Enhancing Information Retrieval in the Literary Domain through Authorship Attribution

Kanishka Silva[1,*]

[1]University of Wolverhampton, Wolverhampton, United Kingdom

Abstract

Information retrieval involves accessing data resources to search for specific information based on user intent or a search query. Enhancing the quality of information retrieval is important, and various cross-domain applications have been proposed to achieve this goal. Authorship Attribution applications have benefitted from the information retrieval techniques, and similarly, information retrieval has also been influenced by the areas such as genre analysis, bot profiling and author identification. These interdependencies have opened up various new research directions in both domains. This paper presents ongoing research that proposes a methodology to improve relevancy, trustworthiness and personalisation factors in information retrieval using authorship attribution techniques of genre analysis, bot profiling and author base ranking, illustrating a case study from the literary domain.

Keywords

Author-based Ranking, Bot profiling, Genre analysis, Information Retrieval, Authorship Attribution

1. Introduction

Information retrieval involves retrieving relevant information from a data collection based on user queries. In recent years, information retrieval has expanded across different domains, and research directions have significantly increased due to global connectivity and accessibility to large repositories easily. Relevancy, trustworthiness and personalisation are essential for enhancing the query response. While various approaches have been developed to measure the relevancy of search results [1], there is still room for improvement, particularly when retrieving information based on the authors. In a personalised setting [2, 3, 4], such as literary work-based querying, users may prefer obtaining results from the same or similar authors. With the rising popularity of bot-generated content, distinguishing original write-ups is essential to ensure their trustworthiness [5, 6].

Authorship Attribution has proven beneficial for information retrieval on many occasions. The work by Kumar et al. [7] utilises information retrieval features for author identification by considering the test document as a query. The authorship attribution in large corpus often incorporates information retrieval techniques. For instance, Yang and Chow [8] present an algorithm that combines profile-based and instance-based approaches for authorship attribution in the forensic domain with thousands of authors.

FDIA'23: Proceedings of the 11th Symposium on Future Directions in Information Access, August 2023, Vienna, Austria
*Corresponding author.
kanishka.silva.92@gmail.com (K. Silva)
0000-0003-2958-9552 (K. Silva)

Similarly, authorship attribution concepts such as recommendation systems have been applied in information retrieval. Alharthi et al. [9] use authorship identification for literary book recommendations as the writing style influences reading preferences. The book recommender model [9] represents books as vectors and evaluates them in a top-k recommendation scenario. Potthast et al. [10] provide a large-scale reproducibility study on authorship identification models, laying the groundwork for integration with information retrieval. They evaluated the reproducibility of all implemented papers and released open-source implementations.

The main research question addressed in this paper is how authorship attribution can be utilised to enhance literary document retrieval considering relevancy, trustworthiness and personalisation factors in information retrieval. Specifically, the proposed methodology was evaluated using a case study in the literary domain with stylometric analysis, genre analysis and author profiling.

The structure of the paper is as follows: Section 2 provides an overview of the related literature on authorship attribution and information retrieval, highlighting research gaps in the area. Section 3 presents the dataset formation progress, and Section 4 outlines the proposed model architecture. Section 5 elaborates on the evaluation plan. Section 6 describes the application of the model architecture for information retrieval in the literary domain. Potential ethical considerations are highlighted in Section 7, followed by the conclusion in Section 8, highlighting the key findings from the literature study and the model architecture analysis.

2. Related Works

Authorship Attribution involves identifying authors of a given document from candidate authors pre-analysed. Research in authorship attribution consists of various approaches, including stylometric feature analysis and comparison mechanisms [11]. The field also explores different disciplines, such as author profile ranking [12, 13, 14] and genre analysis [15, 16, 17] and bot profile identification [18, 19]. In the digital humanities, particularly in the literary domain, authorship attribution addresses unresolved or doubted authorship. The works by Zhao and Zobel [11] were on playwrights in the late 16th and early 17th century of 634 texts and 55 authors, employing features such as POS tagging and function words. Several other literary works are addressed by authorship attribution in digital humanities [20, 21, 22, 23].

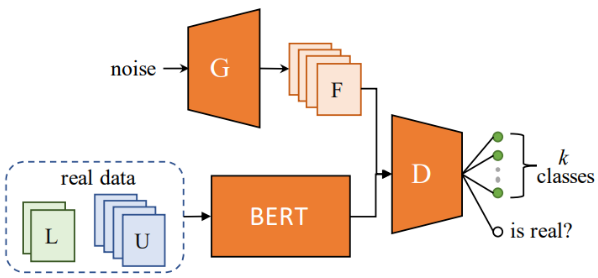


Figure 1: GAN-BERT Model [24]

The GAN-BERT [24] was proposed to improve the text classification in transformer-based models combining BERT-based models and Semi-Supervised GAN [25]. Figure 1 illustrates the GAN-BERT model, where the generator (G) uses random noise to generate forged examples, and the Discriminator (D) performs multi-class classification among authors using BERT embedding. The model is trained on labelled (L) and unlabelled (U) data in the semi-supervised setting. The generator is decoupled from the rest of the model architecture during the testing.

Multimodal architectures combine features in different modalities such as images, categorical, text and numerical features [26, 27, 28, 29] in document classification, sentiment analysis and information extraction, demonstrating outstanding performance compared to single-model architectures. The multimodal transformer proposed by Gu and Budhkar [30] combines text features with categorical or numerical features with feature combination methods, including feature concatenation, attention methods, gating mechanisms, and weighted feature summation. Figure 2 illustrates the architecture of the multimodal transformer.

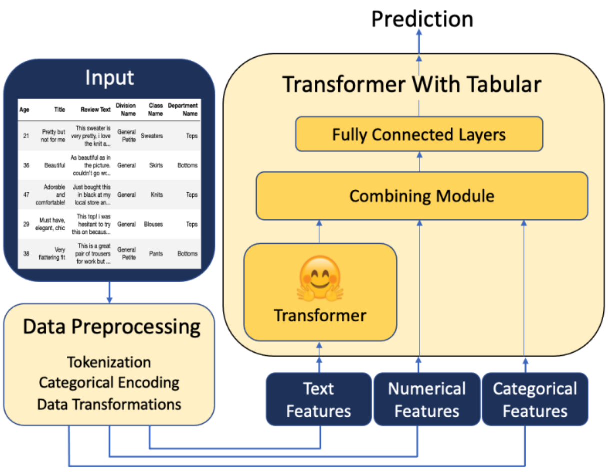
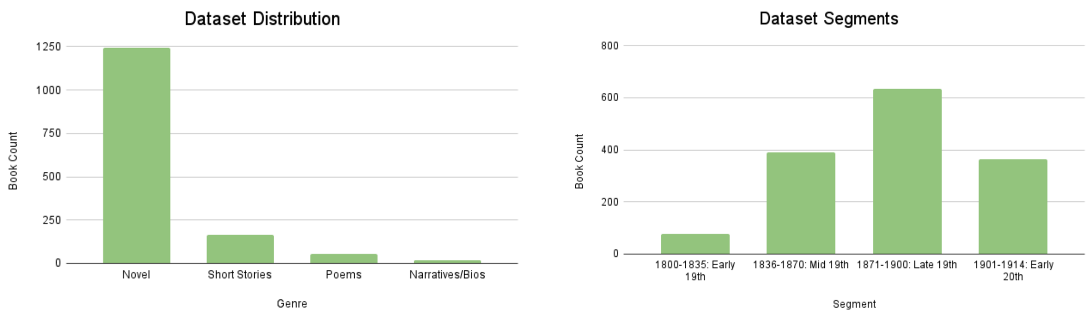


Figure 2: Multimodal Transformer [30]

3. Dataset

A dataset of 1475 documents from the long 19th century (1800-1914) was collected by using Project Gutenberg to support the case study in the literary domain. The dataset contains 1242 Novels, 163 Short Stories, 53 Poems and 17 Narratives, as shown in Figure 3a in text format. For further analysis considering temporal segments, the dataset was divided into different segments: Early 19th century (1800-1835), Mid-19th century (1836-1870), Late 19th century (1871 - 1900) and Early 20th century(1901 - 1914), as illustrated in Figure 3b. The bot profiling model is to be trained with a subset of a dataset comprising corresponding AI-generated novels with similar writing styles, which is to be generated via zero-shot prompting with ChatGPT API.



(a) Dataset Distribution across Different Genres (b) Dataset Distribution across Different Segments

Figure 3: Dataset Statistics for Original Documents

4. Proposed Model Architecture

The proposed model integrates authorship attribution techniques to improve information retrieval, as illustrated in Figure 4. This integration addresses relevancy, trustworthiness and personalisation factors in the information retrieval from literary texts. The model includes genre analysis and bot profiling as filters and uses author profile ranking as an additional sorting mechanism.

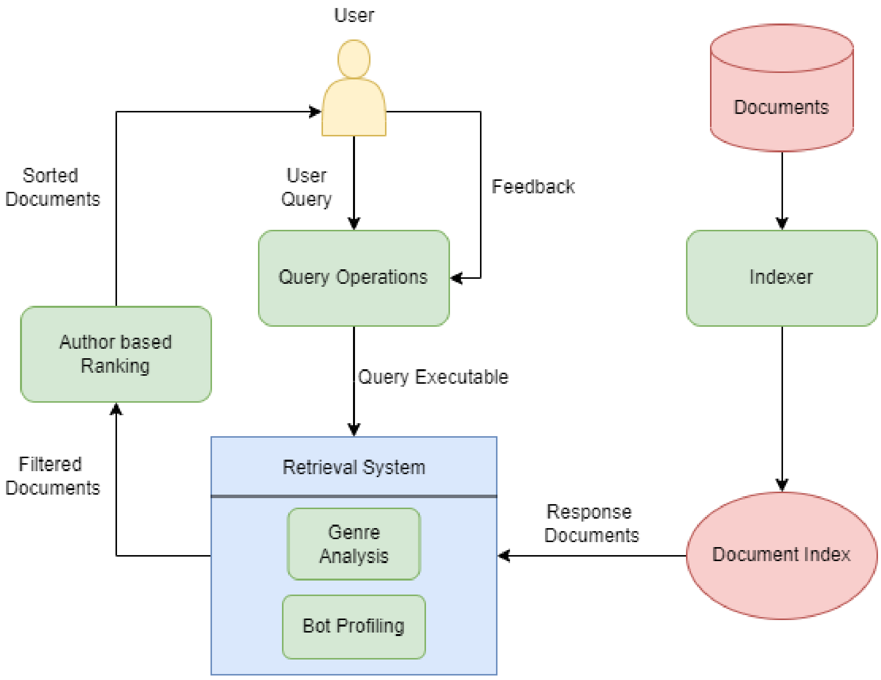


Figure 4: Proposed Model Architecture: Integrated Authorship Attribution Techniques to Information Retrieval Systems

In the author-based ranking, the response documents are sorted according to the authors if mentioned in the query, using stylometric analysis with the GAN-BERT-based model from previous works [31]. The GAN-BERT model encodes the original documents with the BERT model before passing them to the Discriminator. The Discriminator is decoupled from the GAN-BERT model to facilitate the classification in the testing set. The documents are ranked based on the author classification returned from the integrated GAN-BERT model. This integration enhances personalised document retrieval, returning the same or similar author contents.

The returned documents from a search query may not include the relevant genre; therefore, a genre analysis improves the retrieval process to filter out unrelated documents. Recommending a similar author genre based on the filtered documents and the implementation of this component is part of future work.

Given the increasing capability of Large Language Models (LLMs) to generate human-like content, excluding unreliable and misleading bot-generated content in information retrieval is crucial. Based on the previous works [32], a multimodal transformer-based authorship attribution model is suggested to identify AI-generated text content using stylometric analysis. This component is trained with the original and ChatGPT-prompted documents, incorporating stylometric and text-based features in the multimodal transformer. Several handcrafted features such as Average word length (AWL), average sentence length by word (ASW), and functional word counts (CFW) per each document will be considered in the model implementation.

5. Evaluation Plan

A comprehensive evaluation will be conducted to measure the efficiency of the proposed model in enhancing information retrieval within literary documents via authorship attribution methodologies. This evaluation compares the proposed model with a baseline without authorship attribution techniques. Simultaneously, a baseline dataset comprising a collection of retrieved documents per different queries will be utilised and manually annotated to indicate the relevancy or irrelevancy of each document. The proposed model and the baseline model will be evaluated with the baseline dataset using precision, recall, F1-Scare, and Mean Average Precision (MAP) metrics.

6. Case Study Discussion

In this case study related to the literary domain, consider a scenario where a user wants to search for 'The Mahogany Tree' (poem) by an author named 'William' without specifying the author's full name:

1. The genre analysis component filters out documents unrelated to poems. This step ensures that only relevant documents from the poetry genre are considered.
2. The bot profiling component is applied to identify and exclude AI-generated documents, ensuring that unreliable and misleading bot-generated content is excluded from the search results.
3. The author-based ranking is utilised to sort the documents by 'William Makepeace Thackeray' (author), where the sorted documents are presented as the final search results,

with an enhanced retrieval process to provide personalised document recommendations from the same or similar authors.

7. Ethical Considerations

The dataset comprises English novels written between 1800 and 1914, collected from Project Gutenberg [33]. The selected timeframe aligns with Project Gutenberg's criteria for out-of-copyright duration, explicitly adhering to 'Rule 1: Works First Published Before 95 Years Ago and Before 1977' and 'Rule 10(c) - Works of Treaty Parties and Proclamation Countries First Published Between 1923 and 1977' [34]. Although the duration is out-of-copyright regarding literary works, meticulous care was taken to store the data securely with controlled access. Notably, the dataset will remain unreleased.

The author-based ranking model involves generating synthetic documents with similar writing styles to existing authors aiming to cover diverse writing styles comprehensively. Notably, the utilised GAN-BERT model generates embedding representations rather than human-readable text. This design choice substantially limits the possibility of unethical use, as the potential for generating documents using the proposed model without explicit author consent is restricted. Any expansions or applications of this research should conform to the established ethical guidelines. Expressly, any usage of the proposed model should refrain from reproducing or distributing an author's forged content without obtaining appropriate consent.

8. Conclusion

This paper presents ongoing research on using authorship attribution to improve information retrieval in the literary domain. By integrating two previous authorship attribution works with the information retrieval model, this paper aims to improve the retrieval process to provide more relevant and personalised search results. The paper discusses the usage of authorship attribution for information retrieval, with a background study on authorship attribution in author profile ranking, genre analysis, and bot profile identification. Authorship attribution is crucial in addressing unresolved authorship problems in the literary domain. Future directions for this research include implementing the bot profiling, author ranking, and genre analysis models for the case study and prompting ChatGPT to create forged AI content for each document in the dataset. Expanding the dataset to include scientific documents and performing an ablation study for each component is also planned. In conclusion, integrating authorship attribution techniques into information retrieval will improve the performance compared to the state-of-the-art models, particularly in the literary domain.

9. Acknowledgments

This work is supported by the EU Horizon Europe MSCA Staff Exchange project OMINO (ref. 101086321), guarantee-funded by UKRI (ref EP/X040496/1).

References

[1] G. Muresan, C. L. Smith, M. J. Cole, L. Liu, N. J. Belkin, Detecting document genre for personalization of information retrieval, in: 39th Hawaii International International Conference on Systems Science (HICSS-39 2006), CD-ROM / Abstracts Proceedings, 4-7 January 2006, Kauai, HI, USA, IEEE Computer Society, 2006. URL: https://doi.org/10.1109/HICSS.2006.119. doi:10.1109/HICSS.2006.119.

[2] S. Lainé-Cruzel, T. Lafouge, J. P. Lardy, N. B. Abdallah, Improving information retrieval by combining user profile and document segmentation, Inf. Process. Manag. 32 (1996) 305–315. URL: https://doi.org/10.1016/0306-4573(95)00062-3. doi:10.1016/0306-4573(95)00062-3.

[3] S. Kanoje, D. Mukhopadhyay, S. Girase, User profiling for university recommender system using automatic information retrieval, Procedia Computer Science 78 (2016) 5–12. URL: https://www.sciencedirect.com/science/article/pii/S1877050916000041. doi:https://doi.org/10.1016/j.procs.2016.02.002, 1st International Conference on Information Security & Privacy 2015.

[4] R. Arezki, P. Poncelet, G. Dray, D. W. Pearson, Information retrieval model based on user profile, in: C. Bussler, D. Fensel (Eds.), Artificial Intelligence: Methodology, Systems, and Applications, 11th International Conference, AIMSA 2004, Varna, Bulgaria, September 2-4, 2004, Proceedings, volume 3192 of *Lecture Notes in Computer Science*, Springer, 2004, pp. 490–499. URL: https://doi.org/10.1007/978-3-540-30106-6_50. doi:10.1007/978-3-540-30106-6_50.

[5] S. Ramachandran, S. Paulraj, S. Joseph, V. Ramaraj, Enhanced trustworthy and high-quality information retrieval system for web search engines, CoRR abs/0911.0914 (2009). URL: http://arxiv.org/abs/0911.0914. arXiv:0911.0914.

[6] M. Alkhamees, S. Alsaleem, M. Al-Qurishi, M. A. AlRubaian, A. Hussain, User trustworthiness in online social networks: A systematic review, Appl. Soft Comput. 103 (2021) 107159. URL: https://doi.org/10.1016/j.asoc.2021.107159. doi:10.1016/j.asoc.2021.107159.

[7] S. Kumar, S. Rajeswari, M. Srikanth, T. R. Reddy, A new approach for authorship verification using information retrieval features, in: H. S. Saini, R. Sayal, A. Govardhan, R. Buyya (Eds.), Innovations in Computer Science and Engineering, Springer Singapore, Singapore, 2019, pp. 23–29.

[8] M. Yang, K. Chow, Authorship attribution for forensic investigation with thousands of authors, in: N. Cuppens-Boulahia, F. Cuppens, S. Jajodia, A. A. E. Kalam, T. Sans (Eds.), ICT Systems Security and Privacy Protection - 29th IFIP TC 11 International Conference, SEC 2014, Marrakech, Morocco, June 2-4, 2014. Proceedings, volume 428 of *IFIP Advances in Information and Communication Technology*, Springer, 2014, pp. 339–350. URL: https://doi.org/10.1007/978-3-642-55415-5_28. doi:10.1007/978-3-642-55415-5_28.

[9] H. Alharthi, D. Inkpen, S. Szpakowicz, Authorship identification for literary book recommendations, in: E. M. Bender, L. Derczynski, P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, Association for Computational Linguistics, 2018, pp. 390–400. URL: https://aclanthology.org/C18-1033/.

[10] M. Potthast, S. Braun, T. Buz, F. Duffhauss, F. Friedrich, J. M. Gülzow, J. Köhler, W. Lötzsch,

F. Müller, M. E. Müller, R. Paßmann, B. Reinke, L. Rettenmeier, T. Rometsch, T. Sommer, M. Träger, S. Wilhelm, B. Stein, E. Stamatatos, M. Hagen, Who wrote the web? revisiting influential author identification research applicable to information retrieval, in: N. Ferro, F. Crestani, M. Moens, J. Mothe, F. Silvestri, G. M. D. Nunzio, C. Hauff, G. Silvello (Eds.), Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings, volume 9626 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 393–407. URL: https://doi.org/10.1007/978-3-319-30671-1_29. doi:10.1007/978-3-319-30671-1_29.

[11] Y. Zhao, J. Zobel, Searching with style: Authorship attribution in classic literature, in: G. Dobbie (Ed.), Computer Science 2007. Proceedings of the Thirtieth Australasian Computer Science Conference (ACSC2007). Ballarat, Victoria, Australia, January 30 - February 2, 2007. Proceedings, volume 62 of *CRPIT*, Australian Computer Society, 2007, pp. 59–68. URL: http://crpit.scem.westernsydney.edu.au/abstracts/CRPITV62Zhao.html.

[12] M. P. Selvan, A. C. Sekar, Ranking scientific journals based on research author's profile, in: 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2016, pp. 652–654. doi:10.1109/ICCICCT.2016.7988031.

[13] Y. Zhang, M. Wang, M. Zipperle, A. Abbasi, M. Tani, Relrank: A relevance-based author ranking algorithm for individual publication venues, Inf. Process. Manag. 60 (2023) 103156. URL: https://doi.org/10.1016/j.ipm.2022.103156. doi:10.1016/j.ipm.2022.103156.

[14] K. Shu, S. Wang, H. Liu, Understanding user profiles on social media for fake news detection, in: IEEE 1st Conference on Multimedia Information Processing and Retrieval, MIPR 2018, Miami, FL, USA, April 10-12, 2018, IEEE, 2018, pp. 430–435. URL: https://doi.ieeecomputersociety.org/10.1109/MIPR.2018.00092. doi:10.1109/MIPR.2018.00092.

[15] B. Al-Salemi, M. Ayob, G. Kendall, S. A. M. Noah, Multi-label arabic text categorization: A benchmark and baseline comparison of multi-label learning algorithms, Inf. Process. Manag. 56 (2019) 212–227.

[16] S. Ashraf, H. R. Iqbal, R. M. A. Nawab, Cross-genre author profile prediction using stylometry-based approach, in: K. Balog, L. Cappellato, N. Ferro, C. Macdonald (Eds.), Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016, volume 1609 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2016, pp. 992–999. URL: https://ceur-ws.org/Vol-1609/16090992.pdf.

[17] D. A. Pritsos, E. Stamatatos, Open-set classification for automated genre identification, in: P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. M. Rüger, E. Agichtein, I. Segalovich, E. Yilmaz (Eds.), Advances in Information Retrieval - 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings, volume 7814 of *Lecture Notes in Computer Science*, Springer, 2013, pp. 207–217. URL: https://doi.org/10.1007/978-3-642-36973-5_18. doi:10.1007/978-3-642-36973-5_18.

[18] E. van der Walt, J. H. P. Eloff, Using machine learning to detect fake identities: Bots vs humans, IEEE Access 6 (2018) 6540–6549. URL: https://doi.org/10.1109/ACCESS.2018.2796018. doi:10.1109/ACCESS.2018.2796018.

[19] W. Shahid, Y. Li, D. Staples, G. A. Gilkar, S. Hakak, A. A. Ghorbani, Are you a cyborg, bot or human? - A survey on detecting fake news spreaders, IEEE Access 10 (2022) 27069–27083. URL: https://doi.org/10.1109/ACCESS.2022.3157724. doi:10.1109/ACCESS.

2022.3157724.
[20] Y. Yu, W. Liu, Y. Feng, A quantitative study on dream of the red chamber: Word-length distribution and authorship attribution, Complex. 2022 (2022) 9077360:1–9077360:8. URL: https://doi.org/10.1155/2022/9077360. doi:10.1155/2022/9077360.
[21] A. Tikhonov, K. Müller, Scribe versus authorship attribution and clustering in historic Czech manuscripts: a case study with visual and linguistic features1, Digital Scholarship in the Humanities 37 (2021) 254–263. URL: https://doi.org/10.1093/llc/fqaa004. doi:10.1093/llc/fqaa004. arXiv:https://academic.oup.com/dsh/article-pdf/37/1/254/42987140/fqaa004.pdf.
[22] G. Taylor, Finding 'anonymous' in the digital archives: The problem of arden of faversham, Digit. Scholarsh. Humanit. 34 (2019) 855–873. URL: https://doi.org/10.1093/llc/fqy075. doi:10.1093/llc/fqy075.
[23] J. R. Thompson, J. Rasp, Did c. s. lewis write the dark tower?: An examination of the small-sample properties of the thisted-efron tests of authorship, Austrian Journal of Statistics 38 (2016) 71–82.
[24] D. Croce, G. Castellucci, R. Basili, GAN-BERT: generative adversarial learning for robust text classification with a bunch of labeled examples, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 2114–2119. URL: https://doi.org/10.18653/v1/2020.acl-main.191. doi:10.18653/v1/2020.acl-main.191.
[25] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, in: D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, 2016, pp. 2226–2234. URL: https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html.
[26] J. Cheng, I. Fostiropoulos, B. W. Boehm, M. Soleymani, Multimodal phased transformer for sentiment analysis, in: M. Moens, X. Huang, L. Specia, S. W. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, Association for Computational Linguistics, 2021, pp. 2447–2458. URL: https://doi.org/10.18653/v1/2021.emnlp-main.189. doi:10.18653/v1/2021.emnlp-main.189.
[27] L. Yang, J. Na, J. Yu, Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis, Inf. Process. Manag. 59 (2022) 103038. URL: https://doi.org/10.1016/j.ipm.2022.103038. doi:10.1016/j.ipm.2022.103038.
[28] G. V. Singh, M. Firdaus, A. Ekbal, P. Bhattacharyya, Emoint-trans: A multimodal transformer for identifying emotions and intents in social conversations, IEEE ACM Trans. Audio Speech Lang. Process. 31 (2023) 290–300. URL: https://doi.org/10.1109/TASLP.2022.3224287. doi:10.1109/TASLP.2022.3224287.
[29] D. Tomás, R. Ortega-Bueno, G. Zhang, P. Rosso, R. Schifanella, Transformer-based models for multimodal irony detection, J. Ambient Intell. Humaniz. Comput. 14 (2023) 7399–7410. URL: https://doi.org/10.1007/s12652-022-04447-y. doi:10.1007/s12652-022-04447-y.
[30] K. Gu, A. Budhkar, A package for learning on tabular and text data with transformers, in:

Proceedings of the Third Workshop on Multimodal Artificial Intelligence, Association for Computational Linguistics, Mexico City, Mexico, 2021, pp. 69–73. URL: https://www.aclweb.org/anthology/2021.maiworkshop-1.10. doi:10.18653/v1/2021.maiworkshop-1.10.

[31] K. Silva, B. Can, F. Blain, R. Sarwar, L. Ugolini, R. Mitkov, Authorship attribution of late 19th century novels using GAN-BERT, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 310–320. URL: https://aclanthology.org/2023.acl-srw.44. doi:10.18653/v1/2023.acl-srw.44.

[32] K. Silva, I. Frommholz, What if chatgpt wrote the abstract? - explainable multi-authorship attribution with a data augmentation strategy, 2023. (Forthcoming).

[33] Project gutenberg, 2023. URL: https://www.gutenberg.org/.

[34] P. Gutenberg, *Copyright How-To*, May, 2023. https://www.gutenberg.org/help/copyright.html.