality of each representation in Table 3 with means and standard deviations over seeds. BIBERPLUS vectors are 192-dimensional because the tagger reports a mean and a standard deviation per feature. All systems share seeds 0–4 for the random forest over the same cached embeddings and the same PAN train/test split (42,098 training and 5,262 test pairs), so comparisons are paired and the intervals reflect seed variation only. Against LUAR alone, the gains in weighted F1 are +0.036 for NEUROBIBER + LUAR ($t(4)=42.7$, $p<10^{-4}$, 95% CI [0.034, 0.039]), +0.041 for BIBERPLUS + LUAR ($t(4)=61.5$, $p<10^{-5}$), and +0.005 for bi-encoder + LUAR ($p=0.037$). Compared directly, BIBERPLUS + LUAR exceeds NEUROBIBER + LUAR by 0.004 weighted F1 on all five seeds ($p=0.018$). The twin-test feature importances reported by the demo are global random-forest importances, summed over the two per-side blocks and restricted to features present in both texts; the pair representation $[A \mid B \mid A - B]$ is symmetric in its difference block but not in the concatenation.

C Timing Protocol

All CPU timings use the 2.4M-token seven-domain sample from §4.2 on the same machine (Intel Xeon Gold 5317), end to end, with model loading excluded by a warmup pass. Each configuration runs three times and we report the median; single-core runs are pinned to one core with single-threaded BLAS. `pybiber` (v0.3.1 on `spaCy` 3.8 with `en_core_web_sm`) was timed on a shared node, where background load can only lower measured throughput; BIBERPLUS timed under the same conditions stays within 3–10% of the numbers in §4.2, bounding the effect of that load. One of the 7,000 sample texts fails BIBERPLUS’s tagger and is skipped. Raw timings and environment records are in the repository.

System	Dim.	F1 _w	ROC-AUC
Random	–	0.505 ± 0.007	0.504 ± 0.007
StyleDistance	768	0.727 ± 0.002	0.800 ± 0.000
RoBERTa bi-encoder	768	0.786 ± 0.003	0.867 ± 0.001
NEUROBIBER	96	0.795 ± 0.001	0.878 ± 0.001
LUAR	512	0.798 ± 0.002	0.883 ± 0.000
BIBERPLUS	192	0.799 ± 0.001	0.879 ± 0.001
NEUROBIBER + LUAR	608	0.835 ± 0.001	0.916 ± 0.000
BIBERPLUS + LUAR	704	0.839 ± 0.002	0.918 ± 0.001
RoBERTa bi-encoder + LUAR	1280	0.804 ± 0.002	0.889 ± 0.000
NEUROBIBER + StyleDistance	864	0.775 ± 0.001	0.855 ± 0.001

Table 5: **Authorship-verification details.** Embedding dimensionality and mean ± standard deviation over the five paired seeds for every system in Table 3.