DOI:10.59787/2413-5488-2025-52-4-5-18

Svitlana Biloshchytska, Aidos Mukhatayev, Oleksandr Kuchanskyi, Saltanat Sharipova, Nayla Murat, Zhan Amangeldiyev

Astana IT University, Astana, Kazakhstan

METHODS OF DETECTION AND REMOVAL OF CONCEALED BORROWINGS IN ACADEMIC WORKS OF STUDENTS IN THE KAZAKH LANGUAGE

Abstract: The article presents a critical review of modern methods of concealing borrowings in academic works of students in the Kazakh language. Three key directions are considered: semantic (paraphrasing, synonymizing, grammatical transformations), technical (hidden characters, substitution of Cyrillic letters with Latin ones, use of Unicode control characters), and structural (tables, schemes, images). Particular attention is paid to the specifics of the Kazakh language as an agglutinative language, which complicates the task of automatic plagiarism detection. Contemporary resources for hiding borrowing are analyzed. The authors propose methods for neutralizing such concealment, covering not only textual data but also tabular materials, as well as visual elements – diagrams and charts.

Keywords: Academic integrity; plagiarism; concealment of borrowings; Kazakh language; agglutinativity; anti-plagiarism; multimodal models.

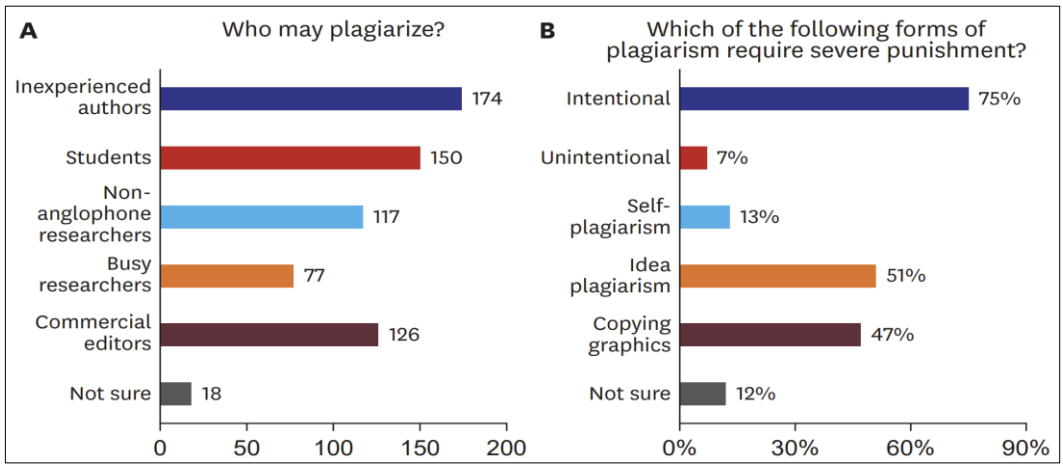
Introduction

In the context of the rapid growth of digital collections and open access to electronic libraries, databases, and various online resources, the issue of academic misconduct has become increasingly relevant. A growing number of students, when writing academic papers, resort to using ready-made materials without citing the source, which leads to a rise in texts with a high level of borrowings (Pudasaini et al., 2024; Boucher & Anderson, 2021). As a result, plagiarism checking of all works has become a mandatory procedure to ensure academic integrity and maintain the quality of education.

Issues of academic misconduct in Kazakhstan have been examinated within the framework of an international online survey devoted to the perception of plagiarism among researchers and journal editors in non-English-speaking countries (Latika Gupta et al., 2021). The results revealed that the most common form of violation is paraphrased plagiarism (69% of cases). Among the risk factors, respondents highlighted students (71%), researchers with limited language proficiency (55%), and representatives of commercial editing agencies (60%) (Fig.1). These data indicate the presence of a systemic problem in Kazakhstan regarding the perception and prevention of plagiarism and emphasize the need to introduce targeted educational courses and modern anti-plagiarism technologies into academic practice.

Figure 1

Physicians and scholars' perception of plagiarism. (Latika Gupta et al., 2021)



Modern anti-plagiarism systems face a serious challenge: students increasingly resort to bypass strategies, including the insertion of hidden characters, the substitution of Cyrillic letters with Latin ones, changes in document structure, or the use of automatic paraphrasing tools (Almuhaideb & Aslam, 2022). Such methods alter the visual appearance of the text, significally complicating its analysis by standard plagiarism detection algorithms. Effective countermeasures include Unicode normalization, removal of hidden characters, and the use of deep learning models trained on Kazakh-language corpora (e.g., KazBERT and Kaz-RoBERTa), which are capable of accounting for morphological complexity and detecting semantic similarities (Togmanov et al., 2022; Toiganbayeva et al., 2021). This intefrated approach helps to substantially reduce the risk of artificially concealed plagiarism and ensures a more objective assessment of the originality of academic texts.

The article by Khaled F. & Sabeeh M. (2021) provides a review of methods and tools for plagiarism detection, including both literal and intellectual forms. The authors present a classification of plagiarism types (textual, source code, mosaic, metaphorical, etc.) and emphasize that intellectual plagiarism – involving paraphrasing, translation, and structural modifications – is significantly more difficult to detect. Both intrinsic and extrinsic detection methods are considered, as well as modern tools ranging from MOSS and Turnitin to newly emerging online services. The authors also discuss datasets used for training and testing systems (WordNet, PAN) and various analysis approaches: n-grams, semantic and stylometric methods, and hybrid models. They conclude that no single method is universal; instead, a balance between accuracy and processing time is required, alongside the comprehensive development of tools to combat the ever-evolving forms of plagiarism.

Traditional methods of text data processing, despite their wide adoption and proven effectiveness in several tasks, demonstrate significant limitations when applied to multimodal documents that include both textual and visual components. Modern electronic documents are often complex structural entities, where information is conveyed not only through linear text but also via diagrams, charts, infographics, and other visual means. This nature of content requires analysis systems to account for the diversity of data representation, which goes beyond the capabilities of traditional text-oriented models.

The aim of this study is to conduct a critical analysis of the methods used to conceal borrowings and approaches to their elimination in academic works of students in the Kazakh language, taking into account the specific features of multimodal documents that include both textual and visual components.

Literature Review

Modern systems for detecting duplicates and near-duplicates are actively applied in various fields - from academic and scientific work to healthcare, electronic document management, and information retrieval.

Over the past decades, many approaches have been developed, relying on different methodological principles. However, in practical applications these solutions reveal significant limitations, especially when the task involves processing multimodal documents in the Kazakh language.

Text processing in Kazakh still faces several specific challenges, particularly in tasks related to the detection of near-duplicates. Unlike English and other languages that are widely represented in corpora, Kazakh is characterized as an agglutinative language, where lexemes change through the sequential addition of affixes to a root. This generates an enormous number of possible word forms, making exact string matching difficult.

Agglutinativity leads to morphological diversity even when expressing the same meaning. For example, the same phrase may appear with different endings depending on case, number, or person. This makes simple methods based on exact or partial token matching ineffective. The lack of high-quality morphological analyzers for the Kazakh language further reduces the accuracy of semantic text comparison.

At present, the number of corpus resources, annotated datasets, and pre-trained language models for the Kazakh language is significantly lower compared to more widely used languages. This hinders the training of effective neural models, including transformers, which rely on large-scale training data. As noted by Bogdanchikov et al. (2022), the lack of high-quality embeddings (word2vec, FastText, BERT-based models) for Kazakh is one of the main reasons for the limited applicability of modern NLP tools. However, in recent years, several specialized resources have been introduced: the KazNERD dataset (Yeshpanov et al., 2022), the KOHTD handwritten corpus (Toiganbayeva et al., 2021), as well as language models such as KazBERT and Kaz-RoBERTa (2023-2025), which have become an important foundation for fine-tuning and adapting plagiarism-detection tasks.

A common issue in Kazakh texts is the mixing of Cyrillic and Latin scripts, especially when words are intentionally distorted to bypass plagiarism detectors. For example, the letters "A", "O", "C", "E", "H", "P", "K" and others can easily be substituted with visually similar Latin counterparts. Without specially designed detection mechanisms, such substitutions often remain unnoticed by duplicate-detection systems. Semantic comparison of Kazakh texts is also complicated due to the insufficient training of models on relevant corpora. Even modern transformers such as multilingual BERT or XLM-RoBERTa demonstrate relatively low accuracy when applied to Kazakh documents, as shown in several recent experiments. At the same time, national Kazakh-specific models, including Kaz-RoBERTa, have demonstrated certain improvements with fine-tuning, though their performance still lags behind larger multilingual systems. This underlines the importance of expanding Kazakh-oriented training corpora and further improving model architectures (Tleubayeva & Shomanov, 2024).

In Bakiyev B. et. al (2022), a method for calculating text similarity in Kazakh is proposed, which incorporates synonyms into an extended TF-IDF model. The author emphasizes that traditional TF-IDF poorly accounts for semantic substitutions of words, which are frequently used in academic writing as a method of concealing plagiarism. Replacing words with synonyms allows borrowings to be hidden while retaining the overall meaning, making it harder for classical algorithms to detect. The proposed approach adds a thesaurus-based processing layer to capture semantic relations between words. This enables the detection of paraphrased text and paraphrase plagiarism. Thus, the study highlights the necessity of accounting for semantic features of the Kazakh language in plagiarism detection.

The article by Rakhimova D. et. al (2021) discusses a hybrid approach to the semantic analysis of Kazakh texts. The authors propose a combination of statistical and neural methods for analyzing the semantic similarity of documents. They note that traditional algorithms are highly sensitive to syntactic modifications (such as word order changes or case substitutions), which are frequently used to conceal borrowings. The hybrid approach mitigates these manipulations by considering not only the surface form but also the deeper semantic content of the text. As a result, the authors demonstrate that combining different analytical methods can improve the effectiveness of detecting hidden plagiarism in Kazakh.

The study by Lizunov, P. et. al (2021) describes a methodology for detecting near-duplicate documents in scientific texts. Particular attention is given to cases where authors alter only minor elements of a document (formatting, sentence reordering, minimal edits) to bypass plagiarism detection. For the Kazakh language, such concealment methods are particularly challenging due to morphological complexity and affixation, which generate a large variety of word forms. The authors propose a combined method that integrates both lexical and structural analysis. This approach makes it possible to identify documents with a high degree of technically disguised similarities. Thus, the study demonstrates an effective strategy for addressing superficial text editing aimed at concealing plagiarism.

The research by Ayazbayev D. et al. (2023) addresses the task of determining semantically similar words in Kazakh using semantic similarity metrics. The authors highlight that the use of synonyms and semantically related words is one of the primary techniques for circumventing anti-plagiarism systems. The proposed methodology enables the automatic detection of such substitutions and the identification of hidden borrowings. The system constructs vector representations of words and compares them to measure semantic similarity. This makes it possible to detect paraphrasing and other intellectual techniques of concealing plagiarism. The work contributes to the development of more accurate systems for analyzing Kazakh texts.

In the article (Prieur M. et al., 2022). The PIKA system for detecting duplicates in the knowledge base is described. Although it does not focus specifically on the Kazakh language, the methods proposed by the authors are also applicable to low-resource languages. PIKA analyzes the structural and semantic characteristics of the text, which makes it possible to identify hidden borrowings even with changes at the level of words or sentences. This is especially important for the Kazakh language, as techniques are often used to change the form of words or replace them with similar terms. The work shows the importance of using more sophisticated duplicate detection algorithms that go beyond a simple lexical match.

A study (Tolegen G. et al., 2020). It is devoted to the recognition of named entities in Kazakh texts using neural networks. At first glance, it is not directly related to plagiarism, but the identification of entities is important for the correct analysis of borrowings. Students and authors often leave borrowed fragments with proper names or terms, which gives away plagiarism, despite the paraphrasing. The NER model allows you to accurately identify such elements and use them as markers to detect plagiarism. The authors show that neural networks are able to adapt to the morphological features of the Kazakh language. This increases the effectiveness of intelligent anti-plagiarism systems.

There are practically no open datasets containing original/incomplete duplicate pairs for the Kazakh language. This limits the possibilities of evaluating models, as well as hinders the reproducibility and comparability of research. Together, these factors require the development of adapted methods of preprocessing, lemmatization, identification of Latin characters in the text, as well as training specialized models on domain names and data in the Kazakh language. Considering all these features will allow us to build a truly effective system for detecting incomplete duplicates in texts in the Kazakh language.

(Togmanov M. et al., 2025) is a benchmark for evaluating language models (Kazakh, Russian, regional knowledge of Kazakhstan), including in relation to text processing tasks. Although it does not directly focus on plagiarism, it demonstrates that modern models do not do well with the Kazakh language in the tasks of understanding and logic. This indirectly points to a problem: the low resource availability of the Kazakh language makes it difficult to build stable loan detection systems. The authors emphasize the need to expand the cases and tests that will allow models to better identify hidden borrowings. Thus, KazMMLU can be considered as a foundation for future research in the field of plagiarism detection.

Classical methods based on the representation of text in the form of a bag of words, vector spatial models (TF-IDF) and shingling were developed primarily for working with pure texts without visual and structural elements. Their advantage lies in the simplicity of implementation and high interpretability, as well as the ability to quickly process large text corpora (Henzinger, M., 2006, Mohammadi, H. & Khasteh, S. H., 2018). However, even with minimal structural changes, such as rearranging paragraphs, reformulating, using synonyms, or embedding text in images, these methods lose their informative value and become a source of false positive or false negative results. One of the key limitations of these approaches is their inability to consider the context and structure of the document, especially if the text is embedded in an image or accompanied by graphic elements. An example would be technical reports, instructions, or scientific publications where basic information is presented in the form of diagrams, flowcharts, and annotated images. When trying to analyze such materials, text-centric methods ignore the visual component, which leads to the loss of semantically significant fragments. In some cases, the contextual meaning of an inscription in a flowchart cannot be determined without analyzing its position, shape, or relationships with other elements, which is completely excluded when using, for example, TF-IDF or MinHash (Henzinger, M., 2006, Fisichella M. et al., 2011).

Documents with a multimodal structure, including text, images, and tables in a single layout, are particularly difficult. Applying traditional methods to them often requires pre-highlighting the text component, which is implemented through optical character recognition (OCR). However, OCR, especially when working with scanned documents or low-quality diagrams, is prone to recognition errors, structural distortions, and loss of important contextual information (Silcock et al., 2022). Thus, even preprocessing becomes a source of noise and unreliable data, which is subsequently processed using methods not designed for this kind of input.

The article (Cha Y. et al., 2005) is devoted to the study of binary similarity measures and their application in the task of recognizing handwritten characters. The authors consider and compare a wide range of metrics, including the Hamming distance, the Jaccard coefficient, the Dice measure, and several others used to analyze binarized images. Experiments have shown that different similarity measures exhibit different resistance to noise and handwriting variability: some metrics are more sensitive to changes in line thickness, while others remain stable with variations. As a result, the authors conclude that choosing the appropriate metric significantly affects the final accuracy of handwritten character recognition. The work contributes to the optimization of binary image processing methods and emphasizes the importance of correctly selecting similarity measures in computer vision tasks.

A study (Wahle et al., 2022) demonstrates that systems based on pre-embeddings and classifiers can detect texts processed with paraphrasing tools, but the effectiveness depends on the degree of change (how strongly phrases are paraphrased, whether synonyms are used, whether the structure is preserved).

A comparative analysis of the possibilities and limitations of traditional methods in the context of multimodal data is presented in Table 1.

Table 1

Problems of using traditional methods in the analysis of multimodal documents

Method/ approach	Advantages	Limitation in multimodal analysis	Sources
TF-IDF	Simple implementation, high processing speed	Ignores word order and document structure; not applicable to visual information	HenzingerM., 2006, Mohammadi H. & Khasteh S. H., 2018
Shingling	Detecting partial text matches	Sensitive to word reordering; incapable of processing diagrams and graphical objects	Henzinger M., 2006, Fisichella M. et al., 2011
MinHash	Scalability for large data volumes	Loss of semantic context; not applicable to images and charts	Fisichella M. et al., 2011
OCR + TF-IDF	Ability to work with scanned documents	Recognition errors; ignores visual layout; sensitive to noise	Silcock et al., 2022, Zhang M. et al., 2023

As can be seen from the table, even when combining methods such as OCR and TF-IDF, the resulting model remains vulnerable to noise, informal structures, and unobvious visual differences. In particular, approaches based on the separation of text from images often ignore the spatial arrangement of objects, which in the case of diagrams can be crucial for understanding the logic of the document. Attempts to improve the situation through text normalization, proposed, for example, in the RETSim architecture (Zhang M. et al., 2023) can reduce sensitivity to OCR artifacts, but they do not solve the problem of the lack of analysis of the layout structure.

Against the background of these limitations, there is an increasing interest in layout-aware models and multimodal neural networks capable of simultaneously analyzing text, images and their mutual arrangement. The LayoutLMv2 model, which has demonstrated effectiveness in the tasks of visual understanding of documents, allows taking into account both textual and visual features, while preserving structural information about the document (Xu Y. et al., 2021). However, the implementation of such approaches requires significant computing resources and the availability of marked-up multimodal datasets, which is currently difficult, especially in the context of low-resource languages and specific formats such as schemes in Kazakh.

The lack of multimodal datasets and the lack of data reflecting local linguistic and cultural characteristics pose a major problem for loan recognition. For the Kazakh language, despite its status as the state language, there are practically no open document bodies marked for duplication, which would include tables and images. Studies such as Bogdanchikov A. et al. 2022. emphasize the difficulties in applying English-language models to the Kazakh context due to agglutinative morphology, free word order, and spelling variations. Models trained in other Turkic languages do not demonstrate sufficient quality in direct translation, and specialized embedding representations adapted to Kazakh vocabulary are available only to a limited extent (Ayazbayev D. et al., 2023)

Thus, traditional methods, despite their historical significance and convenience, demonstrate serious methodological limitations when applied to the tasks of analyzing documents containing both textual and graphical information. Their inability to take into account the visual context, layout structure, and semantic connections between different modalities makes them unsuitable for solving the problems of identifying incomplete duplicates in modern digital archives. This justifies the need for a transition to integrative, multimodal systems, which will be discussed in the following sections.

In the study (Elkhatat A. et al., 2021). It is shown that systems like PlagScan, StrikePlagiarism, Turnitin, etc. Sometimes they are unable to detect so-called "image-text

plagiarism" or borrowings hidden in images, tables, or nonstandard fonts and formats. Students use this by preserving the structure of the document, but changing text fragments using synonyms or hidden characters.

Results and Discussion

To effectively identify borrowings, it is necessary to bring all documents to a single, reference format, eliminating the variability in the presentation of text, tables, diagrams, diagrams and images.

Methods of processing content elements to neutralize methods of concealment of borrowings can be divided into two groups: methods of neutralizing technical (formal) concealment of borrowings (Table. 2) and methods and models for detecting borrowings in semantic (intellectual) content changes (Table 3).

Technical methods include the insertion of invisible characters (zero-width space, soft hyphen), substitution of Cyrillic letters with Latin letters, manipulation of encodings and the introduction of Unicode control characters ("Trojan Source"). These techniques change the appearance of the text for analysis systems, but preserve its readability for humans (Boucher & Anderson, 2021). Unicode normalization (NFC/NFKC), removal of hidden characters, and the use of algorithms for detecting homoglyphs (Almuhaideb & Aslam, 2022) are proposed to combat them.

The developed methods make it possible to minimize the impact of techniques for hiding borrowings, unifying the presentation of content into a reference form. This significantly improves the accuracy of loan detection algorithms and contributes to the objective analysis of electronic documents.

Table 2

Methods of preparing Kazakh-language content to neutralize the concealment of borrowings during technical changes

№	Content preparation method	Tasks of the method	Description
1	Text cleaning from hidden and invisible characters	Removal of hidden characters	Exclusion of characters outside the standard Unicode range (Zero Width Space, Soft Hyphen, Zero Width Joiner, Right-to-Left Override, etc.).
		Normalization of spaces and line breaks	Bringing spaces and line breaks to unifies standard (removing extra spaces, line breaks, tabulations)
		Decoding encoded characters	Converting characters encoded in HTML- or Unicode-formats into standard forms.
2	Conversion of text to a standard alphabet	Replacement of visual analogues	Replacing characters with identical appearance but different encodings (e.g., «а» (U+0430) → «a» (U+0061)).
		Case unification	Converting text to a unified case (e.g., all letters to lowercase)
		Normalization of special characters	Replacing non-standard characters (e.g., non-breaking spaces) with standard ones
		Normalization of diacritical marks	Converting letters with diacritics to a unified standard (e.g., decomposition of composite symbols into NFC form)
3	Document structure adjustment	Structure unification	Removing unnecessary page breaks, unifying paragraphs
		Formatting standardization	Bringing headings, lists, and tables to a unified formatting style
		Metadata normalization	Checking and cleaning hidden data (document properties, comments, bookmarks) that may be used to bypass detection

Semantic techniques include paraphrasing, replacing words with synonyms, grammatical transformations, and translation plagiarism. These methods preserve the general meaning of the text, but change its surface, reducing the effectiveness of shingle and lexical methods. This problem becomes especially serious in the Kazakh language, where agglutinativity creates hundreds of word forms for a single root (Yeshpanov et al.. 2022), In response, the researchers propose the use of contextual language models (XLM-R, KazBERT, Kaz-RoBERTa) capable of taking into account semantic transformations (Conneau et al., 2020).

Table 3

Methods of preparing Kazakh-language content to neutralize the concealment of borrowings with semantic changes

№	Content processing method	Purpose/ description	Example (in Kazakh)
1	Lemmitization (reducing words to their base form)	Removes grammatical forms (case, number, person) to identify the lexical root	«Оқушылар мектептерінде болды» → «оқушы мектеп бол»
2	Synonym normalization	Converts synonyms into standard or frequently used forms	«Ғылыми зерттеу» = «ілімдік ізденіс» → «ғылыми зерттеу»
3	Stop-word removal	Excludes functional words that do not affect meaning, while preserving semantic structure	«Бұл мақалада біз қарастырамыз...» → «мақала қарастыру»
4	Collocation standardization	Brings non-standard expressions to stable, common phrases	«Сабақ барысында оқушылар білім алады» → «оқушылар білім алады»
5	Syntactic normalization	Reconstructs sentence structure to restore original meaning	«Оқушылар бұл тапсырманы орындап шықты» ↔ «Бұл тапсырманы оқушылар орындады»
6	Морфологический morphological analysis	Extracts root and affixes to identify similarity in hidden borrowed words	«үйренгендер», «үйреніп жатыр», «үйрену» → всё сводится к «үйрен»
7	Semantic analysis	Detects hidden borrowings at the meaning level despite formal changes	«Оқушы білім алады» ↔ «Білім оқушыға беріледі»
8	N-gram ananlysis (words or characters)	Compares text fragments (collocations) to identify recurring structures	«Тәуелсіз Қазақстан – болашағы жарқын ел» → «Қазақстан – жарқын болашағы бар ел»
9	Back-transition	Used to detect borrowings hidden by translation from another language	Рус.: «Он имеет большую значимость» → Каз.: «Оның маңызы зор» → Рус.: «Он важен»
10	Phoenetic normalization	Corrects orthofraphic and phoenetic distortions (transliteration, typos, etc.)	«Қаласақ» → «қаласақ», «әлеуметтік» → «әлеуметтік»

A very common practice of hiding borrowings in academic papers is to convert textual information into tabular form. The authors intentionally replace certain sections of the text with tables, which makes it difficult to detect plagiarism when using traditional text-oriented algorithms. Additionally, when working with tables, various masking methods are used: rearranging rows and columns, changing units of measurement (for example, replacing grams with kilograms), paraphrasing descriptive elements, as well as modifying numbering, encodings, or the format of data representation. Such techniques significantly complicate the

automatic comparison of information and require the development of specialized methods for detecting borrowings in tabular structures (Table 4).

Table 4

Methods of preprocessing tables before checking for plagiarism

№	Content processing method	Purpose/ description	Result
1	Extracting text from table cells	Converts table content into linear text (rows → paragraphs); preserves logical structure (header + content); removes hidden characters, extra spaces, and line breaks	The anti-plagiarism system can recognize text from tables, not just a "picture"
2	Formatting unification	Brings fonts, styles, and text cases to a unified form; replaces visually similar characters (Latin/Cyrillic); removes HTML markup and invisible tags	Eliminates masking through different fonts, spaces, or character substitution
3	Table structure normalization	Converts complex tables (with merged cells, nested tables) into a simple matrix; automatically aligns headers and labels; preserves context	Maintains readability and enables correct checking of table content
4	Extraction of numerical and symbolic data	Converts numbers, dates, and formulas into text format; replaces special characters; unifies measurement units and abbreviations	Ensures comparability of numerical data and symbols during verification
5	Segmentation into logical blocks	Splits large tables into subtables; adds "keys" to link headers and values	Increases verification accuracy and preserves the "header–data" relationship
6	Integration into the main text of the document	Incorporates pre-processed tables into the main text body	Provides comprehensive document checking, including tables

The problem of detecting plagiarism in image documents remains one of the least solved problems in anti-plagiarism systems. Unlike texts, where morphological and semantic analysis can be used, images, diagrams and diagrams require special preprocessing methods. To determine whether a visual object is original or borrowed from other sources, it is necessary to transform it into a form suitable for comparison and search through large collections. Such processing includes steps for cleaning, normalizing, and extracting features that will allow images to be compared with existing databases and resources on the Internet (Table 5).

Table 5

Preprocessing images to check for plagiarism

Stage	Description
Extract images from a document	Extract embedded images, diagrams, and charts from DOCX/PDF formats.
Extract text from images	Use OCR (e.g., Tesseract, EasyOCR) for text recognition in diagrams, scans, and charts. The resulting text can be compared with databases for borrowings.
Cleaning and normalization	Convert to standard format (PNG/JPEG), unify size (224×224 px), normalize color palette.
Noise removal	Remove watermarks and artifacts; apply binarization and contrast enhancement for diagrams and charts.
Feature extraction	For photos: CNN embeddings (ResNet, VGG, EfficientNet). For diagrams: OCR text, SIFT/SURF/ORB descriptors, structural features.
Database and Internet search	Compare with local image databases and Internet resources (Google Reverse Image Search, TinEye). Use perceptual hashing methods (pHash).
Comparison and similarity assessment	Calculate similarity (cosine similarity, Euclidean distance) and determine the probability of borrowing.

Despite significant progress in the development of anti-plagiarism systems, the problem of recognizing borrowings in diagrams, images, and drawings remains unresolved. Most modern tools are focused on textual information and demonstrate low efficiency in analyzing visual content. This is especially acute in the case of the Kazakh language: due to the limited corpus resources and the lack of specialized algorithms, verification of multimodal documents is complicated. Thus, the detection of plagiarism in graphic elements and diagrams requires further research and development of methods that consider both the visual and linguistic features of Kazakh content.

In recent years, systems and services have appeared on the market that are directly focused on increasing the uniqueness of the text and hiding borrowings outside the framework of academic standards. They offer functions for paraphrasing, synonymizing, changing style, and replacing text elements in a way that preserves meaning but reduces obvious similarities. A study (Ruben Comas et al., 2023) shows that students actively use online paraphrasing and translation tools to bypass text similarity checking systems and reduce the percentage of overlap with the original text.

Most international solutions are limited to English and Russian, while Kazakh remains a low-resource language. Nevertheless, some services designed to transform a document to conceal borrowings with the Kazakh language are presented in Table 6.

Table 6

Comparative table of 10 popular anti-plagiarism bypass systems

System	Languages	Applied modifications	Techniues (methods)
Ref-n-Write Paraphrasing Tool	English	Paraphrasing, synonym substitution, sentence restructuring	Lexico-syntactic analysis, template database of academic texts
Undetectable.ai	English	Paraphrasing, synonym replacement, generation of new formulations	AI model for generating unique text, statistical analysis
Netus AI Bypasser	English (partial support for other languages)	Paraphrasing, rewriting with LLMs, text restructuring	Deep neural networks, generative models for plagiarism evasion
Antiplagius	Russian, Kazakh (limited)	Hidden characters, character substitution, fragment reordering	Technical manipulations with encodings and symbols
Фокусник	Russian, Kazakh	Character substitution, line reordering, use of encodings	Mechanical transformation of text structure, tables, and symbols
Viper Anti Plagiarism	English	Match detection, basic paraphrasing	Database and online resource search, simple paraphrasing
Антиплагиат Киллер	Russian, Kazakh (adapted)	Paraphrasing, technical bypass techniques (Zero-Width characters, Soft Hyphen)	Technical evasion of plagiarism detection, invisible characters, encodings
AntiplagiatKiller	Russian, Kazakh, English	Paraphrasing	Manual rewriting, text paraphrasing with neural networks
Антиплагиат Фокс	Russian, Kazakh	Automatic word substitution, insertion of hidden characters	Mechanical synonymization, hidden characters, manual rewriting
Antiplagiat.org	Russian, English, Kazakh	Text modification, insertion of invisible characters, reordering	Basic bypass methods: reordering, encodings, hidden characters, manual rewriting
Antiplag.kz	Kazakh, Russian	Paraphrasing, synonymization, text restructuring, character substitution	Web-based encoding

Despite the fact that at the moment the number of systems focused on hiding borrowings with support for the Kazakh language remains limited, the dynamics of information technology development suggests a different future. Given the rapid progress in artificial intelligence and natural language processing, it can be predicted that the number of such services will increase in the coming years. This creates additional challenges for the academic community, as the improvement of such tools will inevitably lead to more complex tasks in identifying borrowings and will require the development of more reliable methods of counteraction.

Conclusion

The review showed that the problem of hiding borrowings in academic works in the Kazakh language remains extremely relevant in the context of digitalization and the growing number of electronic documents. The methods used by students cover a wide range – from technical manipulations (inserting hidden characters, replacing Cyrillic letters with Latin analogues, using Unicode control characters) to semantic and structural transformations (synonymization, paraphrasing, changing the structure of tables and diagrams).

The peculiarities of the Kazakh language as an agglutinative language complicate the task of identifying hidden borrowings, since the variety of word forms and syntactic flexibility make it possible to effectively bypass classical algorithms. An analysis of existing approaches has shown that traditional methods based on bag-of-words, TF-IDF and shingling demonstrate low efficiency when working with multimodal documents that include not only text, but also tables, diagrams and images.

In this regard, modern contextual language models, including KazBERT and Kaz-RoBERTa, as well as multimodal architectures such as LayoutLMv2, are of particular importance. Their use makes it possible to take into account the semantic level of the text, as well as the layout and visual structure of the document, which significantly increases the reliability of anti-plagiarism systems.

At the same time, it was revealed that services offering loan concealment tools (for example, Antiplagiat-Killer, Antiplagiat Fox, Netus AI, etc.) are developing in the market in parallel, including those with support for the Kazakh language. This creates new challenges for the academic community, as improving such services will inevitably lead to more difficult plagiarism detection.

Thus, in order to ensure academic integrity, it is necessary:

1. To develop specialized methods of normalization and preprocessing of Kazakh-language content.
2. Create original/duplicate corpus resources and datasets for training and testing models.
3. Implement multimodal neural network architectures that take into account textual and visual data.
4. Strengthen educational initiatives aimed at fostering a culture of academic integrity.

These measures together will improve the effectiveness of anti-plagiarism systems and minimize the impact of methods of hiding borrowings in academic texts in the Kazakh language.

Conflict of Interest Statement

The authors declare no potential conflicts of interest regarding the research, authorship, or publication of this article.

Funding Information

This research work was carried out within the framework of the scientific project AP23490123 «Development of a system to detect plagiarism using combined methods and models for finding near-duplicate, focusing on the Kazakh language.» for 2024-2026, financed by the Committee of Science Ministry of Science and Higher Education of the Republic of Kazakhstan.

Author Contributions

Svitlana Biloshchytska: Data curation, Writing – Original draft preparation, Project administration. Aidos Mukhatayev: Methods for neutralizing borrowings in text documents. Oleksandr Kuchanskyi: Methods for neutralizing borrowings in tables, charts and diagrams. Saltanat Sharipova: Analysis of concealment of borrowings in text documents. Nayla Murat: Analysis of hidden borrowings in tables. Zhan Amangeldiyev: Analysis of concealment of borrowings in multimodal documents.

References

Almuhaideb, A. M., & Aslam, N. (2022). Homoglyph attack detection model using machine learning and hashing. *Sensors, 22* (14), 5118.

Ayazbayev, D., Bogdanchikov, A., Orynbekova, K., & Varlamis, I. (2023). Defining Semantically Close Words of Kazakh Language with Distributed System Apache Spark. *Big Data and Cognitive Computing*, *7*(4), 160. https://doi.org/10.3390/bdcc7040160

Bakiyev, B. (2022, April). Method for determining the similarity of text documents for the Kazakh language, taking into account synonyms: extension to TF-IDF. In 2022 International Conference on Smart Information Systems and Technologies (SIST) (pp. 1-6). IEEE.

Bogdanchikov, A., Ayazbayev, D., & Varlamis, I. (2022). Classification of Scientific Documents in the Kazakh Language Using Deep Neural Networks and a Fusion of Images and Text. *Big Data and Cognitive Computing*, *6*(4), 123. https://doi.org/10.3390/bdcc6040123

Boucher, N., & Anderson, R. (2023). Trojan source: Invisible vulnerabilities. *Proceedings of the 32nd USENIX Security Symposium*, 123–140. (Original preprint 2021).

Cha, S.-H., Yoon, S., & Tappert, C. C. (2005). On binary similarity measures for handwritten character recognition. In *Proceedings of the Eighth International Conference on Document Analysis and Recognition (ICDAR'05)* (Vol. 1, pp. 4–8). IEEE. https://doi.org/10.1109/ICDAR.2005.173

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2021). Some students' plagiarism tricks, and tips for effective check. *International Journal for Educational Integrity, 17*(1), 15. https://doi.org/10.1007/s40979-021-00082-w

Fisichella, M., Deng, F., & Nejdl, W. (2011). Efficient incremental near duplicate detection based on locality sensitive hashing. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (pp. 1185–1190). ACM. https://www.researchgate.net/publication/221464486

Gupta, L., Tariq, J., Yessirkepov, M., Zimba, O., Misra, D. P., Agarwal, V., & Gasparyan, A. Y. (2021). Plagiarism in non-anglophone countries: a cross-sectional survey of researchers and journal editors. Journal of Korean Medical Science, 36(39).

Henzinger, M. (2006). Finding near-duplicate web pages: A large-scale evaluation of algorithms. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 284–291). ACM. https://doi.org/10.1145/1148170.1148222

Ilham, A. A., Bustamin, A., Aswad, I., & Armin, F. (2020, June). Implementation of clustering and similarity analysis for detecting content similarity in student final projects. In IOP conference series: Materials science and engineering (Vol. 875, No. 1, p. 012039). IOP Publishing.

Khaled, F. & Sabeeh M. Plagiarism Detection Methods and Tools: An Overview. (2021). *Iraqi Journal of Science*, *62*(8), 2771-2783. https://doi.org/10.24996/ijs.2021.62.8.30

Lizunov, P., Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Biloshchytska, S., & Serbin, O. (2021). Development of the combined method of identification of near duplicates in electronic scientific works. Eastern-European Journal of Enterprise Technologies, 4(4), 112.

Mohammadi, H., & Khasteh, S. H. (2018). A fast text similarity measure for large document collections using multi-reference cosine and genetic algorithm. *arXiv preprint arXiv:1810.03102*. https://arxiv.org/abs/1810.03102

Prieur, M., Gadek, G., & Grilheres, B. (2022). Duplicate detection in a knowledge base with PIKA. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022)* (pp. 561–568). SciTePress. https://www.scitepress.org/Papers/2022/107695/107695.pdf

Pudasaini, S., Chhetri, R., Gautam, D., & Joshi, A. (2024). Plagiarism in the age of large language models: A survey. *arXiv preprint arXiv:2407*.13105.

Rakhimova, D., Turarbek, A., & Kopbosyn, L. (2021, April). Hybrid approach for the semantic analysis of texts in the Kazakh language. In Asian Conference on Intelligent Information and Database Systems (pp. 134-145). Singapore: Springer Singapore.

Ruben Comas, Thomas Lancaster, Elvira Curiel, Carmen Touza (2023). Automatic paraphrasing tools: an unexpected consequence of addressing student plagiarism and the impact of COVID in distance education settings Práxis Educativa, Ponta Grossa, v. 18, e21679, p. 1-19, 2023. DOI:10.5212/PraxEduc.v.18.21679.020

Silcock, E., D'Amico-Wong, L., Yang, J., & Dell, M. (2022). Noise-robust de-duplication at scale. *arXiv preprint arXiv:2210.04261*. https://arxiv.org/abs/2210.04261

Tleubayeva, A., & Shomanov, A. (2024). Comparative analysis of multilingual qa models and their adaptation to the kazakh language. *Scientific Journal of Astana IT University*, *19*, 89–97. https://doi.org/10.37943/19WHRK2878

Togmanov, M., Mukhituly, N., Turmakhan, D., Mansurov, J., Goloburda, M., Sakip, A., ... & Koto, F. (2025). KazMMLU: Evaluating Language Models on Kazakh, Russian, and Regional Knowledge of Kazakhstan. arXiv preprint arXiv:2502.12829.

Toiganbayeva, N., Zhumadilov, Y., Abdigaliyev, A., & Khassanov, Y. (2021). KOHTD: Kazakh offline handwritten text dataset. *arXiv preprint arXiv:2110.04075*.

Tolegen, G., Toleu, A., Mamyrbayev, O., & Mussabayev, R. (2020). Neural named entity recognition for Kazakh. *arXiv preprint arXiv:2007.13626*. https://arxiv.org/abs/2007.13626

Wahle, J. P., Ruas, T., Foltýnek, T., Meuschke, N., & Gipp, B. (2022). Identifying machine-paraphrased plagiarism. In M. Smits (Ed.), *Information for a better world: Shaping the global future. iConference 2022. Lecture notes in computer science* (Vol. 13192, pp. 444–456). Cham: Springer. https://doi.org/10.1007/978-3-030-96957-8_34

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, Lidong Zhou (2021). LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*. https://arxiv.org/abs/2012.14740

Yeshpanov, R., Khassanov, Y., & Varol, H. A. (2022). KazNERD: Kazakh named entity recognition dataset. *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 4229–4238.

Zhang, M., Vallis, O., Bumin, A., Vakharia, T., & Bursztein, E. (2023). RETSim: Resilient and efficient text similarity. *arXiv preprint arXiv:2311.17264*. https://arxiv.org/abs/2311.17264

Information about authors:

Svitlana Biloshchytska **–** Doctor of Technical Sciences, School of Artificial Intelligence and Data Science, Astana IT University, bsv@astanait.edu.kz, ORCID: 0000-0002-0856-5474.

Aidos Mukhatayev **–** Candidate of Pedagogical Sciences, Professor, School of General Education Disciplines, Astana IT University, mukhatayev.aidos@gmail.com, ORCID: 0000-0002-0856-5474.

Oleksandr Kuchanskyi **–** Doctor of Technical Sciences, School of Artificial Intelligence and Data Science, Astana IT University, kuchanskyi.o@gmail.com, ORCID: 0000-0003-1277-8031.

Saltanat Sharipova **–** PhD, Center of Competence and Excellence, Astana IT University, saltanat.sharipova@astanait.edu.kz, ORCID: 0000-0001-7267-3261.

Nayla Murat **–** Master student, School of Artificial Intelligence and Data Science, Astana IT University, 242899@astanait.edu.kz.

Zhan Amangeldiyev **–** Master student, School of Artificial Intelligence and Data Science, Astana IT University, 243014@astanait.edu.kz.