Style Change Detection Using Graph and Structural Linguistic Features for Multi-Author Writing Analysis

Notebook for the PAN Lab at CLEF 2025

Ioana-Roxana Boriceanu*, Andra-Elena Băltoiu

National University of Science and Technology POLITEHNICA Bucharest, Romania

Abstract

This paper presents our approach to the Multi-Author Writing Style Analysis task at PAN 2025. The goal is to detect sentence level style changes that may indicate a shift in authorship. We propose a handcrafted, feature based pipeline that integrates graph based properties from Word Adjacency Networks (WANs), lexical and syntactic measures, sentence level context features, and similarity metrics computed over embeddings produced by Sentence-BERT (SBERT). The system is tuned to handle all three levels of difficulty by adapting feature processing and model calibration. Predictions are made using a Gradient Boosting classifier. Results on the validation and test sets show that our interpretable and lightweight method performs competitively across all difficulty levels.

Keywords

Style Change Detection, Authorship Attribution, Word Adjacency Networks,

1. Introduction

Stylometry is the computational analysis of writing style and has been widely applied in authorship attribution [1], plagiarism detection [2], and other related tasks. It relies on the idea that authors leave behind consistent stylistic traces, even when writing about similar topics. These traces can include lexical choices, syntactic preferences, punctuation patterns, and sentence structure, among others.

While traditional stylometric studies focus on attributing entire documents to a known set of authors, recent efforts have explored more fine-grained tasks, such as detecting shifts in writing style within a single document. These style shifts may occur in collaborative writing, edited content, or deceptive texts authored by multiple individuals. Unlike full document attribution, this sentence level task requires high sensitivity to local changes in style and the ability to distinguish them from content or topic variation.

The PAN 2025 lab [3, 4] addresses the problem of style change detection through the Multi-Author Writing Style Analysis task. The goal is to detect changes in writing style at the sentence level within a document written by multiple authors. This task is intrinsic, meaning that no reference texts or author profiles are provided. The system must identify boundaries where the style shifts, using only information from within the document itself. Each sentence is assumed to be written by a single author, and style changes are assumed to occur only between sentences.

The dataset includes three difficulty levels: easy, medium, and hard. These levels differ in how much topical variation is present within each document. In the hard setting, all sentences are on the same topic, so topical cues are not useful. In the easy and medium settings, there is some variation in topic, but relying too heavily on topic shifts can be misleading.

Our system does not try to separate style from content directly. Instead, it uses a diverse set of features that are designed to capture stylistic information. We adjust the feature configuration depending on the difficulty level. For each level, we train a separate Gradient Boosting classifier, apply appropriate feature scaling, and adapt the graph based analysis accordingly. The system is lightweight, interpretable, and performs well across all difficulty settings.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

ioana.boriceanu@stud.acs.upb.ro (I. Boriceanu); andra.baltoiu@upb.ro (A. Băltoiu)

0009-0008-5867-1516 (I. Boriceanu); 0000-0003-3600-0531 (A. Băltoiu)

2. Related Work

Previous research on style change detection has largely been shaped by the PAN shared tasks, which have progressively introduced more fine-grained and realistic scenarios for intrinsic authorship analysis [5, 6]. These tasks have encouraged systems to go beyond surface level features and explore deeper representations of writing style. Recent advancements in style change detection have leveraged a diverse array of methodologies, ranging from deep contextual embeddings to graph based and structural-linguistic features. In the PAN 2024 shared task, numerous teams adopted transformer based models such as BERT [7] and RoBERTa [8] to identify stylistic shifts at the paragraph level.

Many top-performing systems in PAN 2024 relied heavily on deep contextual embeddings from pretrained language models. For instance, one team combined RoBERTa, DeBERTa, and ERNIE models within a majority voting framework, achieving strong results across all difficulty levels [9]. Another system, which ranked first overall, utilized embeddings from the LLaMA-3-8B decoder, fine-tuned using low-rank adaptation (LoRA) to perform label classification, demonstrating the effectiveness of large language models (LLMs) in capturing subtle stylistic variation [10].

In addition to these deep models, another approach integrated transformer based embeddings with handcrafted stylometric features to enhance interpretability and robustness [11]. The team combined RoBERTa representations with features that reflect text formality, grammatical structure, and readability, including metrics such as the Flesch-Kincaid grade level and the SMOG index. The authors also used the Mann–Whitney U test to assess whether differences in feature distributions were statistically significant across authorial boundaries, reaffirming the continued relevance of classical stylistic features in detecting writing style changes.

Graph based representations of text have been explored as an alternative to purely sequential or embedding based approaches in authorship analysis. One such method, introduced in [12], transforms texts into syntactically informed graphs where words are represented as nodes labeled with their part-of-speech (POS) tags. Edges are created based on sentence structure and syntactic grouping, and the resulting graph is characterized using centrality measures such as degree, closeness, betweenness, and eigenvector centrality. These features are then used to train classification models for author identification. A different strategy proposed in [1] applies graph modeling to multi-author documents by constructing Co-Authorship Graphs (CAGs), where text segments are connected based on stylistic similarity computed via modified Hausdorff distance. These models have shown promising results on both synthetic and real world datasets, offering a structurally motivated perspective on style change detection.

3. Approach

3.1. Dataset

The PAN 2025 Style Change Detection dataset comprises English language documents constructed from Reddit comments [13]. Each document is a sequence of sentences authored by multiple individuals, with the objective being to identify the positions at which the author changes. Specifically, for each pair of consecutive sentences, the task is to determine whether a change in authorship has occurred. The dataset is divided into three difficulty levels: easy, medium, and hard. In the easy set, sentences cover a variety of topics, allowing models to utilize topic information as a cue for detecting authorship changes. The medium set contains documents with limited topical variety, compelling models to focus more on stylistic features. The hard set consists of documents where all sentences pertain to the same topic, necessitating reliance solely on stylistic cues for detecting author changes.

3.2. Data Processing

Each document in the dataset is first segmented into individual sentences. These sentences undergo a series of linguistic preprocessing steps including tokenization, part-of-speech (POS) tagging, lemmati-

zation, and removal of stopwords and punctuation. The specific combination of preprocessing steps depends on the difficulty level and is controlled through a configuration schema.

To model the local lexical and syntactic structure of sentences, we construct Word Adjacency Networks (WANs). After preprocessing, each word in a sentence is represented as a node in a directed graph. A directed edge is added between two nodes if the corresponding words appear consecutively in the sentence. If an edge between two nodes already exists, its weight is incremented to reflect the frequency of that word pair. Redundant edges, such as self-loops, are removed to maintain structural clarity. In certain configurations, additional nodes and edges are introduced to represent POS tag transitions, allowing the network to capture grammatical relationships beyond surface word order. These enriched networks offer a stylometric representation of sentence structure and are used to extract graph features such as centrality scores, clustering coefficients, and entropy measures.

A simplified illustration of a WAN is presented in Figure 1, based on the famous phrase "To be, or not to be, that is the question" from Hamlet by William Shakespeare (Act III, Scene 1). Each node represents a token (e.g., word or punctuation mark) in the sentence and may be labeled with its corresponding part-of-speech (POS) tag, depending on the preprocessing configuration. The POS tag, shown below the token, follows the Penn Treebank format and was generated automatically using the spaCy NLP toolkit [14]. To maintain visual clarity, the figure omits separate POS transition nodes and edges. It is intended as a minimal illustration of lexical adjacency rather than a full depiction of the enriched network structure. The first word is highlighted in green and the final word in red.

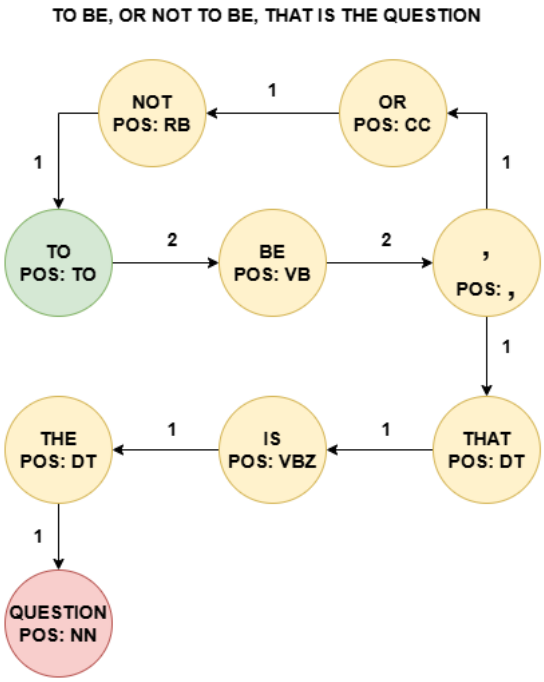

Figure 1: Word Adjacency Network example

To study the impact of preprocessing choices on the strength of stylistic signals, we define sixteen WAN configurations, representing all possible combinations ($2^4 = 16$) of four optional preprocessing steps: punctuation removal, stopword removal, lemmatization, and POS transition modeling. Each configuration can be represented using a four-bit binary mask, where each bit indicates whether a preprocessing step is applied (1) or omitted (0), in the order: punctuation, stopwords, lemmatization, POS tags. For example, the mask 1111 indicates that all steps are applied, while 0000 means no preprocessing is performed. For each difficulty level in the task, we selected a configuration that empirically achieves a balance between preserving stylistic signals and suppressing topical or content related noise.

3.3. Feature Engineering

To detect stylistic shifts at the sentence level, we extract a diverse set of features that reflect lexical, syntactic, semantic, and discourse level patterns. These features fall into six broad categories and were primarily selected based on their demonstrated utility in previous work on authorship attribution and stylistic analysis, as discussed in the Related Work section:

- **Embedding based similarity metrics:** computed over SBERT embeddings, including cosine similarity, Euclidean distance, Manhattan distance, and norm ratio between sentence representations. We also include semantic drift across skip distances and directional embedding angles to capture local variation in meaning.
- **WAN features:** graph metrics such as average degree, clustering coefficient, density, assortativity, and degree entropy; pairwise similarity across centrality measures (degree, closeness, eigenvector); POS transition entropy and drift across sliding windows.
- **Lexical and syntactic features:** type-token ratio, Yule's K, average word length, character-to-token ratios, punctuation burstiness, POS ratios (pronouns, conjunctions, prepositions), dependency depth, passive constructions, and sentence length variance.
- **Contextual features:** named entity and lemma overlap between adjacent sentences, discourse marker detection, subject continuity, and average similarity of embeddings between neighboring sentences.
- **Deep style indicators:** formality score, clause complexity, modality ratio, discourse marker ratio, rhetorical questions, sarcasm markers, person based pronoun usage, sentiment polarity, subjectivity, and LIWC-inspired affective, authentic, certainty, and tentativeness features.
- **Readability and rhythm:** Flesch Reading Ease, Gunning Fog Index, Automated Readability Index, syllable count, and Dale-Chall readability score.

In selecting these features, we aimed to capture diverse aspects of writing style using only the information contained within each document. Since the task is intrinsic and sentence based, external resources such as author profiles or reference texts are not available. All features are therefore designed to detect stylistic change through local linguistic and structural signals.

To guide our design choices, we also conducted manual inspection of the dataset. This close reading helped us better understand the structure and surface properties of the texts across different difficulty levels. In the medium subset, for example, we frequently encountered recurring messages such as:

- *"In general, be courteous to others. Debate/discuss/argue the merits of ideas, don't attack people. Personal insults, shill or troll accusations, hate speech, any suggestion or support of harm, violence, or death, and other rule violations can result in a permanent ban. For those who have questions regarding any media outlets being posted on this subreddit, please click to review our details as to our approved domains list and outlet criteria."*
- *"r/politics is currently accepting new moderator applications. If you want to help make this community a better place, consider !"*
- *"I am a bot, and this action was performed automatically. Please if you have any questions or concerns."*

These recurring fragments often appeared at stylistic boundaries and provided useful cues for feature selection. They also reinforced the need for features that could capture repetition, discourse markers, and formal structure, particularly in the medium setting where topic variation is limited but not fully absent.

To address these aspects, we constructed feature sets that describe lexical richness, punctuation usage, and syntactic preferences. Word adjacency networks (WANs) were used to extract structural features that capture not only the arrangement of words within sentences, but also patterns of grammatical dependencies, local syntactic structure, and characteristic usage preferences that vary between authors,

as demonstrated in prior work on WAN-based authorship attribution [12]. Contextual features evaluate coherence between neighboring sentences, including named entity overlap, subject continuity, and transitions in discourse markers, drawing on strategies effective in stylometric analysis [15]. Deeper stylistic indicators capture tone, formality, and rhetorical patterns, building on prior work that leverages sentiment analysis and LIWC derived features for author profiling and stylistic characterization [16, 17]. Readability and rhythm metrics provide insight into variation in fluency and pacing across authors. To complement these handcrafted features, we also integrated semantic similarity scores derived from SBERT [18] embeddings. Although these embeddings encode both content and style, we found them to be highly effective across all difficulty levels. The ability of SBERT to capture fine grained variation at the sentence level contributed significantly to the overall performance of the system.

By combining quantitative modeling with manual insight, we constructed a feature set that is both comprehensive and tailored to the stylistic patterns observed in the data. This helped the system generalize more effectively across all difficulty levels.

3.4. Classification

For each difficulty level, we train a separate Gradient Boosting classifier [19] using the handcrafted feature set described above. A distinct model is trained for the easy, medium, and hard subsets, allowing us to tailor preprocessing and calibration to the specific characteristics of each case.

Instead of predicting class labels directly, we use the probability scores returned by the classifier. This allows us to apply a custom threshold when converting probabilities into binary predictions. A fixed threshold of 0.35 was selected based on validation performance. This value provided a better balance between false positives and false negatives compared to the default threshold of 0.5, particularly in the hard and medium settings where stylistic changes tend to be subtle and less frequent.

We also adjust preprocessing based on difficulty level. In the easy and medium subsets, we apply min-max scaling to normalize the feature values and ensure consistency across features with different numeric ranges. In the hard setting, we do not apply scaling. Preserving the original feature distributions helped retain stylistic variation that may be weakened through normalization. This proved effective in settings where topic cues are absent and subtle style differences are the only available signal.

4. Results

We conducted extensive experiments to compare multiple classification models, including Support Vector Machines, Random Forest, Naive Bayes, K-Nearest Neighbors, and Gradient Boosting. Among these, Gradient Boosting consistently achieved the highest F1 scores on the validation set across all difficulty levels. It also demonstrated more stable performance in handling subtle stylistic shifts, especially in the hard setting. Based on these observations, we selected Gradient Boosting as the final classifier for our pipeline.

Table 1 presents the F1 scores obtained by the Gradient Boosting classifier on the validation set across all sixteen WAN configurations. Each row corresponds to a specific WAN configuration, expressed as a 4-bit binary mask in the order: punctuation, stopwords, lemmatization, and POS tags. A bit value of 1 indicates that the corresponding preprocessing step is applied. This means that punctuation is removed, stopwords are removed, words are reduced to their base forms through lemmatization, and part-of-speech transitions are included in the Word Adjacency Network. For example, the configuration 1110 applies all steps except POS tag modeling. This representation highlights how performance varies across different preprocessing combinations. The columns represent the three sub-tasks from the PAN 2025 Style Change Detection challenge: Task 1 (easy), Task 2 (medium), and Task 3 (hard), which differ in the amount of topical variation present in the documents.

The best performing configuration for each task is highlighted in bold. In the easy setting (Task 1), multiple WAN configurations achieve nearly identical F1 scores, all within a narrow range around 0.963. This suggests that in the presence of strong topical variation, the specific choice of preprocessing has a limited effect on overall performance. In contrast, performance in the medium and hard settings is more

Table 1
F1 scores for Gradient Boosting on the validation set

WAN Config	Task 1	Task 2	Task 3
1111	0.962	0.795	0.779
1101	**0.963**	0.797	0.780
1001	**0.963**	0.798	0.781
1011	0.961	0.799	0.781
0111	0.962	0.794	0.778
0101	0.962	0.800	0.782
0011	0.962	0.795	0.780
0001	0.962	0.794	0.782
1110	**0.963**	0.798	**0.785**
1100	0.962	0.796	0.784
1000	0.962	0.794	0.778
1010	**0.963**	0.798	0.782
0110	**0.963**	0.797	0.783
0100	0.962	**0.801**	0.784
0010	0.962	0.796	0.781
0000	0.962	0.798	0.781

sensitive to preprocessing. Configuration 0100 (only stopwords removed, no lemmatization, no POS tag modeling) yields the best result in Task 2, while configuration 1110 (all steps except POS modeling) achieves the highest F1 score in Task 3.

A clear pattern in the results is the benefit of stopword removal, which is included in almost all of the best performing configurations. Lemmatization has mixed effects and is often missing from the top setups. Punctuation removal helps in the medium task but seems less important in the hard setting. POS tag modeling is often left out of the best configurations, especially for the hard task. This may be because adding grammatical information makes the networks less sensitive to the specific ways authors use language. Overall, using simpler and more selective preprocessing seems to work better when topic cues are missing and the system needs to rely more on subtle differences in writing style.

Based on these validation scores, the best performing models using configuration 0110 for the easy task, 0100 for the medium task, and 1110 for the hard task were selected and submitted to the TIRA platform [20] for final evaluation in the PAN 2025 shared task.

Table 2 presents the final F1 scores of our system on the official PAN 2025 test set. Our submission was made under the TIRA team name *stylospies*. Our approach consistently outperforms the baseline across all three tasks.

Table 2
F1 scores on the PAN 2025 test set

Model	Task 1	Task 2	Task 3
Our approach	0.959	0.786	0.791
Baseline	0.439	0.440	0.453

The largest improvement is observed in Task 1, where the model achieves an F1 score of 0.959 compared to the baseline score of 0.439. Strong results are also maintained in the more difficult settings, demonstrating the robustness of our method under varying levels of topical variation.

5. Conclusion

This work explored a primarily classical approach to the challenging task of sentence level style change detection. Rather than relying fully on large pretrained language models, we focused on handcrafted features grounded in linguistic structure, lexical patterns, and graph based representations.

We constructed Word Adjacency Networks under sixteen preprocessing configurations and extracted a broad range of linguistic, structural, contextual, and embedding based features. Separate Gradient Boosting classifiers were trained for each difficulty level, using tailored preprocessing and thresholding to account for topical variability across the dataset. The results confirm that even in a high variability and fine grained task such as this one, traditional machine learning methods, when carefully engineered and calibrated, remain competitive. This supports the continued relevance of lightweight and transparent models, especially in scenarios where resource constraints or interpretability are important.

As natural language processing advances, it is increasingly important to consider not only accuracy but also the environmental and practical costs of AI systems. By emphasizing simplicity, transparency, and efficiency, our work contributes to the broader effort toward sustainable AI. Future directions could explore hybrid methods that combine the clarity of handcrafted features with the adaptability of neural representations to further improve performance in stylistically complex tasks.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT to check grammar and spelling. The authors reviewed and edited the content afterward and take full responsibility for the final publication.

References

[1] R. Sarwar, N. Urailertprasert, N. Vannaboot, C. Yu, T. Rakthanmanon, E. Chuangsuwanich, S. Nutanong, *cag*: Stylometric authorship attribution of multi-author documents using a co-authorship graph, IEEE Access 8 (2020) 18374–18393.

[2] A. Saini, M. R. Sri, M. Thakur, Intrinsic plagiarism detection system using stylometric features and dbscan, in: 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), IEEE, 2021, pp. 13–18.

[3] J. Bevendorff, D. Dementieva, M. Fröbe, B. Gipp, A. Greiner-Petter, J. Karlgren, M. Mayerl, P. Nakov, A. Panchenko, M. Potthast, A. Shelmanov, E. Stamatatos, B. Stein, Y. Wang, M. Wiegmann, E. Zangerle, Overview of PAN 2025: Voight-Kampff Generative AI Detection, Multilingual Text Detoxification, Multi-Author Writing Style Analysis, and Generative Plagiarism Detection, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[4] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the Multi-Author Writing Style Analysis Task at PAN 2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[5] A. A. Ayele, N. Babakov, J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, et al., Overview of pan 2024: Multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative ai authorship verification condensed lab overview, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024, pp. 231–259.

[6] J. Bevendorff, I. Borrego-Obrador, M. Chinea-Ríos, M. Franco-Salvador, M. Fröbe, A. Heini, K. Kredens, M. Mayerl, P. Pęzik, M. Potthast, et al., Overview of pan 2023: Authorship verification, multi-author writing style analysis, profiling cryptocurrency influencers, and trigger detection: Condensed lab overview, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023, pp. 459–481.

[7] J. Devlin, Bert: Pre-training of deep bidirectional transformers for language understanding/arxiv preprint, arXiv preprint arXiv:1810.04805 (2018).

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[9] T. Lin, Y. Wu, L. Lee, Team nycu-nlp at pan 2024: integrating transformers with similarity adjustments for multi-author writing style analysis, Working Notes of CLEF (2024).
[10] J. Lv, Y. Yi, H. Qi, Team fosu-stu at pan: Supervised fine-tuning of large language models for multi author writing style analysis, Working Notes of CLEF (2024).
[11] E. Księżniak, K. Węcel, M. Sawiński, Team openfact at pan 2024: Fine-tuning bert models with stylometric enhancements, in: CEUR Workshop Proceedings, volume 3740, 2024.
[12] E. Castillo, O. Cervantes, D. Vilarino, Authorship verification using a graph knowledge discovery approach, Journal of Intelligent & Fuzzy Systems 36 (2019) 6075–6087.
[13] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Pan25 multi-author writing style analysis, 2025. URL: https://doi.org/10.5281/zenodo.15053260. doi:10.5281/zenodo.15053260.
[14] M. Honnibal, spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, (No Title) (2017).
[15] V. W. Feng, G. Hirst, Patterns of local discourse coherence as a feature for authorship attribution, Literary and Linguistic Computing 29 (2014) 191–198.
[16] J. Gaston, M. Narayanan, G. Dozier, D. L. Cothran, C. Arms-Chavez, M. Rossi, M. C. King, J. Xu, Authorship attribution via evolutionary hybridization of sentiment analysis, liwc, and topic modeling features, in: 2018 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2018, pp. 933–940.
[17] G. A. Katsios, N. Sa, T. Strzalkowski, Figuratively speaking: Authorship attribution via multi-task figurative language modeling, arXiv preprint arXiv:2406.08218 (2024).
[18] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).
[19] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics (2001) 1189–1232.
[20] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241.