References

1. Abbasi, A., Javed, A. R., Iqbal, F., Jalil, Z., Gadekallu, T. R., & Kryvinska, N. (2022). Authorship identification using ensemble learning. *Scientific Reports*, 12, Article 9537. https://doi.org/10.1038/s41598-022-13690-4
2. Abburi, H., Suesserman, M., Pudota, N., Veeramani, B., Bowen, E., & Bhattacharya, S. (2023). An ensemble-based approach for generative language model attribution. *arXiv preprint* arXiv:2309.07755v1 [cs.CL]. https://doi.org/10.1007/978-981-99-7254-8_54
3. Alfaro, E., Gamez, M., & Garcia, N. (2013). adabag: An R package for classification with boosting and bagging. *Journal of Statistical Software*, 54(2), 1–35. https://doi.org/10.18637/jss.v054.i02
4. Antosch, F. (1969). The diagnosis of literary style with the verb-adjective ratio. In L. Doleszel & R. W. Bailey (Eds.), *Statistics and Style* (pp. 1–20). New York: American Elsevier.
5. Aoyama, H., & Contable, J. (1999). Word length frequency and distribution in English: Part I Prose. *Literary and Linguistic Computing*, 14(3), 339–359.
6. Arslan, Y., Allix, K., Veiber, L., Lothritz, C., Bissyandé, T. F., Klein, J., et al. (2021). A comparison of pre-trained language models for multi-class text classification in the financial domain. *Companion Proceedings of the Web Conference*, 260–268. https://doi.org/10.1145/3442442.3451375
7. Ashraf, S., Iqbal, H. R., & Nawab, R. M. A. (2016). Cross-genre author profile prediction using stylometry-based approach. In *Working Notes Papers of the CLEF 2016 Evaluation Labs* (pp. 992–999).
8. Baayen, R. H. (2013). *Word Frequency Distributions*. Springer.
9. Bacciu, A., Morgia, M. L., Mei, A., Nemmi, E. N., Nerib, V., & Stefa, J. (2019). Cross-domain authorship attribution combining instance-based and profile-based features notebook for PAN at CLEF 2019. *CEUR Workshop Proceedings*, 2380, 1–12. https://ceur-ws.org/Vol-2380/paper_220.pdf
10. Backera, E., & Kranenburgb, P. V. (2005). On musical stylometry: A pattern recognition approach. *Pattern Recognition Letters*, 26(3), 299–309.
11. Becker, C. (1996). Word lengths in the letters of the Chilean author Gabriela Mistral. *Journal of Quantitative Linguistics*, 3(2), 128–131.
12. Best, K. H. (1996). Word length in Old Icelandic songs and prose texts. *Journal of Quantitative Linguistics*, 3(2), 97–105.
13. Bevendorff, J., Ghanem, B., Giachanou, A., Kestemont, M., Manjavacas, E., Potthast, M., et al. (2020). Shared tasks on authorship analysis at PAN 2020. *Advances in Information Retrieval*, 508–516. Springer.
14. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
15. Brinegar, C. S. (1963). Mark Twain and the Quintus Curtius Snodgrass letters: A statistical test of authorship. *Journal of the American Statistical Association*, 58(301), 85–96.
16. Brinkman, A., Shanahan, D., & Sapp, C. (2016). Musical stylometry, machine learning, and attribution studies: A semi-supervised approach to the works of Josquin. In *Proceedings of the 14th Biennial International Conference on Music Perception and Cognition* (pp. 91–97).

References

17. Burrows, J. F. (1987). *Computation Into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.
18. Cammarota, V., Bozza, S., Roten, C. A., & Taroni, F. (2024). Stylometry and forensic science: A literature review. *Forensic Science International: Synergy*, 9, 100481. https://doi.org/10.1016/j.fsisyn.2024.100481
19. Constable, J., & Aoyama, H. (1999). Word length frequency and distribution in English: Part II. An empirical and mathematical examination of the character and consequences of isometric lineation. *Literary and Linguistic Computing*, 14(4), 507–535.
20. Cox, D. R., & Brandwood, L. (1959). On a discriminatory problem connected with the works of Plato. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(1), 195–200.
21. Dang, H., Lee, K., Henry, S., & Uzuner, Ö. (2020). Ensemble BERT for classifying medication-mentioning tweets. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task* (pp. 37–41). Barcelona, Spain: Association for Computational Linguistics. https://aclanthology.org/2020.smm4h-1.5
22. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805 [cs.CL].
23. Eder, M. (2015). Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, 30(2), 167–182. https://doi.org/10.1093/llc/fqt066
24. Efron, B., & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3), 435–447.
25. Ellegård, A. (1962a). *A Statistical Method for Determining Authorship: The Junius Letters, 1769–1772*. Gothenburg Studies in English, 13. Acta Universitatis Gothoburgensis.
26. Ellegård, A. (1962b). *Who was Junius?* Stockholm: Almqvist & Wiksell.
27. Fabien, M., Villatoro-Tello, E., Motlicek, P., & Parida, S. (2020). BertAA: BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)* (pp. 127–137).
28. Frischen, J. (1996). Word length analysis of Jane Austen's letters. *Journal of Quantitative Linguistics*, 3(1), 128–131.
29. Fucks, W. (1952). On mathematical analysis of style. *Biometrika*, 39(1–2), 122–129.
30. Fucks, W. (1954). On Nahordnung and Fernordnung in samples of literary texts. *Biometrika*, 41(1–2), 116–132.
31. George Mikros, Athanasios Koursaris, Dimitrios Bilianos, & George Markopoulos. (2023). AI-writing detection using an ensemble of transformers and stylometric features. *CEUR Workshop Proceedings* (CEUR-WS.org), September 2023, Jaén, Spain. http://ceur-ws.org
32. Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3), 251–270.
33. Halvani, O., Winter, C., & Graner, L. (2019). Assessing the applicability of authorship verification methods. In *The 14th International Conference on Availability, Reliability and Security (ARES 2019)*. arXiv:1901.00399

References

34. Hatano, K. (1950). *Psychology of Writing*. Tokyo: Shinchosha. (Japanese)
35. He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint* arXiv:2006.03654 [cs.CL].
36. Herdan, G. (1958). The relation between the dictionary distribution and the occurrence distribution of word length and its importance for the study of quantitative linguistics. *Biometrika*, 45(1–2), 222–228.
37. Holmes, D. I. (1994). Authorship attribution. *Computers and the Humanities*, 28(2), 87–106.
38. Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3), 111–117.
39. Holmes, D. I., & Forsyth, R. S. (1995). The Federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing*, 10(2), 112–127.
40. Hoorn, J. F., Frank, S. L., Kowalczyk, W., & Ham, F. (1999). Neural network identification of poets using letter sequences. *Literary and Linguistic Computing*, 14(3), 311–338. https://doi.org/10.1093/llc/14.3.311
41. Hughes, J. M., Graham, D. J., & Rockmore, D. N. (2010). Stylometrics of artwork: Uses and limitations. *Proceedings of SPIE 7531, Computer Vision and Image Analysis of Art*. https://doi.org/10.1117/12.838849
42. Jin, M. (2013). Authorship identification based on phrase pattern. *Behaviormetrika*, 40(1), 17–28. (Japanese)
43. Jin, M. (2014). Using integrated classification algorithm to identify a text's author. *Behaviormetrika*, 41(1), 35–46. (Japanese)
44. Jin, M. (2021). Text analytics fundamentals and practice. *Text Analytics Series*, 1, 223–227. Tokyo: Iwanami Shoten. (Japanese)
45. Jin, M. (1994a). Quantitative studies on patterns in natural language. *Sogo University, Doctoral Dissertation*. (Japanese)
46. Jin, M. (1994b). The use of commas and the stylistic characteristics of authors. *Mathematical Linguistics*, 19(7), 317–330. (Japanese)
47. Jin, M. (1995). Classification of texts based on the distribution of verb length and the ratio of native and compound words. *Natural Language Processing*, 2(1), 57–75. (Japanese)
48. Jin, M. (1996). The distribution of verb length and the writer of texts. *Social Information*, 5(2), 13–22. (Japanese)
49. Jin, M. (1997). Recognition of diary writers based on the distribution of particles. *Mathematical Linguistics*, 20(8), 357–367. (Japanese)
50. Jin, M. (2000). Information processing in natural language using statistical methods. *Statistical Mathematics*, 48(2), 271–287. (Japanese)
51. Jin, M. (2002). Writer identification based on the n-gram model of particles. *Mathematical Linguistics*, 23(5), 225–240. (Japanese)
52. Jin, M. (2002a). Quantitative analysis of the characteristics of writers in the distribution of particles. *Social Information*, 11(2), 15–23. (Japanese)
53. Jin, M. (2002b). Writer identification based on the n-gram model of particles. *Mathematical*

Linguistics, 23(5), 225–240. (Japanese)

54. Jin, M. (2003a). Writer identification and characteristic analysis using self-organizing maps and particle distribution. *Mathematical Linguistics*, 23(8), 369–386. (Japanese)
55. Jin, M. (2003b). Writer identification in Chinese texts. *Second International Conference on Chinese Sociolinguistics and the Founding Conference of the Chinese Sociolinguistics Association*, Macau. (Japanese)
56. Jin, M. (2004). Writer identification using Markov transition information of parts of speech. *32nd Annual Meeting of the Japanese Society for Behavioral Metrics*. (Japanese)
57. Jin, M. (2009a). Estimation of the writing period of texts: A case study of Ryunosuke Akutagawa's works. *Behavioral Metrics*, 36(2), 89–103. (Japanese)
58. Jin, M. (2013). Writer identification based on phrase patterns. *Behavioral Metrics*, 40(1), 17–28. (Japanese)
59. Jin, M. (2014). Writer identification using integrated classification algorithms. *Behavioral Metrics*, 41(1), 35–46. (Japanese)
60. Jin, M. (2021). *Fundamentals and Practice of Text Analytics*. Tokyo: Iwanami Shoten. (Japanese)
61. Jin, M. , & Murakami, M. (2007). Writer identification using random forests. *Mathematical Statistics*, 55(2), 255–268. (Japanese)
62. Jin, M. , Kabashima, T., & Murakami, M. (1993). Commas and the individuality of writers. *Mathematical Linguistics*, 18(8), 382–391. (Japanese)
63. Jin, M. , Kabashima, T., & Murakami, M. (1994). Quantitative analysis of handwritten and word-processed texts. *Mathematical Linguistics*, 19(3), 133–145. (Japanese)
64. Jin, M., & Huh, M.-H. (2012). Author identification of Korean texts by minimum distance and machine learning. *Survey Research*, 13(3), 175–190.
65. Jin, M., & Jiang, M. (2013). Text clustering on authorship attribution based on the features of punctuation usage. *Information (An International Interdisciplinary Journal)*, 16(7B), 4983–4990.
66. Jin, M., & Murakami, M. (1993). Authors' characteristic writing styles as seen through their use of commas. *Behaviormetrika*, 20, 63–76.
67. Jin, M., & Murakami, M. (2017). Authorship identification using random forests. *Proceedings of the Institute of Statistical Mathematics*, 55(2), 255–268. (Japanese)
68. Jin, M., & Zheng, W. (2020). Text corpus mining software MTMineR. *Mathematical Linguistics*, 32(5), 265–276. (Japanese)
69. Kabashima, T. (1955). Regularities in the ratios of classified parts of speech. *Kokugakuin University Journal*, 24(6), 385–387. (Japanese)
70. Kabashima, T. (1963). *Expression Theory: Words and Linguistic Behavior*. Tokyo: Sougeisha. (Japanese)
71. Kabashima, T. (1990). *Japanese Stylebook*. Tokyo: Taishukan Shoten. (Japanese)
72. Kabashima, T., & Jukaku, S. (1965). *The Science of Style*. Tokyo: Sougeisha. (Japanese)
73. Kalgutkar, V., Kaur, R., Gonzalez, H., Stakhanova, N., & Matyukhina, A. (2019). Code

authorship attribution: Methods and challenges. *ACM Computing Surveys*, 52(1), 1–36. https://doi.org/10.1145/3292577

74. Kanda, T., & Jin, M. (2024). An empirical comparison and ensemble learning methods of BERT models on authorship attribution. *Journal of Japan Society of Information and Knowledge*, 34(3), 244–255. (Japanese)
75. Karl, F., & Scherp, A. (2022). Transformers are short text classifiers: A study of inductive short text classifiers on benchmarks and real-world datasets. *arXiv preprint* arXiv:2211.16878v3 [cs.CL].
76. Kjell, B. (1994). Authorship determination using letter pair frequency features with neural network classifiers. *Literary and Linguistic Computing*, 9(2), 119–124. https://doi.org/10.1093/llc/9.2.119
77. Kokensparger, B. (2018). Art stylometry: Recognizing regional differences in great works of art. In *Guide to Programming for the Digital Humanities* (pp. 69–78). Springer.
78. Koppel, M., & Schler, J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis* (pp. 69–72).
79. Kuwano, M., & Jin, M. (2008). Stylistic changes in Mori Ogai before and after his transfer to Kokura. *52nd Annual Meeting of the Japanese Society for Mathematical Linguistics*. Mukogawa Women's University. (Japanese)
80. Lagutina, K., Lagutina, N., Boychuk, E., & Vorontsova, I. (2019). A survey on stylometric text features. In *25th Conference of Open Innovations Association (FRUCT)*. https://doi.org/10.23919/FRUCT48121.2019.8981504
81. Lasotte, Y. B., Garba, E. J., Malgwi, Y. M., & Buhari, M. A. (2022). An ensemble machine learning approach for fake news detection and classification using a soft voting classifier. *European Journal of Electrical Engineering and Computer Science*, 6(2), 1–7. https://doi.org/10.24018/ejece.2021.6.2.409
82. Lee, J., Choi, J., & Jin, M. (2016). Writer identification of Korean texts using phrase patterns. *Information*, 20(1B), 417–428. (Japanese)
83. Lelyveld, J. (1985). A scholar’s find: Shakespearean lyric. *New York Times* (November 24, 1985), 1, 12. With correction of ‘Editor’s Note’ (November 25, 1985), 2.
84. Li, G., & Jin, M. (2019). Stylistic imitation in the novel *Zoku Meian* from a statistical analysis perspective. *Mathematical Linguistics*, 32(1), 19–32. (Japanese)
85. Li, X. (1987). “Hong Lou Meng” ChengShu XinShuo. *Journal of Fudan University (Social Sciences)*, 5, 3–16. (Chinese)
86. Ling, Y. (2023). Bio+Clinical BERT, BERT Base, and CNN performance comparison for predicting drug-review satisfaction. *KDD 2023 Workshop on Applied Data Science for Healthcare*. arXiv:2308.03782v1 [cs.CL]
87. Liu, H., Chan, R. H., & Yao, Y. (2016). Geometric tight frame based stylometry for art authentication of van Gogh paintings. *Applied and Computational Harmonic Analysis*, 41(2), 590–602.

References

88. Liu, X., & Jin, M. (2017a). Quantitative analysis of stylistic changes in Uno Koji before and after his illness. *Mathematical Linguistics*, 31(2), 128–143. (Japanese)
89. Liu, X., & Jin, M. (2017b). Quantitative analysis of stylistic changes in Uno Koji before and after his illness. *Mathematical Linguistics*, 31(2), 128–143. (Japanese)
90. Liu, Y., & Jin, M. (2022). Authorship attribution in the multi-genre mingled corpus. *Bulletin of Data Analysis of Japanese Classification Society*, 11(1), 1–14. (Japanese)
91. Lutosławski, W. (1898). Principes de stylométrie appliqués à la chronologie des œuvres de Platon. *Revue des Études Grecques*, 11(41), 61–81.
92. Manuel, F., Eva, C., Senén, B., & Dinani, A. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15, 3133–3181.
93. Matsumura, T., & Kaneda, Y. (2000). Writer identification in modern Japanese sentences using N-gram distribution. *Mathematical Linguistics*, 22(6), 225–238. (Japanese)
94. Matsuura, T., & Kanada, Y. (2001). Extraction of authors' characteristics from Japanese modern sentences via N-gram distribution. *Lecture Notes in Artificial Intelligence*, 1967, 315–319. Springer. https://doi.org/10.1007/3-540-44418-1_38
95. Mekala, S., Bulusu, V., & Reddy, R. (2018). A survey on authorship attribution approaches. *International Journal of Computational Engineering Research (IJCER)*, 8(9), 48–55.
96. Melka, T. S., & Místecký, M. (2019). On stylometric features of H. Beam Piper's Omnilingual. *Journal of Quantitative Linguistics*, 27(3), 1–40.
97. Mendenhall, T. C. (1887). The characteristic curves of composition. *Science*, 9(214), 237–249.
98. Mendenhall, T. C. (1901). A mechanical solution of a literary problem. *Popular Science Monthly*, 60, 97–105.
99. Merriam, T., & Matthews, R. (1993). Neural computation in stylometry I: An application to the works of Shakespeare and Fletcher. *Literary and Linguistic Computing*, 8(4), 203–209.
100. Merriam, T., & Matthews, R. (1994). Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe. *Literary and Linguistic Computing*, 9(1), 1–6.
101. Meyer, P. (1997). Word-length distribution in Inuktitut narratives: Empirical and theoretical findings. *Journal of Quantitative Linguistics*, 4(1–3), 143–155.
102. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint* arXiv:1310.4546 [cs.CL].
103. Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. (2020). Evaluation of sentiment analysis in finance: From lexicons to transformers. *IEEE Access*, 8, 131662–131682. https://doi.org/10.1109/ACCESS.2020.3009626
104. Morton, A. Q. (1965). The authorship of Greek prose. *Journal of the Royal Statistical Society: Series A (General)*, 128(2), 169–233.
105. Mosteller, F., & Wallace, D. L. (1963). Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302), 275–309.
106. Mosteller, F., & Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley Publishing Company.

References

107. Murakami, M., & Imanishi, Y. (1999). Quantitative analysis of auxiliary verbs in *The Tale of Genji*. *Journal of Information Processing Society of Japan*, 40(3), 774–782. (Japanese)
108. Murakami, M., & Ito, Z. (1991). Mathematical research on Nichiren's writings. *Toyo no Shiso to Shukyo*, 8, 27–35. (Japanese)
109. Murata, T. (2007). Research on conjunctions and particle-equivalent phrases for teaching argumentative writing in specialized Japanese education. *Statistical Mathematics*, 55(2), 269–284. (Japanese)
110. Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., & Woodard, D. (2017). Surveying stylometry techniques and applications. *ACM Computing Surveys (CSUR)*, 50(6), 1–36. https://doi.org/10.1145/3132039
111. Nirasawa, T. (1965). Statistical discrimination of the author of *Yura Monogatari*. *Mathematical Linguistics*, 33, 21–28. (Japanese)
112. Okuda, Y. (1998). Results and considerations on "Commas and the individuality of writers". *Nagoya University Department of Mathematical Sciences, Oobata Laboratory, Graduation Thesis Collection*. (Japanese)
113. Otoom, A., Abdallah, E., Hammad, M., & Bosoul, M. (2014). An intelligent system for author attribution based on a hybrid feature set. *International Journal of Advanced Intelligence Paradigms*, 6(4), 328–345.
114. Palme, H. (1949). Versuch einer statistischen Auswertung des alltäglichen Schreibstils.
115. Prytula, M. (2024). Fine-tuning BERT, DistilBERT, XLM-RoBERTa and Ukr-RoBERTa models for sentiment analysis of Ukrainian language reviews. *Artificial Intelligence*, 285, 85–97. https://doi.org/10.15407/jai2024.02.085
116. Qasim, R., Bangyal, W. H., Alqarni, M. A., & Almazroi, A. A. (2022). A fine-tuned BERT-based transfer learning approach for text classification. *Journal of Healthcare Engineering*, 2022, Article ID 3498123, 1–17. https://doi.org/10.1155/2022/3498123
117. Quiring, E., Maier, A., & Rieck, K. (2019). Misleading authorship attribution of source code using adversarial learning. *USENIX Security Symposium 2019*. https://arxiv.org/abs/1905.1238
118. Riedemann, H. (1996). Word-length distribution in English press texts. *Journal of Quantitative Linguistics*, 3(3), 265–271.
119. Rottmann, O. (1997). Word-length counting in Old Church Slavonic. *Journal of Quantitative Linguistics*, 4(1–3), 252–256.
120. Sara, E., Manar, E., & Kassou, I. (2014). Authorship analysis studies: A survey. *International Journal of Computer Applications*, 86(12), 23–29.
121. Sasaki, K. (1976). The distribution of sentence length. *Mathematical Linguistics*, 78, 13–22. (Japanese)
122. Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1–47.
123. Sherman, L. A. (1888). Some observations upon the sentence-length in English prose. *University Studies (University of Nebraska (Lincoln campus))*, 1(2), 119–130.
124. Sichel, H. S. (1974). On a distribution representing sentence-length in written prose. *Journal of*

References

the Royal Statistical Society: Series A (General), 137(1), 25–34.

125. Smith, M. W. A. (1983). Recent experience and new developments of methods for the determination of authorship. *Association for Literary and Linguistic Computing Bulletin*, 11, 73–82.
126. Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538–556. https://doi.org/10.1002/asi.21001
127. Strøm, E. (2021). Multi-label style change detection by solving a binary classification problem. In *CLEF (Working Notes)* (pp. 2146–2157). https://ceur-ws.org/Vol-2936/paper-191.pdf
128. Sun, H., & Jin, M. (2018). Verification of the ghostwriting suspicion of Kawabata Yasunari's novel *Hanadokei*. *Journal of Information and Knowledge*, 28(1), 3–14. (Japanese)
129. Sun, X., Li, X., Li, W., Wu, F., Guo, S., Zhang, T., et al. (2023). Text classification via large language models. *Findings of the Association for Computational Linguistics: EMNLP*, 8990–9005. arXiv:2305.08377v3 [cs.CL]
130. Suzuki, M., Sakaji, H., Hirano, M., & Izumi, K. (2021). Performance validation of pre-trained BERT in the financial domain. *IEICE Technical Report*, 121/178, 26–29. (Japanese)
131. Tanaka, H., Rui, C., Jing, B., Wen, M., & Hiroyuki, S. (2019). Construction of document feature vectors using BERT. *SIG Technical Reports*, 2019-NL-243/8, 1–6.
132. Tanaka, R., & Jin, M. (2014). Authorship attribution of cell-phone e-mail. *Information*, 17(4), 1217–1226.
133. Tanaka, R., & Jin, M. (2010). An attempt to identify the writer of mobile phone emails. *2010 Joint Statistical Meetings Proceedings*, 332. (Japanese)
134. Taylor, G. (1985). Shakespeare’s new poem: A scholar’s clues and conclusion. *New York Times Book Review* (December 15), 11–14.
135. Tearle, M., Taylor, K., & Demuth, H. (2008). An algorithm for automated authorship attribution using neural networks. *Literary and Linguistic Computing*, 23(4), 425–442.
136. Thisted, R., & Efron, B. (1987). Did Shakespeare write a newly-discovered poem? *Biometrika*, 74(3), 445–455.
137. Tsai, T. J., & Ji, K. (2020). Composer style classification of piano sheet music images using language model pretraining. *International Society for Music Information Retrieval Conference (ISMIR) 2020*. https://arxiv.org/pdf/2007.14587.pdf
138. Tweedie, F. J., Singh, S., & Holmes, D. I. (1995). An introduction to neural networks in stylometry. *Research in Humanities Computing*, 5, 249–263.
139. Tweedie, F. J., Singh, S., & Holmes, D. I. (1996). Neural network application in stylometry: The Federalist papers. *Computers and the Humanities*, 30(1), 1–10.
140. Tyo, J., Dhingra, B., & Lipton, Z. C. (2022). On the state of the art in authorship attribution and authorship verification. *arXiv preprint* arXiv:2209.06869v2 [cs.CL].
141. Valenza, R. J. (1991). Are the Thisted-Efron authorship tests valid? *Computers and the Humanities*, 25(1), 27–46.
142. Vanetik, N., Tiamanova, M., Koga, G., & Litvak, M. (2024). Genre classification of books in

Russian with stylometric features: A case study. *Information*, 15(6), 340. https://doi.org/10.3390/info15060340

143. Wake, W. C. (1957). Sentence-length distributions of Greek authors. *Journal of the Royal Statistical Society: Series A (General)*, 120(3), 331–346.
144. Waugh, S., Adams, A., & Tweedie, F. (2000). Computational stylometrics using artificial neural networks. *Literary and Linguistic Computing*, 15(2), 187–197.
145. Wiener, E., Pedersen, J. O., & Weigend, A. S. (1995). A neural network approach to topic spotting. In *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95)*.
146. Williams, C. B. (1975). Mendenhall's studies of word-length distribution in the works of Shakespeare and Becon. *Biometrika*, 62(1), 207–211.
147. Wu, Z., Liang, J., Zhang, Z., & Lei, J. (2021). Exploration of text matching methods in Chinese disease Q&A systems: A method using ensemble based on BERT and boosted tree models. *Journal of Biomedical Informatics*, 115, 103683. https://doi.org/10.1016/j.jbi.2021.103683
148. Xie, H., Arash, H. L., Nikhill, V., & Dilli, P. S. (2024). Authorship attribution methods, challenges, and future research directions: A comprehensive survey. *Information*, 15(3), 131. https://doi.org/10.3390/info15030131
149. Xu, C., Barth, S., & Solis, Z. (2020). Applying ensembling methods to BERT to boost model performance. *Stanford University*. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/default/15775971.pdf
150. Yanagi, Y., & Jin, M. (2022). Author identification analysis in the case of mixed-genre texts. *Theory and Application of Data Analysis*, 11(1), 1–14. (Japanese)
151. Yasumoto, M. (1958). Stylometric authorship attribution: A case study of the authorship of the Uji chapters of *The Tale of Genji*. *Psychological Review*, 2(1), 147–156. (Japanese)
152. Yasumoto, M. (1958a). On the distribution of sentence length. *Mathematical Linguistics*, 2, 20–25. (Japanese)
153. Yasumoto, M. (1958b). Stylometric authorship attribution: A case study of the authorship of the Uji chapters of *The Tale of Genji*. *Psychological Review*, 2, 147–156. (Japanese)
154. Yasumoto, M. (1959). Basic research on the characterology of writing: Classification of modern writers by factor analysis. *Kokugakuin University Journal*, 6, 19–41. (Japanese)
155. Yasumoto, M. (1994). Three factors determining style. *Language*, 23(2), 22–29. (Japanese)
156. Yasumoto, M. (2009). Stylometry and psychology of writing. In *Encyclopedia of Mathematical Linguistics* (pp. 253–263). Tokyo: Asakura Publishing. (Japanese)
157. Yasumoto, M., & Honda, M. (1988). *Factor Analysis*. Tokyo: Baifukan. (Japanese)
158. Yoshioka, R. (1999). Quantitative analysis of new books. *Humanities and Information Processing*, 20, 51–56. (Japanese)
159. Yukimura, R., Sun, H., & Jin, M. (2018). Feature analysis of paintings using color information of the image. In *Proceedings of Digital Humanities Austria (DHA) 2018* (pp. 55–

61). https://epub.oeaw.ac.at/0xc1aa5576_0x003b398d.pdf

160. Yule, G. U. (1939). On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, 30(3–4), 363–390. https://doi.org/10.1093/biomet/30.3-4.363
161. Yule, G. U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press.
162. Zaitsu, W. (2016). Classification of motives in arson cases over the past 10 years using text mining: A comparison of single and serial arson. *Japanese Journal of Criminal Psychology*, 53(2), 1–13. (Japanese)
163. Zaitsu, W. (2019). Text mining for criminal investigation: Exploring the fingerprints of texts, challenging cybercrime with quantitative stylistic analysis. *Kyoritsu Shuppan*. (Japanese)
164. Zaitsu, W., & Jin, M. (2015). Writer identification of crime-related documents using text mining. *Japanese Journal of Forensic Science and Technology*, 20(1), 1–14. (Japanese)
165. Zaitsu, W., & Jin, M. (2017). Gender estimation of authors using random forests: Toward the realization of criminal profiling. *Journal of Information and Knowledge*, 27(3), 261–274. (Japanese)
166. Zaitsu, W., & Jin, M. (2018). Age group estimation of authors using machine learning: Toward the realization of criminal profiling. *Doshisha University Harris Institute for Science and Technology Reports*, 59(2), 57–65. (Japanese)
167. Zaitsu, W., & Jin, M. (2018). Writer identification based on the usage rate of sentence-ending words: Exploratory multivariate analysis and scoring of the results. *Mathematical Linguistics*, 31(6), 417–425. (Japanese)
168. Zaitsu, W., Jin, M. (2023). Distinguishing ChatGPT(-3.5, -4)-generated and human-written papers through Japanese stylometric analysis. *PLOS ONE*, 18(8), e0288453. https://doi.org/10.1371/journal.pone.0288453
169. Zaitsu, W., Jin, M., Ishihara, S., Tsuge, S., & Inaba, M. (2024). Can we spot fake public comments generated by ChatGPT(-3.5, -4)?: Japanese stylometric analysis expose emulation created by one-shot learning. *PLOS ONE*, 19(3), e0299031. https://doi.org/10.1371/journal.pone.0299031
170. Zamir, M. T., Ayub, M. A., Gul, A., Ahmad, N., & Ahmad, K. (2024). Stylometry analysis of multi-authored documents for authorship and author style change detection. *arXiv preprint* arXiv:2401.06752v1 [cs.CL].
171. Zhang, Y., & Jiang, M. (2021). Authorship identification of text based on attention mechanism. *Journal of Computer Applications*, 41(7), 1897–1901.
172. Zhang, Y., et al. (2024). Pushing the limit of LLM capacity for text classification. *arXiv preprint* arXiv:2402.07470 [cs.CL].
173. Zheng, W., & Jin, M. (2022). A review on authorship attribution in text mining. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15, e1584. https://doi.org/10.1002/wics.1584
174. Zheng, W., & Jin, M. (2023). Is word-length inaccurate for authorship attribution? *Digital Scholarship in the Humanities*, 38(2), 875–890.

References

175. Ziegler, A. (1996). Word length distribution in Brazilian-Portuguese texts. *Journal of Quantitative Linguistics*, 3(1), 73–79.
176. Zuse, M. (1996). Distribution of word length in early modern English letters of Sir Philip Sidney. *Journal of Quantitative Linguistics*, 3(3), 272–276.

日本語文献（上記の英語リストと重複している）:

1. 安本 美典(1958). 文体統計による筆者推定—源氏物語、宇治十帖の作者について—、心理学評論, 2 巻 1 号 p. 147-156
2. 安本 美典(1958a). 文の長さの分布型について，計量国語学，2，20-25.
3. 安本 美典(1958b). 文体統計による筆者推定—源氏物語，宇治十帖の著者について—，心理学評論，2， 147-156.
4. 安本 美典(1959). 文章の性格学への基礎研究—因子分析方による現代作家の分類—国語国文，6， 19-41.
5. 安本 美典(1994). 文体を決める三つの因子，言語，23(2)， 22-29.
6. 安本 美典(2009). 計量文体論・文章心理学,『計量国語学事典』(朝倉出版)，253-263
7. 安本 美典・本多 正久(1988). 因子分析法，培風館.
8. 奥田 康誠(1998). 「読点と書き手の個性」における結果とその考察，名古屋大学理学部数理科学科尾畑伸明研究室，卒業論文集.
9. 樺島 忠夫(1955). 分類した品詞の比率に見られる規則性，国語国文, 24-6, 385-387.
10. 樺島 忠夫(1963). 表現論—言葉と言語行動，綜芸舎.
11. 樺島 忠夫(1990). 日本語のスタイルブック，大修館書店.
12. 樺島 忠夫・寿岳 章子(1965).文体の科学，綜芸舎.
13. 吉岡 亮衛(1999). 新書の数量的分析，人文学と情報処理，20， 51-56.
14. 金 明哲(1994a). 自然言語におけるパターンに関する計量的研究，総合大学院大学，学位論文.
15. 金 明哲(1994b). 読点の打ち方と著者の文体特徴，計量国語学，19(7)， 317-330.
16. 金 明哲(1995). 動詞の長さの分布に基づいた文章の分類と和語および合成語の比率，自然言語処理，2(1)，57-75.
17. 金 明哲(1996). 動詞の長さの分布と文章の書き手，社会情報， 5(2)， 13-22.
18. 金 明哲(1997). 助詞の分布に基づいた日記の書き手の認識，計量国語学，20(8)， 357-367.
19. 金 明哲(2000). 自然言語における統計手法を用いた情報処理，統計数理, 48(2), 271–287.
20. 金 明哲(2002). 助詞の n-gram モデルに基づいた書き手の識別, 計量国語学, 23(5), 225-240.
21. 金 明哲(2002a). 助詞の分布における書き手の特徴に関する計量分析，社会情報，11(2), 15-23.
22. 金 明哲(2002b). 助詞の n-gram モデルに基づいた書き手の識別，計量国語学，23(5), 225-240.
23. 金 明哲(2003). 自己組織化マップと助詞分布を用いた書き手の同定及びその特徴分析,

計量国語学，23(8), 369-386.

24. 金　明哲(2003a). 自己組織化マップと助詞分布を用いた書き手の同定及びその特徴分析，計量国語学，23(8), 369-386.
25. 金　明哲(2003b). 中国文章における書き手の識別,第二届中国社会語言学国際学術検討会暨中国社会語言学会成立大会,マカオ.
26. 金　明哲(2004). 品詞のマルコフ遷移の情報を用いた書き手の同定, 日本行動計量学会第 32 回大会
27. 金　明哲(2009a).文章の執筆時期の推定 ：芥川龍之介の作品を例として,行動計量学, 36(2), 89-103.
28. 金　明哲(2013). 文節パターンに基づいた文章の書き手の識別, 行動計量学, 40(1), 17-28.
29. 金　明哲(2014). 統合的分類アルゴリズムを用いた文章の書き手の識別，行動計量学, 41(1), 35-46.
30. 金　明哲, 村上　征勝(2007). ランダムフォレスト法による文章の書き手の同定, 数理統計，55(2), 255-268.
31. 金　明哲,樺島　忠夫,村上　征勝(1994).手書きとワープロとによる文章の計量分析,計量国語学,19(3),133-145.
32. 金　明哲・樺島　忠夫・村上　征勝(1993). 読点と書き手の個性,計量国語学,18(8), 382-391.
33. 金　明哲(2021). テキストアナリティクスの基礎と実践, 岩波書店.
34. 桑野　麻友子, 金明哲(2008). 小倉左遷前後における森鴎外の文体変化" 日本計量国語学会第 52 回大会. 武庫川女子大学.
35. 佐々木　和枝(1976). 文の長さの分布，計量国語学，78,　13-22.
36. 財津　亘 (2019): 犯罪捜査のためのテキストマイニング: 文章の指紋を探り,サイバー犯罪に挑む計量的文体分析の手法, 共立出版.
37. 財津　亘, 金　明哲 (2017). ランダムフォレストによる著者の性別推定－犯罪者プロファイリング実現に向けた検討－. 情報知識学会誌, 27(3), 261-274.
38. 財津　亘, 金　明哲 (2018). 機械学習を用いた著者の年齢層推定—犯罪者プロファイリング実現に向けて—. 同志社大学ハリス理化学研究報告, 59(2), 57-65.
39. 財津　亘, 金　明哲 (2018). 文末語の使用率に基づいた筆者識別—探索的多変量解析の実施と分析結果に対すスコアリングによる検討—. 計量国語学，31(6)，417-425.
40. 財津　亘(2016). テキストマイニングによる最近 10 年間の放火事件に関する動機の分類-単一放火と連続放火の比較-,　犯罪心理学研究, 53(2), 1-13.
41. 財津　亘, 金明哲（2015）.テキストマイニングを用いた犯罪に関わる文書の筆者識別, 日本法科学技術学会誌, 20(1), 1-14.
42. 松村　司，金田　康正(2000).　n-gram の特徴量を利用した近代日本文の著者識別，計量国語学，22(6)，225-238.
43. 孫　昊, 金　明哲 (2018). 川端康成小説『花日記』の代筆疑惑検証. 情報知識学会誌, 28(1), 3-14
44. 村上　征勝，今西　祐一郎(1999). 源氏物語の助動詞の計量分析,情報処理学会論文誌,40(3),774-782.
45. 村上　征勝，伊藤　瑞叡(1991). 日蓮遺文の数理研究，東洋の思想と宗教，8,　27-35.

References

46. 村田　年(2007). 専門日本語教育における論述文指導のための接続語句・助詞相当句の研究, 統計数理, 55(2), 269–284.
47. 田中 量子, 金 明哲(2010). 携帯電話メールの書き手の判別に関する試み, 2010 年統計関連学会連合大会報告集，332
48. 韮沢 正(1965). 由良物語の著者の統計的判別，計量国語学, 33, 21-28.
49. 波多野 完治(1950)．文章心理学，新潮社.
50. 柳 燁佳, 金 明哲(2022). 異ジャンル文章が混在した場合における著者識別分析.データ分析の理論と応用,11(1), 1–14.
51. 李 広微, 金 明哲 (2019). 統計解析からみた小説『続明暗』の文体模倣. 計量国語学, 32(1), 19-32.
52. 李 鍾賛, 崔 在雄, 金 明哲 (2016). 語節パターンを用いた韓国語文章の著者識別. Information. 20(1B), 417-428.
53. 劉 雪琴, 金 明哲（2017a). 宇野浩二の病気前後の文体変化に関する計量的分析．計量国語学, 31(2), 128-143.
54. 劉 雪琴，金 明哲(2017b). 宇野浩二の病気前後の文体変化に関する計量的分析, 計量国語学, 31(2), 128-143.