MINISTRY OF EDUCATION AND RESEARCH OF THE REPUBLIC OF MOLDOVA

Technical University of Moldova

Faculty of Computers, Informatics, and Microelectronics

Department of Software Engineering and Automation

Approved for defence
Department head:
Ion FIODOROV, phd, associate professor

„___” _________ 2025

RESEARCH OF STYLOMETRY FEATURES FOR INTRINSIC PLAGIARISM DETECTION SYSTEMS

Master's project

Student:	_____________	**Constantin Cazacu, IS-231M**
Coordinator:	_____________	**Mihail Gavrilița, university assistant**
Consultant:	_____________	**Svetlana Cojocaru, university assistant**

Chișinău, 2025

ABSTRACT

The thesis named Research of Stylometry Features for Intrinsic Plagiarism Detection Systems, presented by student Constantin Cazacu, was developed at the Technical University of Moldova.

It is written in English and contains 42 pages, 1 table, 6 figures, and 17 references.

The thesis consists of a list of figures, list of tables, an introduction, three chapters, a conclusion, and a list of references.

The report is structured into three chapters, each addressing a crucial aspect of the work process. The initial chapter serves as an introduction, presenting the identified problem that necessitates resolution. This chapter conducts a thorough exploration of the relevant field, offering comprehensive insights into the underlying issues. Additionally, it evaluates a proposed solution in comparison to existing products and services, thus facilitating a comprehensive analysis of their respective strengths and weaknesses.

The problem definition within this chapter is clear, accompanied by well-defined goals and objectives for the proposed solution. These objectives aim to enhance the quality and impact of doctoral research while streamlining the process for students and faculty. By establishing this analytical foundation, the report paves the way for further research and development, encouraging innovation and progress within the realm of doctoral studies.

The subsequent chapter provides the methodology and proposed approach to the researched field, backed by a detailed comparison of various approaches.

Lastly, the concluding section delves into a comprehensive overview of the system design, encompassing architectural style and components. Additionally, it encompasses the formulation of functional and non-functional requirements, which serve as essential guidelines for system identification, construction, and design.

This document is intended for readers with technical background.

REZUMAT

Teza intitulată Cercetarea Caracteristicilor Stilometrice pentru Sistemele de Detectarea Intrinsecă a Plagiarismului, prezentată de studentul Constantin Cazacu, a fost elaborată la Universitatea Tehnică a Moldovei.

Ea este redactată în limba engleză și conține 42 de pagini, 1 tabel, 6 figuri și 17 referințe.

Teza cuprinde o listă de figuri, o listă de tabele, o introducere, trei capitole, o concluzie și o listă de referințe.

Raportul este structurat în trei capitole, fiecare abordând un aspect crucial al procesului de lucru. Capitolul inițial servește drept introducere, prezentând problema identificată care necesită rezolvare. Acest capitol realizează o explorare aprofundată a domeniului relevant, oferind o perspectivă cuprinzătoare asupra problemelor subiacente. În plus, acesta evaluează o soluție propusă în comparație cu produsele și serviciile existente, facilitând astfel o analiză cuprinzătoare a punctelor forte și slabe ale acestora.

Definirea problemei în cadrul acestui capitol este clară, însoțită de scopuri și obiective bine definite pentru soluția propusă. Aceste obiective vizează îmbunătățirea calității și a impactului cercetării doctorale, simplificând în același timp procesul pentru studenți și cadre didactice. Prin stabilirea acestei baze analitice, raportul deschide calea pentru cercetare și dezvoltare viitoare, încurajând inovarea și progresul în domeniul studiilor doctorale.

Capitolul următor prezintă metodologia și abordarea propusă pentru domeniul cercetat, susținute de o comparație detaliată a diferitelor abordări.

În cele din urmă, secțiunea de încheiere oferă o imagine de ansamblu cuprinzătoare a proiectării sistemului, cuprinzând stilul arhitectural și componentele. În plus, aceasta cuprinde formularea cerințelor funcționale și nefuncționale, care servesc drept orientări esențiale pentru identificarea, construirea și proiectarea sistemului.

Acest document este destinat cititorilor cu pregătire tehnică.

CONTENTS

INTRODUCTION

Plagiarism detection has emerged as a pressing concern in academic, professional, and creative domains, where the integrity of intellectual property must be preserved. The rise of generative AI and large language models (LLMs) has further amplified the urgency of addressing plagiarism. These models, capable of producing human-like text, have blurred the lines between original and borrowed content, complicating traditional plagiarism detection methodologies. The ease with which LLMs can generate contextually relevant and stylistically consistent text has necessitated the evolution of detection systems that move beyond simple content comparison, addressing subtler inconsistencies within a document's writing style. Traditionally, plagiarism detection has relied on extrinsic methods, which involve comparing a given document against external reference materials, such as academic databases or the web. While effective in many cases, these methods are inherently limited to identifying copied content from known sources. This limitation has spurred interest in intrinsic plagiarism detection, which examines the stylistic coherence within a document itself, identifying deviations that might suggest sections authored by someone else. Intrinsic methods focus on the linguistic fingerprints of an author, seeking inconsistencies that signal potential instances of plagiarism, even when external sources are unavailable or inaccessible.

The foundation of this thesis lies in intrinsic plagiarism detection, particularly the use of stylometric features to quantify and analyse writing style. Stylometry, the study of linguistic style, encompasses various features—such as lexical diversity, syntactic structures, sentence length, and frequency of function words—that collectively characterize an author's unique voice. These features, when analysed systematically, can uncover patterns and anomalies that serve as indicators of stylistic inconsistencies. Such an approach offers the potential for a more nuanced detection mechanism, capable of identifying plagiarism in scenarios where conventional methods might fall short. To ground the work in existing knowledge, this research began with an extensive literature review, examining prior studies in intrinsic plagiarism detection and stylometry. Through this analysis, various methodologies and approaches were compared, highlighting their strengths, limitations, and applicability to modern challenges, including those posed by LLM-generated content. The findings from this review informed the conceptualization of a novel solution that balances the technical complexity of plagiarism detection with the simplicity required for an intuitive user experience.

The focus of this thesis is on designing the user experience for a plagiarism detection web application that integrates intrinsic detection techniques. A user-centric approach has been the focus to the development of this conceptual framework, ensuring that the application is tailored to meet the needs and expectations of its target audience—students, educators, and researchers. These users require a solution that is both accessible and efficient, enabling them to upload documents, view results, and interpret findings with minimal effort and maximum clarity. Key elements of the proposed application design include features such as color-coded visualizations of stylistic inconsistencies, which allow users to quickly identify and investigate suspicious sections within their documents. The decision to provide metrics like word count, stylistic inconsistency percentages, and the dominant writing style reflects a deliberate effort to present meaningful insights without overwhelming users. By prioritizing simplicity and transparency, the design ensures that the system remains approachable, even for those with limited technical expertise. The application makes use of the Mistral model from Ollama for processing documents, integrating advanced computational capabilities within a streamlined interface. This model enables the efficient analysis of text chunks, offering high precision in identifying stylistic anomalies while maintaining a lightweight computational footprint. The conceptual design emphasizes visual clarity, minimal interaction, and intuitive workflows, ensuring that users can engage with the system seamlessly. While the scope of this thesis is aimed at the conceptualization and design of the application, the implementation of the underlying logic and functionality is addressed in the work of a collaborator. Nevertheless, this research establishes a comprehensive foundation for the development of a robust, user-friendly plagiarism detection system. By balancing technical sophistication with user-centric design principles, this thesis aims to contribute a meaningful advancement in the field of intrinsic plagiarism detection, addressing the evolving challenges posed by new approaches to plagiarised content.

BIBLIOGRAPHY

[1] H. A. Chowdhury and D. K. Bhattacharyya, “Plagiarism: Taxonomy, Tools and Detection Techniques,” Jan. 19, 2018, *arXiv*: arXiv:1801.06323. doi: 10.48550/arXiv.1801.06323.

[2] R. R. Naik, M. B. Landge, and C. N. Mahender, “A review on plagiarism detection tools,” *International Journal of Computer Applications*, vol. 125, no. 11, pp. 16–22, 2015.

[3] S. Meyer Zu Eissen, B. Stein, and M. Kulig, “Plagiarism Detection Without Reference Collections,” in *Advances in Data Analysis*, R. Decker and H.-J. Lenz, Eds., in Studies in Classification, Data Analysis, and Knowledge Organization. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 359–366. doi: 10.1007/978-3-540-70981-7_40.

[4] M. AlSallal, R. Iqbal, V. Palade, S. Amin, and V. Chang, “An integrated approach for intrinsic plagiarism detection,” *Future Generation Computer Systems*, vol. 96, pp. 700–712, Jul. 2019, doi: 10.1016/j.future.2017.11.023.

[5] E. Stamatatos, “Intrinsic plagiarism detection using character n-gram profiles,” *threshold*, vol. 2, no. 1,500, 2009, Accessed: Jan. 08, 2025. [Online]. Available: https://icsdweb.aegean.gr/stamatatos/papers/PAN2009.pdf

[6] M. Kestemont, K. Luyckx, and W. Daelemans, “Intrinsic Plagiarism Detection Using Character Trigram Distance Scores - Notebook for PAN at CLEF 2011,” presented at the Conference and Labs of the Evaluation Forum, 2011. Accessed: Oct. 06, 2024. [Online]. Available: https://www.semanticscholar.org/paper/Intrinsic-Plagiarism-Detection-Using-Character-for-Kestemont-Luyckx/111cebc98be9d0db00edf24f1084fcc670de8593

[7] G. Ríos-Toledo, J. P. F. Posadas-Durán, G. Sidorov, and N. A. Castro-Sánchez, “Detection of changes in literary writing style using N-grams as style markers and supervised machine learning,” *PLoS ONE*, vol. 17, no. 7, p. e0267590, Jul. 2022, doi: 10.1371/journal.pone.0267590.

[8] B. Stein, N. Lipka, and P. Prettenhofer, “Intrinsic plagiarism analysis,” *Lang Resources & Evaluation*, vol. 45, no. 1, pp. 63–82, Mar. 2011, doi: 10.1007/s10579-010-9115-y.

[9] M. Muhr, R. Kern, M. Zechner, and M. Granitzer, “External and intrinsic plagiarism detection using a cross-lingual retrieval and segmentation system,” in *Notebook papers of CLEF 2010 LABs and workshops*, Citeseer, 2010, p. 22. Accessed: Jan. 08, 2025. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=3e8b05c34d5ee58ed0348f706e46d2762bc0d3b4

[10] M. F. Manzoor, M. S. Farooq, M. Haseeb, U. Farooq, S. Khalid, and A. Abid, “Exploring the Landscape of Intrinsic Plagiarism Detection: Benchmarks, Techniques, Evolution, and Challenges,” *IEEE Access*, vol. 11, pp. 140519–140545, 2023, doi: 10.1109/ACCESS.2023.3338855.

[11] A. Polydouri, E. Vathi, G. Siolas, and A. Stafylopatis, “An efficient classification approach in imbalanced datasets for intrinsic plagiarism detection,” *Evolving Systems*, vol. 11, no. 3, pp. 503–515, Sep. 2020, doi: 10.1007/s12530-018-9232-1.

[12] D. Curran, “An Evolutionary Neural Network Approach to Intrinsic Plagiarism Detection,” in *Artificial Intelligence and Cognitive Science*, L. Coyle and J. Freyne, Eds., Berlin, Heidelberg: Springer, 2010, pp. 33–40. doi: 10.1007/978-3-642-17080-5_6.

[13] G. Oberreuter, G. L’Huillier, S. A. Ríos, and J. D. Velásquez, “Outlier-Based Approaches for Intrinsic and External Plagiarism Detection,” in *Knowlege-Based and Intelligent Information and Engineering Systems*, A. König, A. Dengel, K. Hinkelmann, K. Kise, R. J. Howlett, and L. C. Jain, Eds., Berlin, Heidelberg: Springer, 2011, pp. 11–20. doi: 10.1007/978-3-642-23863-5_2.

[14] R. V. S. P. K. Ranatunga, A. Atukorale, and K. Hewagamage, *Intrinsic Plagiarism Detection with kohonen Self Organizing Maps*. 2011, p. 125. doi: 10.1109/ICTer.2011.6075041.

[15] M. Tschuggnall and G. Specht, “Plag-Inn: Intrinsic Plagiarism Detection Using Grammar Trees,” in *Natural Language Processing and Information Systems*, G. Bouma, A. Ittoo, E. Métais, and H. Wortmann, Eds., Berlin, Heidelberg: Springer, 2012, pp. 284–289. doi: 10.1007/978-3-642-31178-9_35.

[16] H. Ramnial, S. Panchoo, and S. Pudaruth, “Authorship Attribution Using Stylometry and Machine Learning Techniques,” in *Intelligent Systems Technologies and Applications*, S. Berretti, S. M. Thampi, and P. R. Srivastava, Eds., Cham: Springer International Publishing, 2016, pp. 113–125. doi: 10.1007/978-3-319-23036-8_10.

[17] K. Surendran, O. P. Harilal, P. Hrudya, P. Poornachandran, and N. K. Suchetha, “Stylometry Detection Using Deep Learning,” in *Computational Intelligence in Data Mining*, H. S. Behera and D. P. Mohapatra, Eds., Singapore: Springer, 2017, pp. 749–757. doi: 10.1007/978-981-10-3874-7_71.