Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies

Ing. Juraj Petrík

Dissertation Thesis Abstract

Recognition of the similarity of texts, programing code

to obtain the Academic Title of philosophiae doctor“, abbreviated as „PhD.“

In the doctorate degree study programme: Applied Informatics (2511V00)

In the field of study: Computer science

Form of Study: full-time

Bratislava, 2025

Dissertation Thesis has been prepared at Institute of Informatics, Information Systems and Software Engineering, Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava

Submitter: Ing. Juraj Petrík
Institute of Informatics, Information Systems and Software Engineering,
Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava
Ilkovičova 2, 842 16 Bratislava

Supervisor: doc. Mgr. Michal Kováč, MSc., PhD.
Institute of Informatics, Information Systems and Software Engineering,
Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava
Ilkovičova 2, 842 16 Bratislava

Dissertation Thesis Abstract was sent: ....

Dissertation Thesis Defence will be held on …………. *at* ………. *at* Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava

....
prof. Ing. Ivan Kotuliak, PhD.
Dean FIIT STU Bratislava

STU

Abstrakt

Táto dizertačná práca je štúdiou výpočtových metód pre autorskú atribúciu a stylometrickú analýzu so zameraním na rozpoznávanie podobnosti textov a analýzu zdrojového kódu.
Práca ukazuje, ako využiť štandardné unixové nástroje na spracovanie textu na odhalenie plagiátorstva v zdrojovom kóde bez potreby poznať programovací jazyk. Tento prístup fungoval lepšie než známe nástroje ako MOSS, JPlag a SIM. Po druhé, predstavuje hierarchickú architektúru neurónových sietí, ktorá kombinuje konvolučné a rekurentné vrstvy a dosahuje 97,5% presnosť pri určovaní autorstva zdrojového kódu na referenčných datasetoch. Po tretie, je to prvá systematická štúdia efektov časového posunu (angl. temporal drift) v stylometrii zdrojového kódu. Ukazuje, že presnosť v čase výrazne klesá a že existujú zaujímavé asymetrické časové vzory.
Štúdia presahuje rámec analýzy zdrojového kódu a zahŕňa profilovanie obsahu sociálnych médií, analýzu politického diskurzu a autorskú atribúciu naprieč jazykmi. Práca dokazuje, že vyvinuté metódy sú užitočné v mnohých jazykoch a oblastiach účasťou na zdieľaných úlohách PAN@CLEF. Porovnanie tradičných metód strojového učenia a hlbokého učenia ukazuje, že modely hlbokého učenia zvyčajne dosahujú vyššiu presnosť, ale tradičné metódy zostávajú konkurencieschopné vďaka zvyčajne lepšej interpretovateľnosti a menším výpočtovým nárokom.
Dizertácia tiež prispieva výskumnej komunite vytvorením a zverejnením štandardizovaných datasetov (Google Code Jam a Codeforces), ktoré sa často používané ako referenčné datasety v stylometrii zdrojového kódu.

Abstract

This dissertation is a thorough study of computational methods for authorship attribution and stylometric analysis, with a focus on recognizing text similarity and analyzing source code.

The thesis shows how to use standard Unix text processing tools to find plagiarism in source code without having to know the programming language. This worked better than well-known tools like MOSS, JPlag, and SIM. Second, it shows a hierarchical neural network architecture that combines convolutional and recurrent layers and gets 97.5% accuracy in figuring out who wrote the source code on benchmark datasets. Third, it is the first systematic study of the effects of temporal drift in source code stylometry. It shows that accuracy drops significantly over time and shows interesting asymmetric temporal patterns.

The study goes beyond looking at source code to include profiling social media content, analyzing political discourse, and attributing authorship across languages. The work shows that the methods developed are useful in many languages and fields by taking part in PAN@CLEF shared tasks. A systematic comparison of traditional machine learning and deep learning methods shows that deep learning models usually have higher accuracy, but traditional methods are still competitive because usually they are easier to understand and faster to compute.

The dissertation also helps the research community by creating and making public standardized datasets (Google Code Jam and Codeforces) that are now commonly used as benchmarks in source code stylometry.

Introduction and Motivation

The digital transformation of human communication has fundamentally altered the landscape of textual analysis, introducing unprecedented challenges in verifying the authenticity and authorship of digital content. The exponential growth of online platforms, social media networks, and particularly the recent emergence of sophisticated large language models capable of producing human-indistinguishable text has dramatically elevated the importance of stylometry and authorship attribution across multiple domains.

The proliferation of digital content has created new opportunities for both legitimate research and malicious activities. In academic environments, the rise of sophisticated plagiarism techniques and AI-generated content poses significant threats to academic integrity. According to research conducted at Slovak University of Technology in Bratislava in 2010, 33% of students admitted to plagiarizing academic work, while 63% acknowledged providing their work to others for plagiaristic purposes. These statistics underscore the critical need for advanced detection methods that can identify not only direct copying but also sophisticated obfuscation techniques.

In cybersecurity and forensic applications, the ability to identify authors of malicious communications has become crucial for national security and public safety operations. Law enforcement agencies increasingly rely on stylometric techniques to identify threat actors, analyze extremist communications, and attribute cybercrimes to specific individuals or groups. The development of anonymous communication platforms has further intensified the demand for reliable identity attribution systems within cybersecurity frameworks.

The commercial sector faces similar challenges, particularly in intellectual property protection and software patent disputes involving astronomical financial stakes. Events such as the Google vs. Oracle case have demonstrated that software plagiarism extends far beyond academic settings, highlighting the need for robust detection systems capable of identifying stylistic patterns even when code has been significantly modified or obfuscated.

Stylometry, defined as the quantitative analysis of literary style using computational methods, operates on the fundamental premise that authors demonstrate consistent and recognizable writing patterns that can be identified as unique "stylistic fingerprints." These patterns manifest across multiple linguistic dimensions, encompassing character-level statistics, lexical preferences, complex syntactic structures, semantic patterns, and discourse-level features. The field has undergone remarkable evolution since its inception in the late 19th century, when Thomas Mendenhall [10] pioneered the use of statistical analysis of word length distributions to address the famous authorship disputes surrounding Shakespeare's works.

The foundational assumption of stylometry—that authors exhibit quantifiable stylistic patterns—has been validated through numerous studies over more than a century. Early 20th-century contributions from statisticians like George Udny Yule, who developed measures of vocabulary richness and examined relationships between text length and lexical diversity, laid important groundwork. The mid-century breakthrough came with Mosteller and Wallace's seminal analysis [11] of the Federalist Papers, which employed advanced statistical methods to resolve long-standing authorship controversies while establishing rigorous methodological standards that continue to influence stylometric research today.

The computational revolution of the late 20th century transformed stylometric capabilities, enabling analysis of larger corpora and implementation of more sophisticated statistical models. Burrows' introduction of the Delta measure represented a significant methodological advancement, providing a standardized approach for quantifying stylistic distance that gained widespread acceptance in literary scholarship and became the de facto standard for authorship attribution tasks [16].

Modern stylometry encompasses five distinct but interconnected subtasks [12]:

- **Authorship Attribution:** The process of identifying the specific author of a questioned document from a predefined group of candidate authors. This represents the most widely studied area of stylometry due to its broad applicability in forensic, academic, and commercial contexts.
- **Authorship Verification:** A binary classification task that determines whether a questioned document was produced by a particular candidate author. This approach is particularly valuable in scenarios where the authorship question involves a specific individual rather than selection from multiple candidates.
- **Authorship Profiling:** The derivation of demographic, psychological, and behavioral characteristics of authors based on their writing patterns. This includes determination of age, gender, education level, cultural background, and personality traits.
- **Stylochronometry:** The investigation of temporal variations in authorial style, examining how writing patterns evolve over time due to factors such as aging, education, professional development, and changing life circumstances.
- **Adversarial Stylometry:** The study of methods for deliberately obscuring authorial identity or imitating other authors' styles. This includes both defensive techniques for hiding one's identity and offensive methods for impersonating others.

The interconnected nature of these subtasks creates a rich research landscape where advances in one area often inform progress in others. This dissertation addresses challenges across multiple subtasks, examining diverse data types ranging from structured source code to informal social media communications, thereby contributing to our understanding of stylometric techniques across various textual domains.

Dissertation Objectives and Research Questions

This dissertation addresses fundamental challenges in computational stylometry through a systematic investigation of diverse textual domains and methodological approaches. The research is driven by both theoretical curiosity and practical necessity, as evidenced by the increasing importance of authorship attribution in our digitally connected world.

The primary research objectives are formulated as interconnected questions that reflect current paradigms and emerging challenges in computational stylometry:

Research Question 1: Enhancement of Plagiarism Detection Systems How can conventional plagiarism detection methods be enhanced to recognize sophisticated obfuscation techniques that defeat current detection systems? This question emerged from observations that traditional similarity-based detection tools struggle with advanced obfuscation strategies employed by experienced plagiarizers.

The research examines a comprehensive five-level obfuscation hierarchy:

- **Level 1:** Basic copy-paste operations with minimal modification
- **Level 2:** Simple textual transformations including whitespace manipulation, case changes, and comment modifications
- **Level 3:** Structural obfuscation including variable renaming and basic algorithmic restructuring
- **Level 4:** Advanced syntactic transformations including control flow modifications and statement reordering
- **Level 5:** Expert-level obfuscation involving fundamental algorithmic changes, redundant code injection, and semantic preservation with syntactic transformation

Research Question 2: Traditional vs. Deep Learning Paradigms What is the comparative effectiveness of traditional machine learning methods versus modern deep learning architectures for authorship attribution tasks across different textual domains and dataset sizes?

This investigation encompasses:

- Traditional approaches using hand-crafted features (TF-IDF, n-grams, syntactic patterns)
- Classical machine learning classifiers (SVM, Random Forest, Decision Trees)
- Modern deep learning architectures (CNN-RNN combinations, Transformer models including BERT)
- Hybrid approaches combining traditional feature engineering with neural architectures
- Performance evaluation across varying author set sizes (3 to 110 authors)
- Computational efficiency and interpretability trade-offs

Research Question 3: Temporal Dynamics in Stylometry What is the impact of temporal drift on authorship attribution accuracy, and what strategies can mitigate the degradation of performance over time?

This represents one of the first systematic investigations of temporal effects in source code stylometry, examining:

- Accuracy degradation patterns over periods ranging from one to twelve years
- Asymmetric temporal drift phenomena (training on future data vs. past data)
- Programming language-specific temporal effects (C++, Java, Python)
- Mitigation strategies for temporal robustness
- Stylochronometric modeling of authorial evolution

Research Question 4: Feature Engineering and Representation Learning What are the most effective methods for feature extraction and representation across different categories of digital texts, and how do domain-specific characteristics influence optimal feature selection?

This encompasses analysis of:

- **Source code features:** Abstract Syntax Tree (AST) representations, lexical patterns, structural complexity metrics
- **Social media features:** Character-level patterns, emoji processing, hashtag normalization, URL tokenization
- **Short text features:** N-gram selection strategies (global vs. local), frequency-based vs. semantic features
- **Cross-domain features:** Language-agnostic representations, domain adaptation techniques

Research Question 5: Multilingual and Cross-Cultural Adaptation How can authorship attribution methods be modified for effective deployment in multilingual and cross-cultural contexts, particularly in environments with limited language-specific resources?

This investigation addresses:

- Language-independent feature representations
- Cross-lingual transfer learning approaches
- Cultural bias mitigation in stylometric features
- Performance evaluation across English and Spanish datasets
- Preprocessing pipelines for non-standard orthography and multilingual content

Methodological Integration and Contribution Framework

Each research question is systematically addressed through multiple publications that collectively form a comprehensive investigation of computational stylometry. The methodological approach emphasizes:

1. **Empirical Validation:** All proposed methods undergo rigorous testing on established benchmarks and novel datasets, with performance compared against state-of-the-art approaches.
2. **Reproducible Research:** Creation and public release of standardized datasets that have become community benchmarks, facilitating comparative evaluation and advancing the field.
3. **Practical Applicability:** Focus on real-world deployment scenarios, including computational efficiency considerations and interpretability requirements for forensic applications.
4. **Theoretical Advancement:** Systematic investigation of fundamental assumptions in stylometry, particularly the temporal stability of authorial style and the effectiveness of different feature representations.

The interconnected nature of these research questions reflects the multifaceted challenges in modern stylometry, where advances in one area often inform progress in others. For instance, insights from temporal drift analysis inform robust feature selection strategies, while cross-lingual investigations contribute to our understanding of universal stylistic patterns that transcend language-specific characteristics.

Overview of Current State and Contemporary Challenges

Historical Development and Evolution of Stylometry

Stylometry has undergone profound transformation since its inception, evolving from basic statistical approaches to sophisticated computational methodologies that leverage cutting-edge machine learning and artificial intelligence techniques. This evolution reflects not only technological advancement but also deeper understanding of the cognitive and linguistic processes underlying human writing behavior.

Early Foundations (1880s-1950s)

The discipline emerged in the late 19th century with Thomas Mendenhall's groundbreaking work, which represented the first systematic attempt to apply quantitative methods to authorship questions. Mendenhall's analysis of word length distributions in Shakespearean texts established the foundational principle that authors exhibit measurable and consistent stylistic patterns. His methodology, while primitive by contemporary standards, demonstrated remarkable prescience in recognizing that statistical analysis could reveal authorial fingerprints invisible to traditional literary analysis.

The early 20th century witnessed significant methodological refinements through the contributions of pioneering statisticians. George Udny Yule's development of vocabulary richness measures and his investigation of relationships between text length and lexical diversity established important theoretical frameworks that continue to influence modern stylometric research. These early scholars recognized that stylistic analysis required sophisticated statistical tools capable of capturing subtle but consistent patterns in linguistic behavior.

Mid-Century Breakthrough (1950s-1980s)

The field achieved scientific maturity with Mosteller and Wallace's landmark study [10] of the Federalist Papers, which established rigorous methodological standards that transformed stylometry from an art into a science. Their work demonstrated that careful statistical analysis could resolve long-standing authorship controversies with high confidence levels, providing a model for subsequent research in the field.

This period also saw the development of the theoretical foundations for computational stylometry. The recognition that function words—articles, prepositions, conjunctions—carry particularly strong authorial signals led to the development of feature selection strategies that remain influential today. The insight that these seemingly insignificant words reflect unconscious linguistic habits proved crucial for later developments in automated authorship attribution.

Computational Revolution (1980s-2000s)

The advent of computational technology fundamentally transformed stylometric capabilities, enabling analysis of vastly larger corpora and implementation of more sophisticated analytical methods. John Burrows' introduction of the Delta measure in the early 2000s represented a watershed moment, providing a standardized approach for quantifying stylistic distance that gained widespread acceptance across multiple disciplines.

The Delta measure's success stemmed from its elegant simplicity combined with robust performance across diverse texts and languages. By focusing on the most frequent words and calculating standardized distances between their usage patterns, Delta captured fundamental stylistic differences while remaining computationally tractable and theoretically interpretable.

This period also witnessed the emergence of machine learning applications in stylometry. The development of Support Vector Machines (SVMs), decision trees, and ensemble methods provided powerful new tools for classification tasks, enabling researchers to move beyond simple statistical measures to sophisticated pattern recognition approaches.

Modern Era: Deep Learning and Transformer Models (2000s-Present)

The most recent paradigm shift has been driven by advances in deep learning and natural language processing. The development of word embeddings, beginning with approaches like Word2Vec and GloVe, revolutionized text representation by capturing semantic relationships between words in high-dimensional vector spaces.

The introduction of contextualized representations through models like ELMo marked another significant advance, enabling capture of word meanings that vary based on context. However, the most transformative development has been the emergence of Transformer-based models, particularly BERT and its variants, which have achieved state-of-the-art performance across numerous natural language processing tasks.

Recent studies demonstrate that fine-tuned BERT models can achieve exceptional performance in authorship attribution tasks, often surpassing traditional feature-based methods by substantial margins. However, these advances come with significant computational costs and interpretability challenges that limit their applicability in certain domains.

Modern Methodological Frameworks

Contemporary stylometry can be systematically categorized according to multiple taxonomic frameworks that reflect different aspects of methodological approach, feature representation, and analytical objectives.

Feature-Based Categorization

The landscape of feature extraction and representation defines the primary axis along which modern stylometric techniques are differentiated:

Lexical Features encompass vocabulary-based approaches including word frequency distributions, lexical richness measures, and word usage patterns. These features capture an author's vocabulary preferences and have proven particularly effective for longer texts where sufficient lexical variety is available. Advanced lexical features include measures of semantic similarity, word embeddings, and topic modeling approaches that capture meaning-related aspects of vocabulary usage.

Syntactic Features analyze part-of-speech patterns, phrase structure distributions, and grammatical complexity measures. These features have proven especially valuable for cross-domain authorship attribution since they demonstrate less sensitivity to topic variations compared to lexical features. Recent advances in syntactic parsing have enabled more sophisticated analysis of hierarchical structure patterns and grammatical relationships, opening new avenues for capturing authorial preferences in sentence construction and discourse organization.

Character-Level Features include n-gram distributions, punctuation patterns, and character-based statistics that exhibit high consistency across languages and contexts. Character-level approaches have demonstrated remarkable effectiveness in multilingual authorship attribution and in handling texts with significant noise or non-standard orthography. The language-agnostic nature of character-level features makes them particularly valuable for cross-lingual applications and analysis of informal communications.

Semantic Features represent the newest frontier in stylometric feature engineering, incorporating topic modeling, word embeddings, and contextualized representations that capture meaning-related dimensions of writing style. These approaches show particular promise for scenarios where authors discuss similar topics while maintaining distinctive stylistic signatures.

Structural Features encompass document-level organizational patterns, paragraph structure, and discourse markers that reflect authors' preferences for text organization and information presentation. In specialized domains like source code analysis, structural features might include Abstract Syntax Tree patterns, control flow characteristics, and architectural design preferences.

Computational Paradigms

The methodological landscape of modern stylometry can be partitioned into several primary computational paradigms:

Traditional Machine Learning Approaches continue to demonstrate robust baseline performance across diverse datasets and applications. Methods such as Support Vector Machines, Random Forests, and Naive Bayes classifiers offer interpretable models with efficient computational requirements. These approaches remain particularly valuable in scenarios where interpretability is crucial, such as forensic applications where decision rationale must be clearly explainable.

Deep Learning Architectures have achieved breakthrough performance in numerous stylometric applications. Convolutional Neural Networks effectively capture local patterns in text, while Recurrent Neural Networks excel at modeling sequential dependencies in writing style. The introduction of attention mechanisms has enabled models to focus on the most relevant stylistic patterns, improving both accuracy and interpretability.

Transformer-Based Models represent the current state-of-the-art for many stylometric tasks, with models like BERT, RoBERTa, and their variants achieving unprecedented performance levels. These models leverage massive pre-training corpora to develop sophisticated representations of linguistic patterns, though they require substantial computational resources and present significant interpretability challenges.

Hybrid Approaches that combine traditional feature engineering with modern neural architectures have shown considerable promise for practical applications. These methods benefit from the representational learning capabilities of deep learning while maintaining the interpretability and domain knowledge incorporation capabilities of traditional approaches.

Contemporary Challenges and Limitations

Despite remarkable advances in computational capability and methodological sophistication, the field continues to confront several fundamental challenges that constrain both practical applications and theoretical understanding.

Scalability and Open-Set Attribution

The scalability challenge represents one of the most significant barriers to practical deployment of stylometric systems. While most published research focuses on closed-set scenarios with limited numbers of authors (typically fewer than 100), real-world applications often require attribution among thousands or tens of thousands of potential authors.

The open-set authorship attribution problem—characterized by the possibility that test documents are authored by individuals not represented in the training dataset—presents particularly acute challenges. Traditional classification techniques frequently exhibit substantial inaccuracies in open-set scenarios since they are designed to assign every input to one of the training classes regardless of actual authorship. Recent research has begun addressing this challenge through confidence calibration, outlier detection, and one-class classification approaches, but significant gaps remain.

Short Text Attribution Challenges

The proliferation of social media platforms and microblogging services has created unprecedented demand for attribution of short texts, yet these brief communications present fundamental challenges for traditional stylometric approaches. Social media posts, text messages, and comments contain limited stylistic information, making reliable attribution extremely difficult with conventional methods.

The challenge extends beyond simple information scarcity to include the informal register typical of social media communications, frequent use of non-standard orthography, multimedia integration, and platform-specific conventions. Recent research has explored specialized approaches including aggregation of multiple short texts, development of features optimized for informal communication, and adaptation of deep learning models for short text analysis. However, performance on short texts generally remains substantially lower than on longer documents, limiting practical applications in many important domains.

Temporal Drift and Style Evolution

The assumption of temporal stability in authorial style—implicit in most stylometric research—is increasingly challenged by empirical findings demonstrating significant temporal drift across various domains. Authors' writing styles evolve due to factors including aging, education, professional development, changing life circumstances, and exposure to new linguistic influences.

This temporal instability poses serious challenges for practical stylometric systems, particularly in forensic applications where training and test data may be separated by substantial time periods. The investigation of temporal effects through time-aware models and stylochronometric analysis represents an emerging research area, though progress has been constrained by the scarcity of longitudinal datasets spanning significant time periods.

The challenge is compounded by the fact that different stylistic features exhibit varying degrees of temporal stability. While some characteristics like function word usage patterns may remain relatively stable, others such as vocabulary preferences and topic interests may change rapidly. Understanding these differential temporal dynamics is crucial for developing robust attribution systems.

Cross-Domain and Cross-Linguistic Robustness

The performance of stylometric systems across different domains (genres, topics, registers) and languages presents ongoing challenges that limit the generalizability of research findings. Most studies focus on monolingual, single-domain scenarios, while real-world applications frequently require attribution across diverse contexts.

Cross-domain attribution difficulties arise from the fact that variations in topic and genre can mask authorial stylistic signals. An author's writing style may appear substantially different when composing academic papers versus social media posts, or when discussing familiar versus unfamiliar topics. This domain sensitivity requires careful consideration of feature selection and model design to ensure that systems capture authorial rather than topical signals.

Cross-linguistic attribution introduces additional challenges related to different writing systems, grammatical structures, and cultural conventions. While some research has explored multilingual stylometry, most work continues to focus exclusively on English texts, thereby limiting the global applicability of stylometric methods. The development of language-agnostic features and cross-lingual transfer learning approaches represents an important frontier for expanding stylometric capabilities.

Adversarial Attacks and Security Vulnerabilities

The vulnerability of stylometric systems to adversarial attacks and intentional style obfuscation represents a critical concern for security-sensitive applications. Adversarial stylometry encompasses techniques for deliberately modifying writing style to evade detection by attribution systems while preserving semantic content and readability.

Recent research has demonstrated that sophisticated adversarial attacks can significantly compromise the effectiveness of attribution systems, raising questions about the reliability of stylometric evidence in high-stakes applications. The arms race between increasingly sophisticated attack methods and defensive countermeasures continues to evolve, with implications for the practical deployment of stylometric systems in forensic and security contexts.

The challenge is exacerbated by the relative ease with which many stylometric features can be manipulated through conscious effort or automated tools. Simple modifications such as synonym substitution, sentence restructuring, or punctuation alteration can substantially degrade attribution performance, while more sophisticated attacks can virtually eliminate authorial signals while maintaining text quality and readability.

Research Methodology

The research in this dissertation is based on comparative experimental methodology that systematically evaluates different computational approaches.

Data and Processing

Various datasets were used for experiments:

- **Programming Contest Datasets:** Solutions from **Google Code Jam (2009-2020)** and **Codeforces** were used, providing data with temporal information.
- **Social Media Datasets:** Within participation in **PAN@CLEF** competitions, datasets focused on profiling bots, gender, celebrities, and political discourse were analyzed.
- **Academic Data:** Anonymized student programming assignments from FIIT STU were used for plagiarism detection experiments.

Each type of data required a specific preprocessing pipeline, from language-agnostic tokenization for source codes to normalization of non-standard language in social media texts.

Feature Extraction and Modeling

The work implemented and compared several strategies and models:

- **Feature Extraction:** Statistical (n-grams, TF-IDF), structural (derived from AST), stylistic (punctuation, sentence length), and semantic features were used.
- **Modeling Approaches:** Traditional classifiers (SVM, Random Forest), an approach using **Unix tool chaining**, and deep learning models (**hierarchical CNN-RNN networks**, **transformers like BERT**) were compared.

Validation and Interpretation

Cross-validation and multiple training/testing with stability monitoring were used to ensure result robustness. The performance of plagiarism detection methods was compared with established tools (e.g., MOSS, JPlag). Furthermore, some stylometric methods were compared within international competitions (PAN@CLEF). For better understanding of transformer model decisions, interpretability methods like SHAP and LIME were used.

Overview of Publications and Achieved Results

The core of the dissertation consists of nine scientific publications.

Publication 1: Unix-Based Source Code Plagiarism Detection

- **Contribution:** The work presented a language-independent method for detecting plagiarism in source code using Unix tools and a five-level obfuscation detection system.
- **Results:** In scenarios with expert-level obfuscation, the method achieved better results compared to MOSS, JPlag, and SIM tools.

Publication 2: Deep Learning for Source Code Attribution

- **Contribution:** The work proposed a hierarchical neural network (CNN-RNN) for source code authorship attribution. During the research, the **Google Code Jam dataset** was created and published.
- **Results:** The architecture achieved high accuracy of **97.5%** on a dataset with 110 authors.

Publication 3: Bot and Gender Detection (@PAN challenge)

- **Contribution:** The hierarchical CNN-RNN architecture was applied to a dual task: distinguishing content generated by humans and bots and determining the author's gender on social media data.
- **Results:** The method achieved **90%** accuracy in bot detection and 77.6% in gender profiling in English.

Publication 4: Celebrity Profiling with TF-IDF (@PAN challenge)

- **Contribution:** The study showed that traditional methods (TF-IDF and Random Forest) can achieve competitive performance with lower computational complexity. The work also focused on addressing imbalanced data using SMOTE technique.
- **Results:** The approach demonstrated that traditional methods achieved comparable performance with contemporary deep learning methods.

Publication 5: Temporal Drift Analysis

- **Contribution:** This is one of the first systematic analyses of temporal effects in source code authorship attribution.
- **Results:** Experiments demonstrated accuracy decline over time (from >90% to <50%) and revealed **asymmetric temporal drift patterns**.

Publication 6: Political Discourse Analysis

- **Contribution:** The research focused on demographic profiling in the context of political discourse on Twitter with emphasis on interpretability.
- **Results:** The method using Random Forest classifier achieved **85%** accuracy in gender classification.

Publication 7: Advanced Source Code Attribution

- **Contribution:** The work compares performance between traditional ML models and modern DL architectures (including BERT) on different source code representations (text vs. AST).
- **Results:** It was shown that while deep learning models achieve higher accuracy, traditional methods are competitive and offer better interpretability.

Publication 8: Local and Global N-gram Feature Analysis

- **Contribution:** The study focused on comparing global and local n-gram feature selection for short text authorship attribution.
- **Results:** It was shown that **local selection of most frequent n-grams** outperforms global methods, bringing up to 2% increase in weighted F1-score.

Publication 9: Interpretable Style Change Detection

- **Contribution:** The work addresses style change detection (SCD) and compares traditional ML models with hand-crafted features with transformer-based models, with emphasis on interpretability using SHAP and LIME.
- **Results:** The analysis shows that while transformer models achieve the highest accuracy, traditional models are competitive and offer the advantage of decision transparency.

Summary and Discussion

The dissertation presents a comprehensive contribution to the field of computational stylometry. During the research, new methodological frameworks were developed, resources were created for the community, and analyses were provided that contributed to both theoretical understanding and practical applications. The work brought language-agnostic approaches to source code analysis, systematically explored temporal drift effects, and compared traditional and modern computational approaches, thereby providing guidance for practical deployment. By creating publicly available datasets, it contributed to research reproducibility in this field.

Current research in the field is taking several directions:

- **Foundation Models:** Large pre-trained language models achieve good results, but their computational complexity and limited interpretability are still challenges.
- **Multimodal and Behavioral Stylometry:** There is growing interest in combining textual analysis with additional signals (visual, temporal, behavioral), which can increase attribution accuracy.

- **Interpretable Stylometry (Explainable Stylometry):** Development of interpretable models is crucial for applications where decisions must be transparent and justifiable, such as in legal and forensic sciences.

References

[1] ALSANOOSY, T. et al. Authorship Attribution for English Short Texts. In *Engineering, Technology & Applied Science Research*. 2024. Vol. 14, no. 5, pp. 16419–16426.

[2] BARLAS, G. - STAMATATOS, E. Cross-domain Authorship Attribution Using Pre-trained Language Models. In *Information Processing & Management*. 2024. Vol. 61, no. 2, pp. 103–127.

[3] BERBER SARDINHA, T. Stylochronometry: A New Approach to Temporal Analysis in Authorship Studies. In *Literary and Linguistic Computing*. 2024. Vol. 39, no. 1, pp. 156–178.

[4] BRENNAN, M. et al. Adversarial Stylometry: Circumventing Authorship Recognition. In *Digital Investigation*. 2023. Vol. 46, pp. 101–115.

[5] BURROWS, J.F. Word-patterns and story-shapes: The statistical analysis of narrative style. In *Literary and Linguistic Computing*. 1987. Vol. 2, no. 2, pp. 61–70.

[6] FABIEN, M. et al. BERTology Meets Stylometry: Transformer Models for Authorship Attribution. In *Natural Language Engineering*. 2023. Vol. 29, no. 3, pp. 584–615.

[7] FRÖHLING, L. - ZUBIAGA, A. Robustness of Authorship Attribution Against Adversarial Attacks. In *Proceedings of the 31st International Conference on Computational Linguistics*. 2023. pp. 2314–2328.

[8] LAGUTINA, K. et al. A Survey on Stylometric Text Features. In *2019 25th Conference of Open Innovations Association (FRUCT)* [online]. Helsinki, Finland: IEEE, 2019. pp. 184–195. [cited 2025-07-03]. Available on internet: https://ieeexplore.ieee.org/document/8981504/.

[9] LIAO, S. et al. Temporal Stylometry: Modeling Writing Style Evolution. In *Computational Linguistics*. 2023. Vol. 49, no. 3, pp. 645–672.

[10] MENDENHALL, T.C. The Characteristic Curves of Composition. In *Science*. 1887. Vol. 9, no. 214, pp. 237–249.

[11] MOSTELLER, F. - WALLACE, D.L. *Inference and Disputed Authorship: The Federalist*. [s.l.]: Addison-Wesley, 1964.

[12] NEAL, T. et al. Surveying Stylometry Techniques and Applications. In *ACM Computing Surveys*. 2023. Vol. 56, no. 1, pp. 1–36.

Publications and citations

J. Petrik, D. Chudá, B. Steinmuller, "Source code plagiarism detection: The Unix way," in *2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 2017, pp. 000467–000472.

Karnalim, O. and Chivers, W., 2019, December. Similarity detection techniques for academic source code plagiarism and collusion: a review. In *2019 IEEE international conference on engineering, technology and education (TALE)* (pp. 1-8). IEEE.

Aniceto, R.C., Holanda, M., Castanho, C. and Da Silva, D., 2021, October. Source code plagiarism detection in an educational context: A literature mapping. In *2021 IEEE Frontiers in Education Conference (FIE)* (pp. 1-9). IEEE.

Karnalim, O. and Kurniawati, G., 2020. Programming style on source code plagiarism and collusion detection. *International Journal of Computing*, *19*(1), pp.27-38.

Svajlenko, J. and Roy, C.K., 2020. A survey on the evaluation of clone detection performance and benchmarking. *arXiv preprint arXiv:2006.15682*.

Bhanuse, R., Bawankar, U., Sharma, D.M., Patle, M., Narsapurkar, V. and Sawate, S., 2022, March. A coding platform: For all programming labs and practical examination. In *AIP conference proceedings* (Vol. 2424, No. 1, p. 060001). AIP Publishing LLC.

Jain, A.D., Gupta, A., Choudhary, D., Nayan and Tiwari, A., 2022. A comprehensive source code plagiarism detection software. In *Innovations in Computational Intelligence and Computer Vision: Proceedings of ICICV 2021* (pp. 343-350). Singapore: Springer Nature Singapore.

Manahi, M., Sulaiman, S. and Bakar, N.S.A.A., 2022, March. Source code plagiarism detection using Siamese BLSTM network and embedding models. In *Proceedings of the 8th International Conference on Computational Science and Technology: ICCST 2021, Labuan, Malaysia, 28–29 August*(pp. 397-409). Singapore: Springer Singapore.

Gupta, N., Gandhi, V., Hariya, C. and Shelke, V., 2018, January. Detection of code clones. In *2018 International Conference on Smart City and Emerging Technology (ICSCET)* (pp. 1-4). IEEE.

Pyles, C., van Schalkwyk, F., Gorman, G.J., Beg, M., Stott, L., Levy, N. and Gilad-Bachrach, R., 2021. PyBryt: auto-assessment and auto-grading for computational thinking. *arXiv preprint arXiv:2112.02144*.

Karnalim, O., 2022. *Building awareness of programming plagiarism and collusion through similarity feedback generation* (Doctoral dissertation, University of Newcastle).

Niswar, M., 2022. Application of JavaScript Code Similarity Detection for Assessment of Web Programming Assignment. *EPI International Journal of Engineering*, *5*(2), pp.81-85.

J. Petrik, D. Chuda, "Bots and Gender Profiling with Convolutional Hierarchical Recurrent Neural Network.," in *CLEF (Working Notes)*, 2019.

Ma, W., Liu, R., Wang, L. and Vosoughi, S., 2020. Towards improved model design for authorship identification: A survey on writing style understanding. *arXiv preprint arXiv:2009.14445*.

Labadie-Tamayo, R., Castro-Castro, D. and Ortega-Bueno, R., 2020. Fusing stylistic features with deep-learning methods for profiling fake news spreaders. In *CEUR Workshop Proceedings* (Vol. 2696).

I. Gulis, D. Chudá, J. Petrik, "Plagiarism Detection in Students' Assignments Written in Natural Language," in *International Conference on e-Learning*, 2016, pp. 141.

Kuropiatnyk, O.S., 2019. Constructive and object-oriented modeling text for detection of text borrowings. *Системні технології*, *4*(123), pp.34-47.

Lu, Y., K, M.K., Mohammed, N. and Wang, Y., 2019, November. Homoglyph attack detection with unpaired data. In *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing* (pp. 377-382).

J. Petrik, D. Chudá, "Source code authorship approaches natural language processing," in *Proceedings of the 19th International Conference on Computer Systems and Technologies*, 2018, pp. 58–61.

Bogdanova, A., 2021, October. Source code authorship attribution using file embeddings. In *Companion Proceedings of the 2021 ACM SIGPLAN International Conference on Systems, Programming, Languages, and Applications: Software for Humanity* (pp. 31-33).

Bogdanova, A., Farina, M., Kholmatova, Z., Kruglov, A., Romanov, V. and Succi, G., 2022, November. Analysis of source code authorship attribution problem. In *2022 International Conference on Computers and Artificial Intelligence Technologies (CAIT)* (pp. 109-115). IEEE.

Alalawi, S.H., 2024. UNVEILING THE ORIGINS OF SOURCE CODE THROUGH AUTHORSHIP ATTRIBUTION: A COMPARATIVE STUDY OF AI AND HUMAN CODING PATTERNS.

Misini, M.S.A., Kadriu, A. and Canhasi, E., Authorship Analysis in Albanian Texts.

J. PETRIIK. "Perfectplaggie: Source code plagiarising tool," in *12th IIT. SRC*, 2016.

Duracik, M., Hrkut, P., Krsak, E. and Toth, S., 2020. Abstract syntax tree based source code antiplagiarism system for large projects set. *IEEE Access*, *8*, pp.175347-175359.

Alexandra-Cristina, C. and Olteanu, A.C., 2022, July. Material survey on source code plagiarism detection in programming courses. In *2022 International Conference on Advanced Learning Technologies (ICALT)* (pp. 387-389). IEEE.

J. Petrik, D. Chuda, "Twitter Feeds Profiling with TF-IDF.," in *CLEF (Working Notes)*, 2019.

Moreno-Sandoval, L.G., Pomares-Quimbaya, A. and Alvarado-Valencia, J.A., 2021. Celebrity profiling through linguistic analysis of digital social networks. *Computational Social Networks*, *8*(1), p.16.

Kavadi, D.P., Al-Turjman, F., Reddy, K.A.N. and Patan, R., 2021. A machine learning approach for celebrity profiling. *International Journal of Ad Hoc and Ubiquitous Computing*, *38*(1-3), pp.111-126.

Adi Narayana Reddy, K., Laskari, N.K., Shyam Chandra Prasad, G. and Sreekanth, N., 2022. Fusion-Based Celebrity Profiling Using Deep Learning. In *Intelligent System Design: Proceedings of INDIA 2022* (pp. 107-113). Singapore: Springer Nature Singapore.

Kalli, S.N.R., Kumar, B.N. and Jagadeesh, S., 2022. A Term Weight Measures based Approach for Celebrity Profiling. *Journal of Algebraic Statistics*, *13*(2), pp.2377-2391.

Kim, M.H. and Jang, B., 2020. Performance Evaluations of Text Ranking Algorithms. *한국컴퓨터정보학회논문지*, *25*(2), pp.123-131.

J. Petrik, D. Chuda, "The effect of time drift in source code authorship attribution: Time drifting in source code-stylochronometry," in *Proceedings of the 22nd International Conference on Computer Systems and Technologies*, 2021, pp. 87–92.

Jelodar, H., Meymani, M. and Razavi-Far, R., 2025. Large language models (llms) for source code analysis: applications, models and datasets. *arXiv preprint arXiv:2503.17502*.

Gong, S. and Zhong, H., 2025. Incremental learning of code authors over time. *Journal of Systems and Software*, p.112527.

Guo, H., Cheng, S., Jin, X., ZHANG, Z., Shen, G., Zhang, K., An, S., Tao, G. and Zhang, X., 2025. Profiler: Black-box AI-generated Text Origin Detection via Context-aware Inference Pattern Analysis.

Coursey, A., Tennyson, M. and Krotov, V., 2024. R Code Authorship Attribution using the ASAP Tool. *Journal of the Midwest Association for Information Systems (JMWAIS)*, *2024*(2), p.3.

A. Skurla, J. Petrik, "Authorship Profiling in Political Discourse on Twitter: Age and Gender Determination," in *Proceedings of the International Conference on Computer Systems and Technologies 2024*, 2024, pp. 82–86.