Musik in bayern

ISSN: 0937-583x Volume 89, Issue 8 (Aug -2024)

https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*256*

Evolving Landscape of Authorship Attribution in Digital Texts

R. Sowmiya [1] and B. Lavanya [2]

[1] Research Scholar, Department of Computer Science, University of Madras.

[2] Associate Professor, Department of Computer Science, University of Madras.

To Cite this Article

R. Sowmiya, B. Lavanya "Evolving Landscape of Authorship Attribution in Digital Texts".*Musik in Bayern*, *89*(8), 15–19. https://doi.org/10.15463/gfbm-mib-2024-256

Article Info

Received: 02-04-2024 Revised: 2-06-2024 Accepted: 5-07-2024 Published: 16-08-2024

Abstract. In the changing internet landscape, the guarantee that a message genuinely originates from its purported sender is highly in question. Authorship of any digital textual content is to be attributed properly since manipulating it can have dire consequences. Given its cruciality, this work presents a comprehensive study of the various applications and challenges of historic and modern authorship attribution of digital text and its impact. Spanning the root of forensic linguistics, software engineering, and content security in social media platforms, the problems of authorship are reviewed. Additionally, popular datasets have been catalogued with their sources; the usage of several representative features and Large Language Models (LLMs) have been organised to promote future research. This study underscores the necessity of continued research in this direction to protect writer's rights and enhance their online security.

Keywords: natural language processing, authorship attribution, survey.

1. Introduction

In computational linguistics, stylometry is a key term owing to its indispensable roles in legal, social, and academic arenas, among others. It deals with the study of the writing style of documents and authors, usually to reveal other related information. In forensic studies, the linguists handle the task of unveiling details/authors of anonymous notes and letters. After the advent of computational stylometric methods, computational linguists have been involved in this task to produce more accurate results efficiently. However, there persisted challenges regarding their admissibility and eligibility, especially in the legal domain [1].

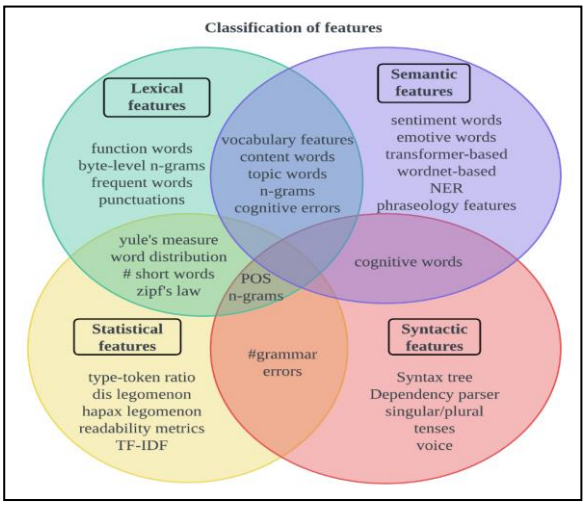
In order to attribute the writings to their proper authors, one needs to identify the pattern or the style or the write-print of the author, which is tricky, more often than not. This task of identifying an author's writing with the evidence of his previous works is called Authorship Attribution (AA). To achieve this, the features under consideration must be capable of capturing their styles. This led to research questions like, "Which feature sets more definitely represent the author's style?", "Which method – Machine Learning (ML), Deep Learning (DL) or Large Language Models (LLM) – is effective in accomplishing this task?" and "Which aspect of the writing reveals more about the writer: Lexical, syntactic or others?"

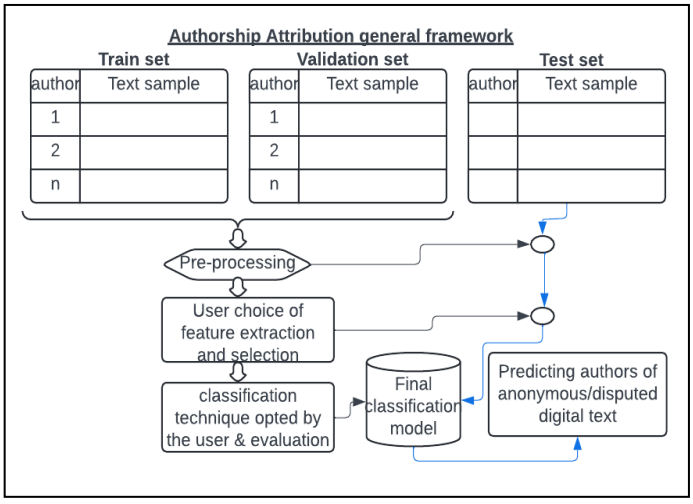
In this survey, the various domains finding applications and future scopes in AA, and the feature varieties as well as Artificial Intelligence (AI) employed for AA are discussed. This paper is structured in the following way:

Section 2 presents the framework of AA and discusses the history of AA problems. Section 3 dives deeper into various branches of AA features and framework and Section 4 presents the popular datasets, Section 5 walks through applications of AA in various domains followed by the conclusion in Section 6.

2. History at a glance

The AA can be approached in two ways: Profile-based and Instance-based methods [2]. Traditional approaches considered the notion of profile-based methods where cumulative corpora made of all the writings of each author are analysed to arrive at a conclusion about the authorship of the disputed document. In contrast, in instance-based methods, each independent sample of writing is utilised to learn the authorial style. Notably, the centuries-old work of [3] attempted to resolve whether some literary works were written by Shakespeare, John Fletcher or Francis Bacon; It established the word length distribution of a document to be representative of an author's style. The authors of [4] investigated the frequency of words and applied the





Bayesian theorem to judge the writers of the 12 disputed papers out of 85 Federalist papers published by the New York Press, during 1787-1788. Predominantly, statistical methods were explored in traditional AA. A statistical method, not based on the frequency of words or common words, but rather the measures of the central tendency of sentence length was proposed by [5]. Yet another method incorporating the statistical technique of multivariate (cluster) analysis with modified feature selection of frequent words was introduced in [6]. An important breakthrough in AA [7] took place in the early 21st century when Burrow's delta measure (employing the z-score) was introduced. From these evidences, the weightage given to the statistics and lexical component of text to distinguish the style of the authors can be well-perceived.

3. AA framework and classification of features

The distinction between a few feature groups is often fuzzy. In this work [8], the authors classify vocabulary features under semantic features whereas the authors of [9] mention it under lexical features. In yet another study, character, word and Parts-of-the-Speech (POS) n-grams are grouped under content features [10] whereas it could be traced in the lexical feature section in [8]. Thus, to classify features, the perspective of the author and the corpus under consideration play a crucial role. The classification of some AA features carried out along with the AA framework is given in Figure 1.

Figure 1. Classification of features and general framework of AA

4. Catalogue of Benchmark and Recent Datasets

Benchmark datasets in AA may include CCAT 50, CCAT 10, Federalist papers and Enron mail to mention a few whereas recent datasets such as ElectAI [11] for the attribution of AI-writing/ LLM-writing apart from the human writing are also catalogued along with several other datasets in Table 1.

Table 1: Datasets – sources and details

Dataset type	Source	Details and Year of Publishing
Enron mail AA dataset	https://www.cs.cmu.edu/~./enron/	0.5 Million messages from 150 users (2015)
Plagiarism detection dataset	https://doi.org/10.5281/zenodo.3250095	Amazon Mechanical Turk Vs computer-created plagiarised data (2011)
Blog AA	https://www.kaggle.com/datasets/rtatman/blog-authorship-corpus/data	681,288 posts from 19,320 bloggers (2004)
Cross-domain multilingual AA	https://zenodo.org/records/3530313	The English language is also included (2019)
AI – SOCO 2020 - Code AA	https://zenodo.org/records/4059840	6553 problems – 1000 users (2020)
Github – Java – Code AA	https://ieee-dataport.org/documents/github-dataset-authorship-attribution	172,919 Java source codes from 3,128 authors (2021)
COLL - Code AA	http://hdl.handle.net/102.100.100/166	4777 files coded by 558 authors, in C, C++ and Java (2007-2014)
ElectAI – Tweet and AI - AA	https://github.com/LanguageTechnologyLab/ElectAI/tree/main	1550 tweets by human and AI authors (2024)
Writing prompts used with human-written text - AI AA	https://www.kaggle.com/datasets/ratthachat/writing-prompts/data	300K human-written stories paired with writing prompts from an online forum (2018)
Victorian writers AA	https://archive.ics.uci.edu/dataset/454/victorian+era+authorship+attribution	50 well-known authors with 93600 instances (2018)
Federalist papers dataset	https://www.kaggle.com/datasets/tobyanderson/federalist-papers/data	85(attributed) + 11(disputed) papers authors Hamilton, Madison, or Jay (1818)
Judgement AA – legal dataset	https://umlt.infotech.monash.edu/?page_id=152	Legal dataset – 3 judges with 2313 judgements (2011)
Tweets dataset	https://github.com/theocjr/social-media-forensics	128 million messages (tweets) from 50,000 Twitter users during 2018
CCAT 50 (Reuters dataset)	https://doi.org/10.5281/zenodo.3759068	2500 texts from 50 different authors (2015)
CCAT 10 (Reuters dataset)	https://doi.org/10.5281/zenodo.3759064	500 texts from 10 different authors (2015)

5. Application, Challenges and Scopes of AA in various domains

5.1. Forensic AA

Researchers are working on methods to find the source of fraudulent emails by examining sentence arrangement, word selections and stylometric attributes [12]. By doing this, email service companies can filter out spam and shield customers from scams. AA can also be utilised in forensic examinations to determine the sender of emails and short messages (SMS) that contain threats or harassment. Nevertheless, the efficacy of these methods depends on the intricacy of the email content and the accession of substantial email samples [13].

5.2. Code AA

Failing to address the problem of code AA will not only lead to copyright infringement of the works of the original author but also hinder the effort to trace the sources of viruses and malicious codes [14]. In the study of style in code AA multiple techniques are emerging: To capture the syntax of the program, Abstract syntax trees and parse trees can be opted for, and as for lexical features byte-level n-grams, bit-level features

(for binary code) and other tokens are commonly utilised. Code obfuscation is a rising concern in this field of study where malicious coders format or hide their style to conceal their information.

5.3. LLM-based AA

Interesting outbreaks in the frontiers of AA dealing with LLMs have been rising in recent times. One such research examines [15] how much the LLMs producing written samples in English (mimicking humans) deceive the LLM trained to classify authors based on their styles. Their findings confirm that LLM's mimicking was successful, cautioning the research community to work on preventive measures for this problem. Studies show interest in discrimination and attribution of AI vs Human authorship in social media content, from the political perspective also [11]. Another recent study dives deeper to explore whether LLMs could score better in AA, and how well they could help with explainable classification with linguistic features in writing [16].

5.4. Social media AA

Social networking networks provide anonymity, therefore strong AA methods are required. This is especially important to spot the Astroturfers (i.e. people who (are usually paid to) promote an impression or deceptive opinion about an organisation or a political party through social media platforms) [17], harmful individuals and fraudulent posts. AA in social media posts, for instance, can help curb the dissemination of false information by confirming the legitimacy of digital text. The primary difficulty in social media text authorship is the informal, evolving and brief nature of the digital text [18].

6. Conclusion

This work comprehensively studies and organises widely dispersed data about the historic and emerging trends in the domain of authorship attribution. The recent innovations in LLM over Human authorship have been discussed. More research in this aspect may alleviate AI plagiarism issues in scientific articles. A wide range of relevant datasets have been catalogued in this work to encourage future research in this direction.

7. Acknowledgements

This work is supported by the University Grants Commission of India under the Junior Research Fellowship (JRF) scheme.

8. References

[1] L M. Solan, Ph.D. Intuition Versus Algorithm: The Case of Forensic Authorship Attribution. *Journal of L & Pol'y*. 2013, **21** (2): 551-576. Available at: https://brooklynworks.brooklaw.edu/jlp/vol21/iss2/13

[2] E Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*. 2009, **60** (3): 538-556.

[3] T.C. Mendenhall. The Characteristic Curves of Composition. *Science*, 1887, **9** (214): 237-49.

[4] F. Mosteller, and D.L. Wallace. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *Journal of the American Statistical Association*. 1963, **58** (302): 275-309. Available at: https://doi.org/10.2307/2283270

[5] G. U. Yule. On Sentence-Length as a Statistical Characteristic of Style in Prose: With Application to Two Cases of Disputed Authorship. *Biometrika*. 1939, **30** (3/4): 363-390. https://doi.org/10.2307/2332655

[6] D. L. Hoover. Multivariate Analysis and the Study of Style Variation, *Literary and Linguistic Computing*. 2013, 18(4): 341–360. https://doi.org/10.1093/llc/18.4.341

[7] S. Evert, T. Proisl, F. Jannidis, I. Reger, S. Pielström, C. Schöch and T. Vitt. Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*. 2017, **32** (2): ii4–ii16.

[8] X. He, A.H. Lashkari, N. Vombatkere and D.P. Sharma. Authorship Attribution Methods, Challenges, and Future Research Directions: A Comprehensive Survey. *Information*. 2024, **15** (131): 1-42.

[9] A. Modupe, T. Celik, V. Marivate and O.O. Olugbara. Post-Authorship Attribution Using Regularized Deep Neural Network. *Appl. Sci*. 2022, **12** (7518): 1-24. https://doi.org/10.3390/app12157518

[10] H. Wu, Z. Zhang and Q. Wu. Exploring syntactic and semantic features for authorship attribution. *Applied Soft Computing*. 2021, **111** (107815): 1-11. https://doi.org/10.1016/j.asoc.2021.107815

[11] A. Dmonte, M. Zampieri, K. Lybarger and M. Albanese. Classifying Human-Generated and AI-Generated Election Claims in Social Media. *arXiv* preprint arXiv:2404.16116. 2024.

[12] K.A. Apoorva and S. Sangeetha. Deep neural network and model-based clustering technique for forensic electronic mail author attribution. *SN Applied Sciences*. 2021, **3** (3): 348

[13] F. Iqbal, R. Hadjidj, B.C. Fung and M. Debbabi. A novel approach of mining write-prints for authorship attribution in e-mail forensics. *Digital Investigation*. 2008, **5**: S42-S51. https://doi.org/10.1016/j.diin.2008.05.001

[14] V. Kalgutkar, R. Kaur, H. Gonzalez, N. Stakhanova and A. Matyukhina. Code authorship attribution: Methods and challenges. *ACM Computing Surveys (CSUR)*. 2019, **52** (1): 1-36. https://doi.org/10.1145/3292577

[15] K. Jones, J.R. Nurse and S. Li. May. Are you robert or roberta? Deceiving online authorship attribution models using neural text generators. *Proc. of the International AAAI Conference on Web and Social Media*. 2022, **16**, pp. 429-440.

[16] B. Huang, C. Chen and K. Shu. Can Large Language Models Identify Authorship? *arXiv*:2403.08213. 2024.

[17] J. Peng, R.K.K. Choo and H. Ashman. August. Astroturfing detection in social media: Using binary n-gram analysis for authorship attribution. *Proc. 2016 IEEE Trustcom/BigDataSE/ISPA*. 2016, pp. 121-128.

[18] O. Fourkioti, S. Symeonidis and A. Arampatzis. Language models and fusion for authorship attribution. *Information Processing & Management*. 2019, **56** (6): 102061.

Authors' background

Your Name	Title*	Research Field	Personal website
R. Sowmiya	Phd candidate	Natural Language Processing, Stylometry and Text mining	
B. Lavanya	Associate professor	Natural Language Processing, Data Mining and Bioinformatics	

*This form helps us to understand your paper better, the form itself will not be published.

*Title can be chosen from: master student, Phd candidate, assistant professor, lecture, senior lecture, associate professor, full professor