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 defense
Department head:
Ion FIODOROV, phd, associate professor

„___" _____ 2025

ANALYSIS AND IMPLEMENTATION OF STYLOMETRIC FEATURES FOR INTRINSIC PLAGIARISM DETECTION SYSTEMS

Master's project

Student: _____ **Furdui Alexandru, IS-231M**

Coordinator: _____ **Gavrilita Mihail, university assistant**

Consultant: _____ **Cojocaru Svetlana, university assistant**

Chișinău, 2025

ABSTRACT

This thesis investigates the analysis and implementation of stylometric features for intrinsic plagiarism detection systems, addressing the increasing demand for effective methods to detect unoriginal content in written works. Intrinsic plagiarism detection operates by examining the internal stylistic characteristics of a document, identifying potential plagiarism based on changes in the author's writing style without requiring comparison to external sources.

The first chapter focuses on presenting the background research and motivation for this study. It discusses the relevance of intrinsic plagiarism detection in today's digital world, where a vast amount of information is easily accessible. A thorough examination of the current landscape in plagiarism detection technologies is provided, along with the limitations and challenges faced by traditional, external comparison methods.

The second chapter delves into related works, offering an in-depth review of previous studies in the fields of stylometry and plagiarism detection. It explores the various stylometric features that are essential for analyzing text, including word frequency, sentence length, and syntactic patterns, and how these are employed to detect stylistic shifts indicative of plagiarism.

In the first half of the third chapter, the technical design and implementation of the system are discussed. This section explains the feature extraction methods used, the algorithms for stylometric analysis, and the architecture of the proposed system. Special attention is given to the software tools and libraries used to build a reliable and scalable system.

The second half of the implementation chapter provides an evaluation of the system's performance. The experiments conducted are outlined, showcasing how the system detects stylistic anomalies in various documents. The results are analyzed based on performance metrics like precision, recall, and accuracy, highlighting the effectiveness and potential limitations of the system in different scenarios.

The conclusion presents a summary of the findings and their implications for plagiarism detection in academic and professional contexts. Future directions for improving the system, such as incorporating multilingual capabilities and enhancing accuracy, are also discussed.

REZUMAT

Această lucrare investighează analiza și implementarea caracteristicilor stilometrice pentru sistemele de detectare a plagiatului intrinsec, răspunzând cererii tot mai mari de metode eficiente pentru detectarea conținutului neoriginal în lucrările scrise. Detectarea plagiatului intrinsec funcționează prin examinarea caracteristicilor stilistice interne ale unui document, identificând potențialul plagiat pe baza schimbărilor în stilul de scriere al autorului, fără a fi necesară compararea cu surse externe.

Primul capitol se concentrează pe prezentarea cercetării de fond și a motivației pentru acest studiu. Se discută relevanța detectării plagiatului intrinsec în lumea digitală de astăzi, în care o cantitate vastă de informații este ușor accesibilă. Este oferită o examinare detaliată a peisajului actual al tehnologiilor de detectare a plagiatului, împreună cu limitările și provocările cu care se confruntă metodele tradiționale, bazate pe comparații externe.

Capitolul al doilea explorează lucrările conexe, oferind o revizuire detaliată a studiilor anterioare în domeniile stilometriei și detectării plagiatului. Acesta analizează diversele caracteristici stilometrice esențiale pentru analiza textului, inclusiv frecvența cuvintelor, lungimea propozițiilor și tiparele sintactice, și modul în care acestea sunt utilizate pentru a detecta schimbările stilistice indicatoare de plagiat.

În prima jumatate din al treilea capitol, sunt discutate aspectele tehnice ale designului și implementării sistemului. Această secțiune explică metodele de extragere a caracteristicilor utilizate, algoritmii pentru analiza stilo metrică și arhitectura sistemului propus. O atenție deosebită este acordată instrumentelor software și bibliotecilor folosite pentru a construi un sistem fiabil și scalabil.

A doua jumatate a capitolului trei oferă o evaluare a performanței sistemului. Experimentele efectuate sunt detaliate, demonstrând modul în care sistemul detectează anomaliile stilistice în diverse documente. Rezultatele sunt analizate pe baza unor metrici de performanță precum precizia, revocarea și acuratețea, evidențiind eficiența și posibilele limitări ale sistemului în diferite scenarii.

Concluzia prezintă un rezumat al constatărilor și implicațiilor acestora pentru detectarea plagiatului în context academic și profesional. Sunt discutate direcții viitoare pentru îmbunătățirea sistemului, cum ar fi integrarea capacităților multilingve și creșterea preciziei.

CONTENTS

INTRODUCTION 8
1 DOMAIN ANALYSIS 10
1.1 Topic Importance 11
1.2 Plagiarism Detection Methods 12
1.2.1 Extrinsic Plagiarism Detection 13
1.2.2 Intrinsic Plagiarism Detection 14
1.2.3 Hybrid Methods 16
1.3 Text Processing Steps for Plagiarism Detection 18
1.4 Chunking Process in Text Processing 19
1.4.1 Sentence Chunking 20
1.4.2 Phrase Chunking 22
1.4.3 N-Gram Chunking 23
1.4.4 Window-based Chunking 24
2 COMPUTER SYSTEM MODELING AND DESIGN 26
2.1 Behavioral Description of the System 26
2.2 Software Requirements 28
2.2.2 Non-Functional Requirements 31
2.3 Hardware Requirements 32
2.4 Software Interfaces 33
3 IMPLEMENTATION 35
3.1 Backend Implementation 35
3.1.1 API Design 37
3.1.2 Database Design 39
3.1.3 Frontend Development 39
3.1.4 Workflow Integration 40
3.2 Frontend Implementation 41
3.3 System Performance 43
CONCLUSION 47
BIBLIOGRAPHY 48

INTRODUCTION

Plagiarism, the act of using another's work without proper attribution, has long been a significant concern in academic, professional, and creative fields. With the rise of digital information and the ease with which content can be accessed and duplicated, detecting and preventing plagiarism has become an even more pressing issue. The proliferation of internet sources, databases, and online publications has made it increasingly challenging to ensure that written work is original. Consequently, institutions, educators, publishers, and companies rely heavily on advanced plagiarism detection systems to maintain the integrity of their content.

This paper explores the various approaches to plagiarism detection, focusing on the analysis and implementation of stylometric features for intrinsic plagiarism detection systems. Stylometry, the study of linguistic style, plays a vital role in detecting plagiarism by analyzing text for shifts in writing patterns that could indicate that the content was not written by the same author. Stylometric features such as word frequency, sentence length, and syntactic patterns allow for a detailed analysis of authorship within a single document, even without access to external sources for comparison. These features are particularly useful in intrinsic plagiarism detection systems, which identify unoriginal content based on internal inconsistencies in writing style rather than external matching.

The aim of this study is to provide a comprehensive understanding of the domain of plagiarism detection, covering extrinsic, intrinsic, and hybrid detection methods. The research examines the effectiveness of these methods and outlines the text processing steps necessary for implementing an efficient plagiarism detection system. By focusing on chunking as a core process in text processing, the study delves into how various chunking methods—sentence chunking, phrase chunking, and N-gram chunking—can enhance the detection of stylistic shifts and plagiarism.

The first section of this paper, provides a detailed overview of the scope and relevance of plagiarism detection in today's digital landscape. This section sets the foundation for understanding why plagiarism detection remains a critical issue and highlights the growing demand for sophisticated tools to combat it. Following this, the section delves into the significance of this research, emphasizing the need for developing systems that can detect plagiarism more accurately and efficiently.

The following sections explore the primary methods used for plagiarism detection. The chapter on "Plagiarism Detection Methods" outlines the differences between extrinsic and intrinsic approaches, examining how each method works and its limitations. Extrinsic methods rely on external databases and sources to identify copied content, while intrinsic methods, such as stylometric analysis, focus on the internal characteristics of the text itself. A third category, hybrid methods, blends both extrinsic and intrinsic techniques to achieve a more comprehensive detection process, providing a balance between external comparison and internal analysis.

The final sections of this study provide insights into the practical applications of these methods and discuss how they can be implemented in real-world systems. A detailed examination of sentence chunking, phrase chunking, and N-gram chunking reveals the importance of choosing the right chunking method for different types of text. These techniques are crucial for identifying plagiarism in large documents where stylistic shifts may not be immediately apparent. The study also discusses the limitations of existing systems and proposes improvements for enhancing detection accuracy.

Overall, this research aims to contribute to the ongoing development of plagiarism detection tools by offering a thorough analysis of current methods and techniques. As the digital landscape continues to evolve, so too must the tools used to ensure the originality and integrity of written content. This paper provides both a theoretical framework and practical insights into how modern plagiarism detection systems can be designed and improved to meet these challenges.

BIBLIOGRAPHY

[1] T. Kučečka, D. Chudâ, and P. Samuhel, “Selective chunking — Easy and effective way to estimate text similarity,” in 2013 IEEE 14th International Symposium on Computational Intelligence and Informatics (CINTI), Nov. 2013, pp. 381–385. doi: 10.1109/CINTI.2013.6705226.

[2] A. Prakash and S. K. Saha, “Experiments on Document Chunking and Query Formation for Plagiarism Source Retrieval.,” in CLEF (Working Notes), 2014, pp. 990–996. Accessed: Jan. 08, 2025. [Online]. Available: https://downloads.webis.de/pan/publications/papers/prakash_2014.pdf

[3] B. Stein and S. M. zu Eissen, “Intrinsic Plagiarism Analysis with Meta Learning.,” PAN, vol. 276, 2007, Accessed: Jan. 08, 2025. [Online]. Available: https://webis.de/events/pan-07/pan07-papers-final/stein07-intrinsic-plagiarism-analysis-with-meta-learning.pdf

[4] L. Seaward and S. Matwin, “Intrinsic plagiarism detection using complexity analysis,” in Proc. SEPLN, 2009, pp. 56–61. Accessed: Jan. 08, 2025. [Online]. Available: https://www.academia.edu/download/38491177/pan09-proceedings.pdf#page=64

[5] O. Htun and Y. Mikami, “Demonstration of text similarity metric for plagiarism detection,” Transactions on GIGAKU: Scope and Policy, 2012, Accessed: Jan. 08, 2025. [Online]. Available: https://www.researchgate.net/profile/Ohnmar-Htun/publication/267338211_Demonstration_of_Text_Similarity_Metric_for_Plagiarism_Detection/links/544dab3b0cf2d6347f45ca2f/Demonstration-of-Text-Similarity-Metric-for-Plagiarism-Detection.pdf#page=32

[6] A. Bohra and N. C. Barwar, “A Deep Learning Approach for Plagiarism Detection System Using BERT,” in Congress on Intelligent Systems, M. Saraswat, H. Sharma, K. Balachandran, J. H. Kim, and J. C. Bansal, Eds., Singapore: Springer Nature, 2022, pp. 163–174. doi: 10.1007/978-981-16-9113-3_13.

[7] G. Oberreuter, G. L’Huillier, S. A. Ríos, and J. D. Velásquez, “Approaches for intrinsic and external plagiarism detection,” Proceedings of the PAN, vol. 4, no. 5, p. 63, 2011.

[8] M. Apoorva, “The Application of Machine Learning in Detecting Plagiarism in Academic Works,” International IT Journal of Research, ISSN: 3007-6706, vol. 1, no. 1, Art. no. 1, Nov. 2023.

[9] H. Arabi and M. Akbari, “Improving plagiarism detection in text document using hybrid weighted similarity,” Expert Systems with Applications, vol. 207, p. 118034, Nov. 2022, doi: 10.1016/j.eswa.2022.118034.

[10] A. Vasuteja, A. V. Reddy, and A. Pravin, “Beyond Copy Paste: Plagiarism Detection using Machine Learning,” in 2024 International Conference on Inventive Computation Technologies (ICICT), Apr. 2024, pp. 245–251. doi: 10.1109/ICICT60155.2024.10544470.

[11] S. Gawali, D. S. Thakore, S. D. Joshi, and V. S. Shinde, “Plagiasil: A Plagiarism Detector Based on MAS Scalable Framework for Research Effort Evaluation by Unsupervised Machine Learning – Hybrid Plagiarism Model,” in Applied Machine Learning for Smart Data Analysis, CRC Press, 2019.

[12] I. Mukherjee, B. Kumar, S. Singh, and K. Sharma, “Plagiarism detection based on semantic analysis,” International Journal of Knowledge and Learning, vol. 12, no. 3, pp. 242–254, Jan. 2018, doi: 10.1504/IJKL.2018.092316.

[13] L. Ahuja, V. Gupta, and R. Kumar, “A New Hybrid Technique for Detection of Plagiarism from Text Documents,” Arab J Sci Eng, vol. 45, no. 12, pp. 9939–9952, Dec. 2020, doi: 10.1007/s13369-020-04565-9.

[14] V. K and D. Gupta, “Text plagiarism classification using syntax based linguistic features,” Expert Systems with Applications, vol. 88, pp. 448–464, Dec. 2017, doi: 10.1016/j.eswa.2017.07.006.

[15] O. Kamat, T. Ghosh, K. J, A. V, and R. P, “Plagiarism Detection Using Machine Learning,” Dec. 09, 2024, arXiv: arXiv:2412.06241. doi: 10.48550/arXiv.2412.06241.

[16] S. Alzahrani and H. Aljuaid, “Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: A study on Arabic-English plagiarism cases,” Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 4, pp. 1110–1123, Apr. 2022, doi: 10.1016/j.jksuci.2020.04.009.