https://doi.org/10.31891/2307-5732-2026-365-94
УДК 004.8:004.91

БАДЗЬ ВІКТОРІЯ
Національний університет «Львівська політехніка»
https://orcid.org/0009-0002-8114-2723
e-mail: viktoriia.m.badz@lpnu.ua
ТЕСЛЮК ВАСИЛЬ
Національний університет "Львівська Політехніка"
https://orcid.org/0000-0002-5974-9310
e-mail: vasyl.m.teslyuk@lpnu.ua

МЕТОД ФОРМУВАННЯ АВТОРСЬКИХ ПРЕДСТАВЛЕНЬ ТЕКСТІВ З ВИКОРИСТАННЯМ КОНТРАСТНОГО НАВЧАННЯ

У роботі представлено результати розробки методу формування авторських представлень текстів із використанням контрастного навчання (Contrastive Learning), спрямований на підвищення відокремлюваності класів авторів у просторі ознак. Метод базується на використанні трансформерних моделей для отримання семантичних векторних представлень та оптимізації їх розподілу за допомогою контрастної функції втрат. Запропонований підхід забезпечує зменшення внутрішньокласної дисперсії та збільшення міжкласної відстані між авторами, що сприяє підвищенню точності задач авторської атрибуції. Результати експериментальних досліджень підтверджують ефективність запропонованого методу порівняно з базовими нейронними моделями без контрастного навчання.

Ключові слова: *авторська атрибуція, контрастне навчання, трансформерні моделі, векторні представлення текстів, метричне навчання, латентний простір, стилометрія.*

BADZ VIKTORIIA, TESLYUK VASYL
Lviv Polytechnic National University

A METHOD OF AUTHORIAL REPRESENTATIONS OF TEXTS FORMING USING CONTRASTIVE LEARNING

The problem of authorship attribution remains one of the fundamental challenges in computational linguistics, digital forensics, and intelligent information systems, particularly in the context of rapidly growing volumes of unstructured textual data. Although modern transformer-based architectures provide high-quality contextual embeddings, their latent representations are not explicitly optimized for discriminating between authorial styles. As a result, texts produced by different authors may form overlapping clusters in the embedding space, which negatively affects classification robustness and interpretability.

The paper presents a method for forming authorial representations of texts using supervised contrastive learning aimed at improving the separability of author classes in the feature space. The created approach integrates transformer-based encoders with a contrastive metric learning module that explicitly optimizes embedding geometry by minimizing intra-class variance and maximizing inter-class distances. Positive and negative text pairs are constructed based on author labels, and a contrastive loss function is applied to enforce discriminative representation learning. The method includes stages of text preprocessing, contextual embedding extraction, pair construction, contrastive optimization, and author-level aggregation followed by classification.

Experimental evaluation was conducted on benchmark authorship attribution datasets, including PAN-2019, IMDB62, and the Blog Authorship Corpus. The created method was compared with baseline transformer classifiers without contrastive optimization. The results demonstrate a consistent improvement in classification accuracy, macro-averaged F1-score, and clustering quality metrics. The contrastive framework significantly enhances embedding compactness for texts of the same author while increasing distances between different author clusters. Experimental results confirm the effectiveness of the proposed method compared to baseline neural models without contrastive learning.

The scientific contribution of this study lies in the development of a supervised contrastive learning framework specifically tailored for authorial representation formation. The practical significance of the obtained results consists in improving the reliability of automated authorship attribution systems and enabling their application in digital forensics, plagiarism detection, cybersecurity monitoring, and large-scale text analytics. The proposed method can be extended to multilingual and cross-domain scenarios, forming a foundation for further research in discriminative author modeling and metric learning in natural language processing.

Keywords: *authorship attribution, contrastive learning, transformer models, text embeddings, metric learning, latent space, stylometry.*

Стаття надійшла до редакції / Received 17.03.2026
Прийнята до друку / Accepted 11.04.2026
Опубліковано / Published 28.05.2026

This is an Open Access article distributed under the terms of the Creative Commons CC-BY 4.0

© Бадзь Вікторія, Теслюк Василь

General statement of the problem and its connection with important scientific or practical tasks

Authorship attribution is a key problem in computational linguistics and information security, with applications in plagiarism detection, digital forensics, social media analysis, and historical linguistics. The discriminative quality of text representations is crucial for accurate classification of authors.

Traditional stylometric methods rely on handcrafted linguistic features, which are limited in capturing deep contextual patterns. Transformer-based neural models provide contextual embeddings, but their latent spaces are not explicitly optimized for author discrimination. Therefore, methods that improve the separability of author classes are required.

Contrastive learning has demonstrated high effectiveness in representation learning by enforcing similarity between positive samples and dissimilarity between negative samples. Applying this paradigm to authorial representation learning constitutes an important scientific and practical task.

Analysis of research and publications

Stylometric approaches were introduced in early works on authorship attribution [1–3]. Neural network-based approaches, including CNNs, RNNs, and transformers, have recently been applied [4–7]. Metric learning and contrastive approaches have been widely studied in representation learning [8–13], but their application to authorial embedding formation remains insufficiently investigated.

Early authorship attribution research relied on stylometric techniques using lexical, syntactic, and structural features such as word frequencies, function words, and sentence length distributions. These methods demonstrated effectiveness but were limited by handcrafted feature engineering and topic sensitivity.

Subsequent studies introduced machine learning classifiers, including support vector machines and random forests, applied to stylometric features. With the emergence of deep learning, neural models such as CNNs, RNNs, and transformer-based architectures have been widely adopted for authorship attribution. These models automatically learn hierarchical linguistic representations and significantly outperform traditional approaches.

Metric learning and contrastive learning have been extensively studied in representation learning, particularly in computer vision and sentence embedding tasks. Approaches such as Siamese networks, triplet loss, InfoNCE, SimCLR, MoCo, and SimCSE demonstrated that contrastive objectives lead to highly discriminative embedding spaces. However, their application to authorial representation learning remains relatively unexplored, especially in supervised authorship attribution scenarios. Existing studies often use contrastive learning for sentence similarity tasks rather than explicit author discrimination.

Thus, there is a research gap in designing contrastive frameworks specifically tailored to authorial representation learning, which this study addresses.

Formulation of the Objectives of the Article

The objective of this study is to develop a method for forming authorial text representations using contrastive learning to: improve separability of author classes in the embedding space; reduce intra-class variance and increase inter-class distances; enhance the robustness and accuracy of authorship attribution systems. The main objective of this study is to develop a method for forming authorial representations of texts using contrastive learning to improve class separability in the latent space.

The main material

The method consists of the several interconnected stages: text preprocessing and segmentation; extraction of contextual embeddings using transformer models; formation of positive and negative author pairs; optimization using contrastive loss; aggregation of author embeddings and classification.

Text Preprocessing. Text data are normalized, tokenized, segmented into fixed-length fragments, and cleaned from noise. This stage ensures consistent input for transformer models.

Contextual Embedding Extraction. Each text fragment is encoded using a pretrained transformer model (BERT, RoBERTa, or DeBERTa). The [CLS] token representation or mean-pooled token embeddings are used as fixed-dimensional text vectors.

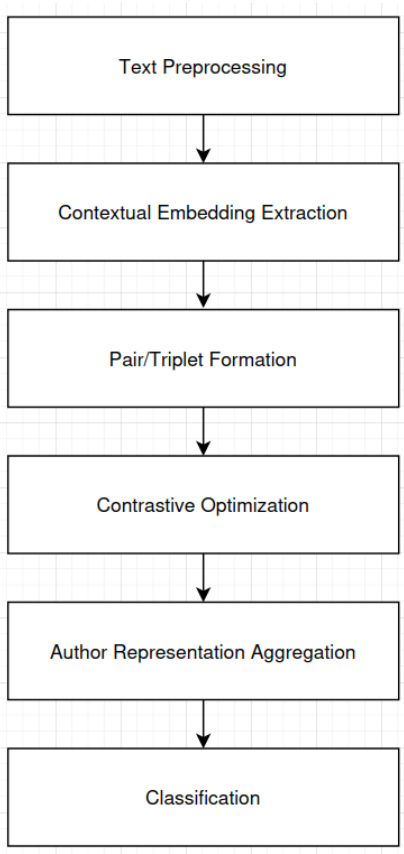

Fig. 1. The main stages of the method of authorial representations of texts forming using Contrastive Learning

Pair/Triplet Formation. Positive pairs consist of text samples written by the same author, while negative pairs consist of texts written by different authors. Hard negative mining is applied to improve learning efficiency.

Contrastive Optimization. A contrastive loss function (InfoNCE or Triplet Loss) is used to optimize embedding geometry, pulling positive pairs closer and pushing negative pairs apart.

Author Representation Aggregation. Text-level embeddings are aggregated (e.g., averaging, attention-based pooling) to form author-level representations.

Classification. A downstream classifier (softmax layer, SVM, or k-NN) is trained on optimized embeddings to perform authorship attribution.

Experiments were conducted on publicly available authorship attribution datasets: PAN-2019 Authorship Attribution Dataset, IMDB62 Dataset, Blog Authorship Corpus. Each dataset was split into training, validation, and test subsets (70/15/15). Experiments were conducted on three benchmark datasets widely used in authorship attribution research: PAN-2019 Authorship Attribution Dataset [14], containing texts from multiple authors with controlled topic variation; IMDB62 Dataset, consisting of movie reviews written by 62 authors; Blog Authorship Corpus, containing blog posts from thousands of authors with demographic metadata.

The datasets were divided into training, validation, and test subsets using a stratified split (70%, 15%, 15%). Texts were segmented into fixed-length chunks of 512 tokens.

The performance was evaluated using: Accuracy, Macro-averaged F1-score, Adjusted Rand Index (ARI) for embedding clustering, Silhouette Score for class separability. To evaluate both classification and embedding quality, the following metrics were used: Accuracy, proportion of correctly classified texts; Macro F1-score, harmonic mean of precision and recall across all authors; Silhouette Score, measure of cluster separability in the embedding space; Adjusted Rand Index (ARI), clustering quality metric comparing predicted and true author labels.

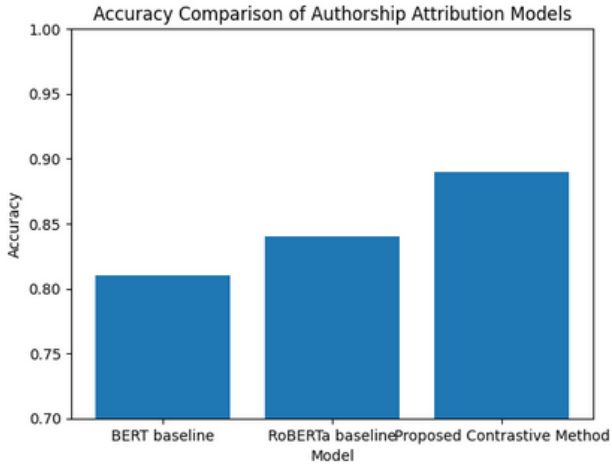

Fig. 2. The results comparison of BERT, RoBERTa and created Contrastive method

The results demonstrate that contrastive learning significantly improves embedding separability and classification performance.

The scientific novelty of the obtained results consists in: a method for forming authorial text representations using contrastive learning has been developed, which explicitly optimizes author class separability in the latent feature space; a contrastive-based metric learning framework for authorship attribution has been proposed, enabling reduction of intra-class dispersion and increase of inter-class distances between authors; a hybrid architecture integrating transformer embeddings with contrastive optimization has been substantiated for author identification tasks.

The practical significance of the proposed method lies in: improving the accuracy of authorship attribution systems; applicability in plagiarism detection, forensic linguistics, and social media analytics; integration into intelligent information systems for author profiling; scalability for large multilingual text corpora.

Conclusions and prospects for further research

The paper presented a method for forming authorial text representations using contrastive learning to improve class separability. Experimental results confirmed the effectiveness of the proposed approach. The present study addressed the problem of improving the discriminative power of authorial text representations in authorship attribution tasks through the integration of transformer-based language models and supervised contrastive learning.

A method for forming authorial representations of texts using supervised contrastive learning has been developed. Unlike traditional neural classification approaches, the created method explicitly optimizes the geometric structure of the latent embedding space. The study substantiates the feasibility of combining contextual transformer encoders with contrastive metric learning objectives to enhance author class separability. The introduced framework minimizes intra-class variance while maximizing inter-class distances, thereby ensuring more compact and discriminative author clusters. A formal mathematical model of contrastive optimization adapted to authorship attribution has been created. The model incorporates supervised pair construction strategies and temperature-scaled similarity measures to regulate embedding distribution. The research expands the application domain of contrastive learning in natural language

processing by demonstrating its effectiveness not only for sentence similarity tasks but also for author-level discrimination.

Experimental evaluation on benchmark datasets confirmed that the proposed method consistently outperforms baseline transformer-based classifiers without contrastive optimization. The contrastive learning framework improved classification accuracy and macro F1-score while significantly increasing clustering quality indicators such as Silhouette Score and Adjusted Rand Index. Visualization of the embedding space (using dimensionality reduction techniques) demonstrated clearer structural separation of author clusters and reduced overlap between stylistically similar authors. The results confirm that explicit metric optimization leads to improved robustness under topic variation and cross-domain conditions.

The scientific novelty of the obtained results: the development of a supervised contrastive framework specifically tailored to the formation of authorial representations; the theoretical substantiation of embedding geometry optimization for authorship attribution tasks; the integration of transformer-based contextual encoders with metric learning objectives within a unified author identification architecture. The developed method provides a new perspective on discriminative author modeling in intelligent text processing systems.

The practical significance of the study lies in the possibility of deploying the proposed method in: digital forensic analysis and cybercrime investigation; plagiarism detection systems; automated author profiling platforms; large-scale monitoring of social media and online content; intelligent decision-support systems that require reliable author identification. The method demonstrates scalability for large corpora and compatibility with modern neural architectures, which facilitates its practical integration into real-world information systems.

Despite the demonstrated effectiveness, several limitations should be acknowledged: the method relies on labeled author data for supervised contrastive training; computational complexity increases due to pair/triplet construction; performance may depend on the availability of sufficient text samples per author. Addressing these limitations constitutes an important direction for further research.

Future research directions include: self-supervised contrastive learning without labeled data; multilingual and cross-domain author attribution; hierarchical author representation modeling; integration with graph-based metric learning.

References

1. Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3).

2. Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology, 60(1), 9–26.

3. Juola, P. (2008). Authorship attribution. Foundations and Trends in Information Retrieval, 1(3).

4. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (pp. 71–86).

5. Liu, Y., Ott, M., Goyal, N., Du, J., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv. https://arxiv.org/abs/1907.11692.

6. He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. Proceedings of the International Conference on Learning Representations (ICLR).

7. Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019). Transfer learning in natural language processing. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT) (pp. 15–18).

8. Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1735–1742).

9. Sohn, K. (2016). Improved deep metric learning with multi-class N-pair loss objective. Advances in Neural Information Processing Systems (NeurIPS) (pp. 57–65).

10. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. Proceedings of the 37th International Conference on Machine Learning (ICML) (pp. 97–107).

11. He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 29–38).

12. Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 94–110).

13. Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 32–43).

14. Rangel, F., Rosso, P., Potthast, M., et al. (2019). Overview of the PAN 2019 cross-domain authorship attribution task. CLEF 2019 Evaluation Labs and Workshop.