IRSTI 20.19.27 https://doi.org/10.26577/jpcsit2025336

Assel Ospan* , **Kanat Auyesbay** ,
Talshyn Sarsembayeva , **Aman Mussa**
Al Farabi Kazakh National University, Almaty, Kazakhstan
*e-mail: asselyaospan@gmail.com

APPLICATION OF RULE-BASED METHOD FOR AUTOMATIC EXTRACTION OF TAGS FROM COLUMN-STYLE PDF-DOCUMENTS

Abstract. This study presents a rule-based hybrid pipeline for the automated extraction of structured metadata from PDF versions of Kazakh-language newspaper articles, focusing on the national newspaper Egemen Qazaqstan. The primary goal is to support the development of a machine-readable knowledge base for future use in training large language models (LLMs) and building an AI-powered assistant for data journalism in Kazakhstan. The pipeline integrates three open-source parsers – pdfminer.six, PyMuPDF, and pdfplumber – to extract key elements such as title, author, date, abstract, text, journal name, and category. To evaluate extraction quality, we compared the results of the automated parser against manually annotated reference files across three real-world issues of the newspaper. The evaluation employed three complementary metrics: Precision, Textual Semantic Similarity (TSS), and Holistic Precision, which jointly assess both exact and semantic matches. The experimental results show that three structured tags – date, journal, and category – achieved perfect Holistic Precision (1.00), while the remaining tags obtained high scores (author 0.93, abstract 0.98, text 0.91, title 0.85), yielding a macro-average Holistic Precision of 0.95. The validated pipeline was then applied to the full corpus of 2,140 newspaper PDFs published between 2017 and March 2025, successfully converting 159,135 articles into structured JSON format. This enriched corpus serves as a foundational knowledge base for Kazakh-language AI systems in journalism and media analysis.

Keywords: PDF parsing, Rule-based extraction, Metadata extraction, Document structure recognition, Text mining, Low-resource language processing, Knowledge base for LLMs.

1. Introduction

Currently, the development of an AI assistant for data journalism in the Kazakh language is gaining particular relevance due to the rapid advancement of large language models (LLMs), such as OpenAI GPT-4 [1], Google Gemini [2], Meta LLaMA [3], and others. One of the key factors determining the effectiveness of such models is the availability of a high-quality corpus in the target language. However, for the Kazakh language, there is a significant shortage of digital and structured texts, especially in the domain of official and analytical journalism. Most previously published materials exist only as scanned or digital PDF versions of newspapers, lacking HTML markup and accessible APIs, which significantly complicates the task of automatic data extraction and annotation.

In this context, it is particularly important to enrich the model's knowledge base with only verified information from trusted sources. Newspaper articles that have undergone editorial and journalistic reviews serve as a reliable foundation for building such corpora. To choose an optimal source, our team consulted professional journalists. Based on their recommendation, archival issues of the newspaper Egemen Qazaqstan [4] were selected as the base corpus – a leading national publication known for its credibility and stylistic quality.

The use of PDF versions of newspapers is due to the fact that the official Egemen Qazaqstan website provides access to archives only from 2017 onward. A significant portion of the materials from 2017 to 2020 was published mainly in print or as PDF files, without HTML versions or machine-readable markup. This limits the ability to automatically collect and structure texts using standard web tools.

As a result, there is a need to apply specialized methods for extracting text and metadata from PDF documents with column layouts, embedded illustrations, and Kazakh fonts. Developing an effective hybrid pipeline to process such materials becomes a key step in building a high-quality corpus

suitable for training Kazakh-language LLMs aimed at analysis and generation tasks in data journalism.

In scientific literature, various approaches to text extraction from PDFs are proposed, primarily divided into rule-based and learning-based methods.

Rule-based methods rely on predefined rules using regular expressions, keywords, positional and typographic cues to identify meaningful text elements. Their main advantages include interpretability, flexibility in adapting to specific documents, and no need for training data. In our prior work, we have repeatedly relied on rule-based and hybrid pipelines to structure data from unstructured sources. In [5], authors introduced TableProcessor, which used rule-driven parsers and schema mapping to interpret statistical tables from the Bureau of National Statistics (Word/Excel) and convert them into JSON suitable for GIS and knowledge-base ingest. In [6], it was proposed an ontology-guided, iterative method that combines deterministic rules with neural NER to classify cells and expand a domain ontology, yielding RDF triples from semi-structured tables. Building on the same design principles, [7] presented Qurma, a modular system that extracts tables from HTML/PDF/images and applies rule-based normalization and semantic interpretation within a Clean Architecture for scalability and domain portability. These experiences directly inform the present newspaper pipeline – specifically our use of regular expressions, keyword lexicons, and positional/typographic cues – while preserving interpretability and eliminating the need for training data. A successful rule-based approach is presented in [8], where Alamoudi et al. extract metadata from PDF books using PDFBox and logical rules to identify key elements such as book title, author name, and ISBN; their rule set encodes deterministic patterns and constraints to maximize accuracy, yielding overall accuracies of 94.62% (training) and 90.27% (test) on their dataset, which underscores the effectiveness of rule-based designs for document-specific metadata extraction. Similarly, [9] describes a method for citation metadata extraction using Hidden Markov Models, classifying token sequences based on reference structure patterns. In the PDFBoT tool [10], the authors propose a system for high-precision extraction of the main text from academic PDFs using HTML replication, formatting analysis, and syntactic cues. The method achieves high accuracy while preserving sentence and paragraph structure, removing auxiliary elements like tables and images.

In contrast, learning-based methods (based on machine learning and neural network architectures) are used to automatically train models to extract data from documents with complex or unstable structures. The authors of [11] proposed an end-to-end neural architecture for image-based sequence recognition, which has been successfully applied to scene text recognition tasks. In the research work [12], an OCR-free Document Understanding Transformer was developed, which uses a transformer architecture to understand document structure without relying on traditional OCR. The authors of [13] introduced TrOCR, a transformer-based optical character recognition model pre-trained on large-scale corpora, enabling high accuracy across a wide range of document types. Additionally, for learning-based methods, researchers created DocLayNet [14], a large annotated dataset for document layout segmentation, which is used to train models for automatic document structure analysis.

Given the limited availability of annotated data, the specific layout of newspapers (columns, fonts, illustrations), and the need for fine-tuned customization for a single source (Egemen Qazaqstan), a rule-based approach proves to be the most rational and effective solution for automatic extraction of structured tags from Kazakh-language PDF documents.

Based on this, we formulate the following research question: how can rule-based methods be adapted for the automatic extraction of semantically significant tags from Kazakh-language newspaper PDFs, considering their complex visual structure and formatting features?

The goal of this study is to develop and implement a hybrid pipeline based on the integration of rule-based tools – pdfminer.six, PyMuPDF, and pdfplumber – aimed at extracting structured elements (url, title, author, date, abstract, text, journal, category) from PDF versions of Kazakh-language newspaper articles.

Why were these particular tools chosen? To justify this, we refer to the study [15], where the authors conducted a comprehensive comparative analysis of ten popular PDF parsing tools across six types of documents: financial reports, manuals, scientific articles, laws and regulations, patents, and government tenders. The results showed that rule-

based tools such as pdfminer.six [16], PyMuPDF [17], and pdfplumber [18] perform well in extracting text from standard documents, while deep learning models such as Nougat [19] and Table Transformer (TATR) [20] significantly outperform them in more complex cases, especially when dealing with mathematical formulas or non-standard tables.

The structure of this paper is as follows: Section II presents the experimental algorithm and data extraction methods from the PDF corpus, including a detailed description of the rule-based pipeline; Section III provides quantitative and qualitative results of the selected parsers; Section IV discusses their strengths and weaknesses in the context of Kazakh-language newspapers; Section V concludes the study and outlines directions for further development of the hybrid solution.

The outcome of this work is the construction of a machine-readable JSON-format corpus suitable for subsequent training of large language models and the development of an intelligent assistant in the field of Kazakh-language data journalism.

2. Materials and methods

For the experimental study, a corpus of PDF documents from the Egemen Qazaqstan newspaper was used. The processing was carried out using the tools PyMuPDF, pdfminer.six, and pdfplumber, in accordance with established computational standards. Based on empirical observations, a set of heuristic rules was developed for the automatic extraction of semantically meaningful tags such as title, author, date, abstract, text, category, and journal.

2.1. Description of dataset

For the purposes of this study, a specialized corpus was compiled, consisting of all PDF issues of the Egemen Qazaqstan newspaper from January 2017 to March 2025 inclusive. A total of 2,140 PDF documents were collected, each representing a full issue of the national newspaper.

Each PDF file contains an average of 16-18 pages, formatted in a multi-column format with illustrations and multilingual inserts. On average, there are about 4.6 full articles per page, so the total number of text units in the corpus is 159,135 articles. These materials cover a wide range of topics – with a total of 208 unique categories. The primary categories include: Economics, Politics, Society, Culture, Education, History, Literature, Religion, Health, Spirituality, Regions, Agriculture, Technology, and Sports.

2.2. Rule-based methodology

This section outlines the rule-based methodology developed for the automatic extraction of semantically meaningful tags and metadata from the PDF versions of the collected newspapers. During the preprocessing stage, text cleaning and normalization methods were applied, including removal of OCR artifacts, merging of hyphenated line breaks, and Unicode normalization.

The core of the approach is a set of heuristic rules that leverage visual, lexical, and structural features to identify key tags such as: title, author, date, category, text, abstract, and journal. Figure 1 visually highlights all the main components corresponding to the target tags, which are subject to automatic extraction using the rule-based method.

The heuristic rules for extracting each tag were manually developed based on the analysis of the structure and visual characteristics of the PDF documents. All applied heuristics and filters are summarized in Table 1. The proposed rule set is tailored for the automatic annotation of newspapers with column-based layout and can serve as a foundation for building an annotated corpus or for subsequent training of machine learning models.

Figure 1 – Example of a newspaper article from the PDF issue of Egemen Qazaqstan dated January 1, 2021, featuring column-based layout.

Table 1 – Heuristic extraction rules for newspaper tags.

№	Tag	Rules	Rules description
Rule 1	title	if 10 < len(text) < 180 and block_index == 0 and span["size"] > 15	The title is identified as the first text block on the page that contains between 10 and 180 characters and is printed in a font size larger than 15 pt.
Rule 2	author	if re.search(r"[А-ЯӘӨҮҰІ][а-яәөүұі]+ [А-ЯӘӨҮҰІ][а-яәөүұі]+", text)	The author is identified by the presence of a full name in the format "First Last," starting with capital letters and matching a predefined regular expression.
Rule 3	date	if re.search(r"\d{1,2} \D+ 20\d{2}", text)	The date is extracted as a string containing a day, month name, and four-digit year (e.g., "15 шілде 2023"), conforming to the "day month year" pattern.
Rule 4	category	if any(keyword in text.upper() for keyword in ["Экономика", "Саясат", "Қоғам", …, "Спорт"] and span["size"] < 11)	The category is recognized by the presence of an uppercase keyword from a predefined list of 208 unique topics.
Rule 5	text	cleaned_text = unicodedata.normalize('NFC', re.sub(r"\s+", " ", re.sub(r"(\w+)-\s*\n\s*(\w+)", r"\1\2", re.sub(r"cid:\d+", "", text.replace('\n', ' ').replace('\r', "")))).strip()) if isinstance(text, str) else ""	The article text is merged from multiple columns into a single string, with hyphenated line breaks removed, extra spaces and line breaks cleaned, and Unicode normalization applied for text standardization.
Rule 6	abstract	abstract = text.split(".")[0] if "." in text else text[:200]	The abstract is extracted as the first sentence (up to the first period); if no period is found, the first 200 characters are used.
Rule 7	journal	if block_index < 3 and span["size"] > 10 and text.isupper() and len(text.split()) <= 4:	The journal name is identified as an uppercase string located in one of the top three text blocks on the page, consisting of no more than four words and printed in a font larger than 10 points.

The rules presented in Table 1 will be further integrated into a hybrid pipeline that leverages PyMuPDF for analyzing the visual structure of the document, pdfminer.six for extracting the main text, and pdfplumber for refining block positions when necessary.

2.2.1 Pipeline architecture

Before designing the architecture of the hybrid pipeline, a comparative evaluation of three popular Python libraries – PyMuPDF, pdfplumber, and pdfminer.six – was conducted to assess their effectiveness in extracting various tags from newspaper-style PDF documents (Table 2).

The results indicate that none of the libraries alone ensures high accuracy across all tag categories: for instance, pdfminer.six performs better at extracting main text and abstracts, PyMuPDF shows the best performance for title and date tags, while pdfplumber demonstrates low robustness when dealing with the multi-layered structure of newspaper layouts. These differences stem from the internal architecture of the libraries: PyMuPDF relies on visual block positioning, pdfminer.six focuses on linear text extraction, and pdfplumber applies character-level parsing.

Based on this comparative analysis, a well-grounded decision was made to integrate all three tools into a single rule-based pipeline, leveraging the strengths of each library to achieve optimal accuracy for individual tag extraction tasks. The proposed pipeline implements a rule-based method for automatic extraction of semantically labeled tags from Kazakh-language newspaper PDF documents (see Figure 2).

Table 2 – Comparison of Tag Extraction Accuracy Using PyMuPDF, pdfplumber, and pdfminer.six.

Tag	PyMuPDF (%)	Pdfplumber (%)	pdfminer.six (%)
Title	90	50	60
Author	70	30	60
Abstract	60	40	80
Text	80	60	90
Date	90	40	40
Category	75	30	60
Journal	80	50	50

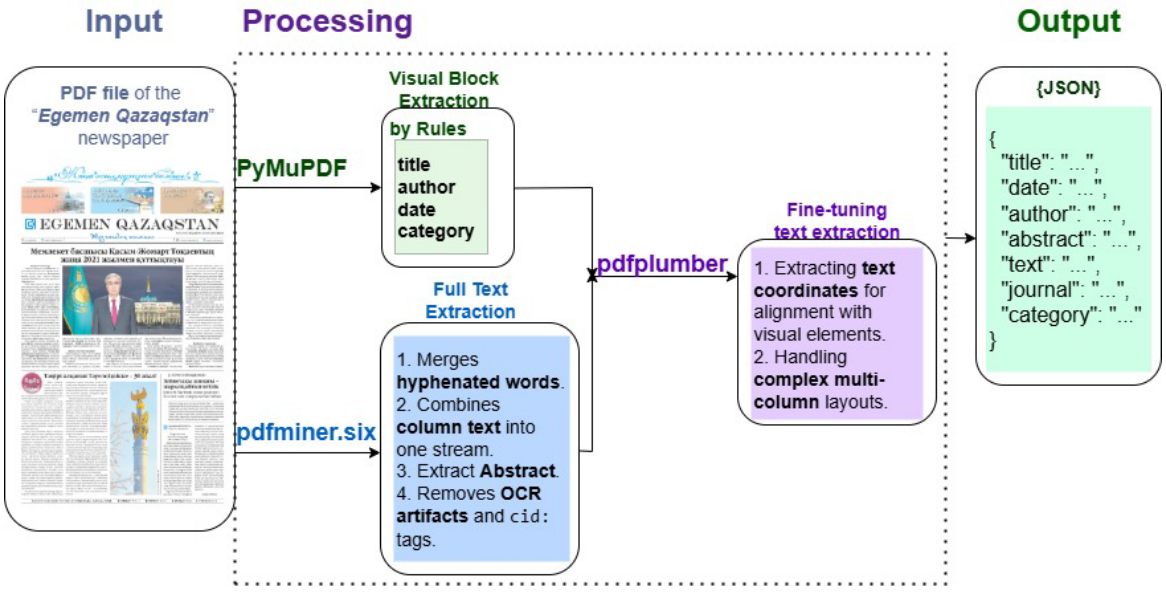

Figure 2 – Hybrid PDF Tagging Architecture using PyMuPDF, pdfminer.six and pdfplumber.

The input to the system is a PDF file representing a print issue of the Egemen Qazaqstan newspaper, featuring a multi-column layout with both textual and graphical elements. The first stage employs the PyMuPDF library to extract visual text blocks by analyzing coordinates, font sizes, and spatial positioning. Based on a set of heuristics, this module identifies and extracts key tags such as title, author, date, and category. To retrieve the full textual content of the articles, the pipeline utilizes pdfminer.six, which disregards visual layout constraints and merges columnar content into a unified text string. At this stage, hyphenated word breaks are resolved, Unicode normalization is applied, and OCR artifacts (e.g., “cid:123”) are removed.

The resulting content is then used to generate the abstract and to extract the journal name based on the top blocks of the page. Additionally, pdfplumber is used for geometry refinement, coordinate verification, and correction of extracted tags in cases with complex visual layouts. This enhances tag-level precision through improved structural parsing. The output is a machine-readable JSON structure containing all extracted tags (url, title, author, date, abstract, text, journal, category), suitable for knowledge base enrichment, language model training, and intelligent Kazakh-language text analysis systems.

2.2.2 Algorithm

The developed algorithm formalizes this pipeline into a sequential metadata extraction procedure: Rules 1–4 are applied to visual blocks via PyMuPDF, followed by Rules 5–6 for text and abstract generation using pdfminer.six, and Rule 7 for journal identification. Pdfplumber serves as an auxiliary module for final verification and correction. This hybrid rule-based pipeline demonstrates the effective integration of three PDF processing tools, ensuring automated and interpretable extraction of textual information from Kazakh-language newspaper documents.

Algorithm: Rule-based Metadata Extraction from PDF Newspapers

```
Input: PDF file D of Egemen Qazaqstan newspaper
Output: JSON object J with fields: url, title, author, date, abstract, text, journal, category
function EXTRACT_METADATA(D)
 V ← extract_visual_blocks(D) using PyMuPDF
 INITIALIZE title, author, date, category
 for each text span T in V do
 T ← clean_and_normalize_text(T)
 if title == "" and apply Rule 1 on T then
 title ← T
 else if author == "" and apply Rule 2 on T then
 author ← T
 else if date == "" and apply Rule 3 on T then
 date ← T
 else if category == "" and apply Rule 4 on T then
 category ← T
 end for
text_raw ← extract_text_pdfminer(D)
text ← normalize_hyphens_and_columns(text_raw) ← apply Rule 5
abstract ← extract_abstract(text) ← apply Rule 6
for each top span T in first 3 blocks of V do
 if journal == "" and apply Rule 7 on T then
 journal ← T
end for
J ← {
 "title": title,
 "author": author,
 "date": date,
 "abstract": abstract,
 "text": text,
 "journal": journal,
 "category": category
 }
J ← refine_with_pdfplumber(J, D)
return J
end function
```

2.2.3 Algorithmic complexity

Let P denote the number of pages, B=|V| the number of visual blocks, S the number of text spans, C the total number of characters, and R=7 the number of heuristic rules. Enumerating blocks across all pages is O(B), while establishing per-page reading order can require O(B log B) in the worst case. Cleaning and normalizing the text for all spans is linear in the total character count, O(C). The single pass that applies Rules 1–4 examines up to S spans until all four fields are filled, and per-span costs are constant for metadata checks tied to span/block attributes (title), O(|T|) for non-pathological regular expressions (author, date), and O(|T|) for uppercasing combined with average O(1) hash-set membership over a fixed 208-label category, which aggregates to O(S+C) in the worst case. Full-text extraction followed by column stitching, de-hyphenation, and whitespace repair is also linear, O(C). Abstract derivation is O(C) in the worst case but typically sublinear due to early termination at the first sentence. Journal identification is restricted by design to the first three blocks and therefore runs in O(1) on average, although it can degrade to O(B) if fallbacks expand the search region. The optional geometry refinement applies local checks and can be bounded by O(B).

Consequently, the algorithm runs in $O(B \log B + S + C)$ per issue in the worst case, which in practice tends toward near-linear O(B+C) thanks to page-zone filtering and early stopping (e.g., title and journal found early; abstract stops at the first period). Viewed per tag, title selection is O(1) per span via metadata with an O(B) worst case if page scans are required; author and date are O(C) due to regex evaluation; category is O(C) using the constant-size lexicon; text and abstract are O(C); and journal is O(1) on average with an O(B) worst case.

2.3. Evaluation Methodology

Evaluation of the performance of rule-based information extraction (IE) systems, unlike machine learning models, does not require a training phase; however, it demands meticulous verification of the correctness of the extracted data. The most common approaches include manual inspection (see Formulas 1–3), semantic similarity measurement between extracted and reference text fragments (see Formula 4), and expert evaluation using a checklist (see Formulas 5–7). These methods are widely applied in scenarios where labeled datasets are unavailable.

Table 3 presents three quality assessment methods for evaluating the performance of rule-based information extraction systems.

Table 3 – Methods and formulas for evaluating the quality of rule-based information extraction (IE) systems.

№	Evaluation method	Formula	Description
1	Manual inspection	For each entity $E=\{w_1,w_2,\ldots,w_k\}$, where w_i are whitespace-separated words, count $\lvert E_{true} \cap E_{pred}\rvert$, then compute $Precision = \frac{\lvert E_{true} \cap E_{pred}\rvert}{E_{pred}}$ (1), $Recall = \frac{\lvert E_{true} \cap E_{pred}\rvert}{E_{true}}$ (2), $holistic\ F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$ (3)	It is used as a baseline quality metric for rule-based approaches in the absence of training data and allows for direct comparison with the gold-standard annotation [21].
2	Textual semantic similarity (TSS)	Modified cosine similarity measure: $Sim(T_1, T_2) = \frac{\overrightarrow{T_1} * \overrightarrow{T_2}}{\lvert\lvert\overrightarrow{T_1}\rvert\rvert * \lvert\lvert\overrightarrow{T_2}\rvert\rvert}$ (4), where $T_1, T_2 \in R^n$ – vector representations of texts, $Sim(T_1, T_2) \in [0,1]$, tags are considered correct if $Sim \geq 0.85$	It evaluates the degree of semantic similarity between the reference and extracted text based on vector representations. A modified cosine similarity measure with a threshold ≥ 0.85 is applied [22].
3	Expert checklist evaluation	1. Exploration: $H_e = \frac{T_e - U_e}{T_e}$ (5), where H_e – homogeneity of vocabulary, T_e – total number of words in the selected entity texts, T_e – the number of unique words in the same texts. 2. Term frequency analysis: $TF.IDF_{t,e} = \frac{f_{t,e}}{\sum_{t' \in e} f_{t',e}} * log\left(\frac{\lvert E\rvert}{\lvert E_t\rvert}\right)$ (6),	The REST-tool is designed to optimize resources in information extraction (IE) tasks by analyzing whether rule-based approaches can be effectively applied to each entity or whether machine learning (ML) methods should be used, involving multiple expert reviewers [23].

Continuation of the table

№	Evaluation method	Formula	Description
		where $f_{t,e}$ – the number of occurrences of term t in the selections of entity e, |E| – total number of entities, $\|E_t\|$ – the number of entities containing the term t. 3. Expert Evaluation: $Precision = \frac{TP}{TP+FP}$ (7), where TP – true positive, FP – false positive. Then do checklist – assessment of the applicability of the rules: *TH: Text Highlights ≥ 25%,* *LH: Linguistic Homogeneity ≥ 10%,* *ER: Entity Recall ≥ 75%,* *EP: Entity Precision ≥ 75%,* If all 4 criteria are met, the rules apply, otherwise – ML.	

In [21], the authors present a system for structured information extraction from scientific texts that combines manual annotation with automated entity extraction. They introduce the *holistic F1* metric, which accounts for both exact matches and semantic similarity between extracted and gold-standard fragments. This methodology emphasizes comprehensive expert evaluation, making it especially applicable to rule-based systems without training data.

The TSS method evaluates the degree of semantic alignment between reference and extracted text using vector representations, applying a modified cosine similarity threshold of ≥ 0.85 [22]. The study in [22] demonstrates that transformer models pre-trained and fine-tuned on clinical data achieve a high correlation with expert judgments (Pearson r ≈ 0.89–0.91), confirming the effectiveness of this approach for measuring semantic similarity in clinical texts.

In [23], the authors developed a rule-based Evaluation and Support Tool (REST) to optimize resource allocation in information extraction tasks. Their workflow began with fully manual expert annotations, followed by the introduction of metrics to assess rule adequacy, and the use of expert checklists to determine whether rule-based methods should be applied to specific entity types.

After reviewing these three effective evaluation metrics (see Table 3), the authors of this study selected two – Precision and TSS. We compute TSS with the Sentence-Transformers library [24] using the multilingual model sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 [25]. The model maps sentences to 384-dimensional embeddings, then it is used the model's Hugging Face AutoTokenizer (BERT-style) with max_seq_length = 128 and apply means pooling over token embeddings to obtain sentence vectors. Similarity is measured by cosine similarity between the reference and extracted text. We count a pair as semantically equivalent when TSS ≥ 0.85. This threshold was selected on a held-out, manually labeled set by maximizing F1 and was robust in the 0.80–0.88 range. The model is multilingual (≈50 languages, including Kazakh), so no additional language adaptation was required; we only apply Unicode NFC normalization prior to encoding [24, 25]. We then integrated these two metrics into a unified Holistic Precision formula for a more comprehensive assessment of the extracted tags (see formulas 8 and 9):

$$holistic\ Precision_{tag} = \\ = \alpha * Precision_{tag} + (1-\alpha) * TSS_{tag} \quad (8)$$

$$\alpha \in [0,1] \quad (9)$$

where $Precision_{tag}$ – proportion of exactly matching extracted values, TSS_{tag} – average cosine similarity between reference and extracted text, α – weighting coefficient (in our case 0.5 for equal contribution of both metrics).

Both metrics are normalized to [0,1] and capture complementary error modes: Precision penalizes string-level deviations – critical for structured fields (e.g., date, category, journal) – consistent with standard IE/NER evaluation where scoring is performed at the entity level [26]. TSS rewards meaning-preserving paraphrases – critical for free-text fields – so that variants of title, abstract, or full

text that differ lexically yet preserve meaning are not unduly penalized. These practices reflect long-standing IE/NER evaluation conventions and explain why the two signals are complementary.

In the absence of domain-specific cost asymmetries, we set α=0.5 as a neutral prior for three reasons: (i) it treats the two components symmetrically; (ii) it avoids degeneration to a single metric at the extremes $\alpha \in [0, 1]$; and (iii) it is invariant under affine rescalings of either component. This choice aligns with the classical balanced setting in IR/IE, where the Precision measure corresponds to equal weighting ($\beta = 1 \Rightarrow \alpha = \frac{1}{2}$), and with the REST rationale to balance exact and semantic signals in rule-based IE evaluation [23].

Holistic Precision provides a more flexible and realistic estimate of extraction quality, particularly useful when analyzing rule-based systems where semantically correct but not form-identical extractions are possible.

2.4. Rights, licensing, and intended use

Copyright in the original newspaper content remains with the rights holders. No full article texts or images are redistributed by this work.

Published metadata is licensed under a CC BY-NC 4.0 (Creative Commons Attribution–NonCommercial 4.0) license. Any reuse must be related to this article/dataset, remain non-commercial, and comply with applicable law and publisher policies. We honor copyright holder takedown requests; requests must include the publication date, author, and title so that relevant records can be removed.

The dataset and code are intended for academic research and benchmarking of information-extraction methods for low-resource languages and to support development of a non-commercial AI assistant for journalists. Use of the original article texts for model training lies outside the scope of this release and requires appropriate permissions.

3. Results

As part of the experimental evaluation of the proposed rule-based approach, we manually annotated a sample of 113 articles obtained from three issues of the Egemen Qazaqstan newspaper. The annotation covered key semantic tags, including title, author, abstract, text, date, journal, and category. To assess the extraction quality, we applied our proposed integrated metric – holistic Precision (see Formula 8) – which takes into account not only the exact match (Precision), but also the TSS between the reference and extracted fragments. This dataset was used as a test sample for the experiment. All training and testing files are available in the GitHub repository at https://github.com/AsselOspan/Rule_based_PDF2JSON. Figure 3 presents an example of the original print layout of the newspaper and the corresponding gold-standard JSON file created through manual annotation.

To evaluate the quality of data extraction, a comparative analysis was conducted between the results obtained using the rule-based parsing approach and manually annotated gold-standard files. The comparison was performed across each of the seven key tags: title, author, abstract, text, date, journal, and category. During the experiment, the values of Precision, TSS (Textual Semantic Similarity), and the final metric Holistic Precision – which integrates both indicators – were calculated. The results of the experiment based on three selected PDF files are presented in Tables 4–7.


EGEMEN QAZAQSTAN

Атажұртқа алғашқы көш

1991 жылғы 17 наурызда келді

Апатты үйлер

мұнайлы Атыраудың абыройына сын емес пе?

Қызылордалық полицейлер «делдалдықпен» айналыса ма?

(a)


```
[{
  "title":"Атажұртқа алғашқы көш 1991 жылғы 17 наурызда келді",
  "date":"17 Наурыз, 2021",
  "author":"Бекен ҚАЙРАТҰЛЫ",
  "abstract":"Ел Тәуелсіздігінің табыстары туралы жиі айтып та,
  "text":"Отыз жылға жуық уақыт ішінде атажұртқа оралған қандасы
  "journal":"Egemen Qazaqstan",
  "category":"Қазақстан"
},
{
  "title":"Апатты үйлер мұнайлы Атыраудың абыройына сын емес пе?
  "date":"17 Наурыз, 2021",
  "author":"Жолдасбек КӨШЕРБАЙҰЛЫ",
  "abstract":"Қазақ жеріндегі «қара алтынның» алғашқы тамшысы 18
  "text":"Содан бері 122 жыл өткенімен, қазыналы аймақтан мұнай
  "journal":"Egemen Qazaqstan",
  "category":"Аймақтар"
},
{
  "title":"Қызылордалық полицейлер «делдалдықпен» айналыса ма?",
  "date":"17 Наурыз, 2021",
  "author":"Мұрат ЖЕТПІСБАЕВ",
  "abstract":"Соңғы уақытта әлеуметтік желіде қоғам белсендісі,
  "text":"Өткен жылдың 24 желтоқсанында облыс орталығында жол-көл
  "journal":"Egemen Qazaqstan",
  "category":"Аймақтар"
},
```


(b)

Figure 3 – PDF presentation of the Egemen Qazaqstan newspaper in JSON format: (a) original printed page layout; (b) corresponding JSON representation.

Table 4 – Results of the experiment for the newspaper "Egemen Qazaqstan" from January 23, 2020 on three quality indicators.

Tag	Precision	TSS	Holistic Precision
title	0,68	0,87	0,77
author	0,74	0,87	0,80
date	1,00	1,00	1,00
abstract	1,00	1,00	1,00
text	0,88	0,93	0,91
journal	1,00	1,00	1,00
category	1,00	1,00	1,00

Table 5 – Results of the experiment for the newspaper “Egemen Qazaqstan” from January 1, 2021 on three quality indicators

Tag	Precision	TSS	Holistic Precision
title	0,80	0,93	0,86
author	1,00	1,00	1,00
date	1,00	1,00	1,00
abstract	0,98	0,96	0,97
text	0,94	0,93	0,93
journal	1,00	1,00	1,00
category	1,00	1,00	1,00

Table 6 – Results of the experiment for the newspaper "Egemen Qazaqstan" from May 20, 2021 on three quality indicators

Tag	Precision	TSS	Holistic Precision
title	0,87	0,93	0,90
author	1,00	1,00	1,00
date	1,00	1,00	1,00
abstract	0,97	0,96	0,96
text	0,87	0,92	0,90
journal	1,00	1,00	1,00
category	1,00	1,00	1,00

Table 7 – Average results of the experiment

Tag	Precision	TSS	Holistic Precision
title	0,78	0,91	0,85
author	0,91	0,96	0,93
date	1,00	1,00	1,00
abstract	0,98	0,97	0,98
text	0,90	0,93	0,91
journal	1,00	1,00	1,00
category	1,00	1,00	1,00
Average	**0,94**	**0,97**	**0,95**

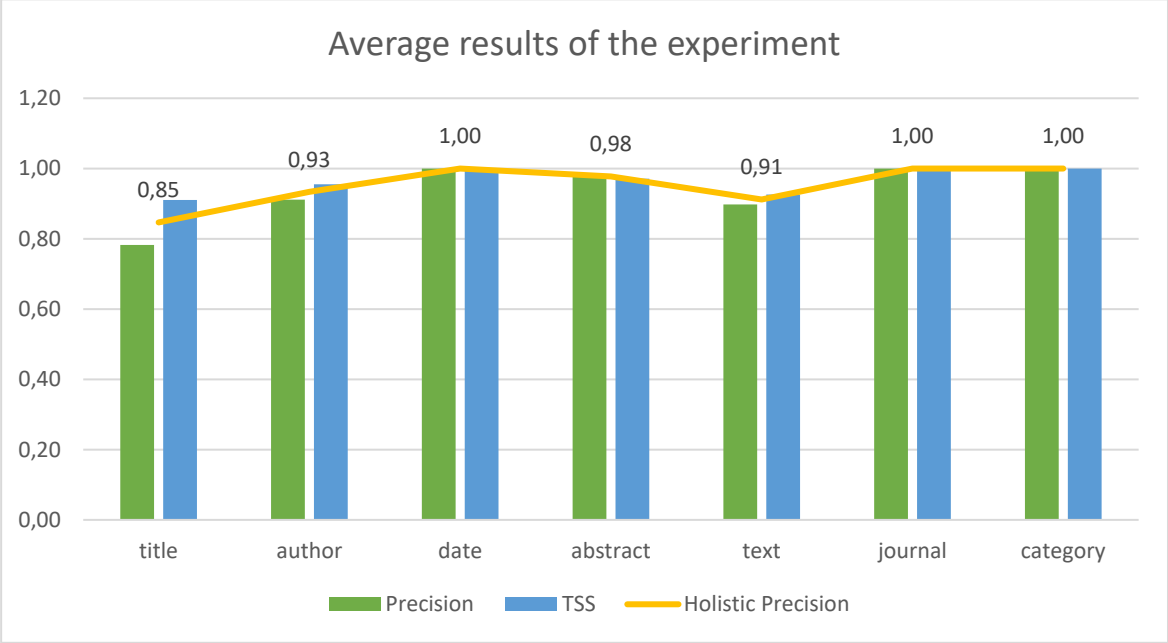

Figure 4 – Graph of averaged results of the experiment on assessing the quality of semantic tag extraction using the rule-based method.

As shown in Figure 4, the TSS and Holistic Precision metrics for most tags fall within the 0.95–1.00 range, indicating high accuracy of the proposed rule-based approach. Slightly lower Precision values were observed for the title tag, which can be attributed to the variability of headlines and minor stylistic or syntactic differences.

The results demonstrate excellent extraction accuracy for formalized tags such as date, journal, and category, where the integrated Holistic Precision metric reached a perfect score of 1.000. High semantic similarity was also observed for less formal fields such as title, abstract, and text, with Holistic Precision values ranging from 0.85 to 0.98.

Following the experimental evaluation and validation of the method's effectiveness, the approach was scaled to the entire PDF corpus of the Egemen Qazaqstan newspaper from 2017 to March 2025. As a result of automated extraction and structuring, 2,140 PDF files were processed, and 159,135 articles were successfully converted into JSON format (see Table 8).

Table 8 – Number of Egemen Qazaqstan PDF issues from 2017 to March 2025 and the number of articles successfully converted to JSON format.

Year	Number of PDF newspapers	Number of articles in JSON
2017	272	17 147
2018	285	16 000
2019	265	22 647
2020	280	21 630
2021	301	22 365
2022	337	28 855
2024	340	26 491
2025	60	4 000
Total number	**2 140**	**159 135**

The resulting collection of structured materials will serve as a knowledge base for the development of an intelligent AI assistant tailored to data journalism tasks.

4. Discussion

The goal of this study was to determine how rule-based methods can be adapted for the automatic extraction of semantically meaningful tags from PDF issues of Kazakh-language newspapers–despite their complex visual structure and formatting–and whether high-quality results can be achieved. Although there exist widely used PDF parsing methods [16–20], our experiments showed that they underperformed on Kazakh texts and struggled with multi-column layouts. We therefore targeted these challenges specifically, since building a knowledge base for downstream NLP inevitably involves such complex files. The methods we designed are detailed in the Methodology section, with custom heuristic rules as the key to improving extraction quality.

We evaluated extraction quality with three complementary metrics – Precision, TSS, and their integrated Holistic Precision – each addressing a distinct need in Kazakh-language processing. Precision quantifies exact matches between extracted and reference tags and is crucial for highly formalized elements (e.g., date, category, journal), where correctness can be measured without contextual interpretation. TSS is especially important for Kazakh's rich morphology and agglutination: if strict Precision alone were used, many partially matching yet semantically valid outputs (e.g., for title, abstract and text) would be incorrectly deemed wrong. Holistic Precision combines both signals to provide a balanced view of system effectiveness, which is particularly relevant for rule-based systems in low-resource settings where large annotated datasets are unavailable.

A comparative analysis of tag-level results across the three component parsers – PyMuPDF, pdfplumber, and pdfminer.six – shows that no single component dominates all tags. PyMuPDF performs best for title, date, category, and journal thanks to its use of layout cues, while pdfminer.six leads on abstract and text due to its linear text analysis that is more resilient to multi-level layouts; pdfplumber is the weakest on most tags. Our Hybrid Rule-Based Pipeline, which integrates all three modules, achieves the highest overall performance, with Holistic Precision ($\alpha = 0.5$) reaching 91–100% on six of the seven tags (and 85% on title). Figure 5 summarizes these outcomes by reporting per-tag Holistic Precision for the three single-component variants (bars) and the proposed hybrid pipeline (line).

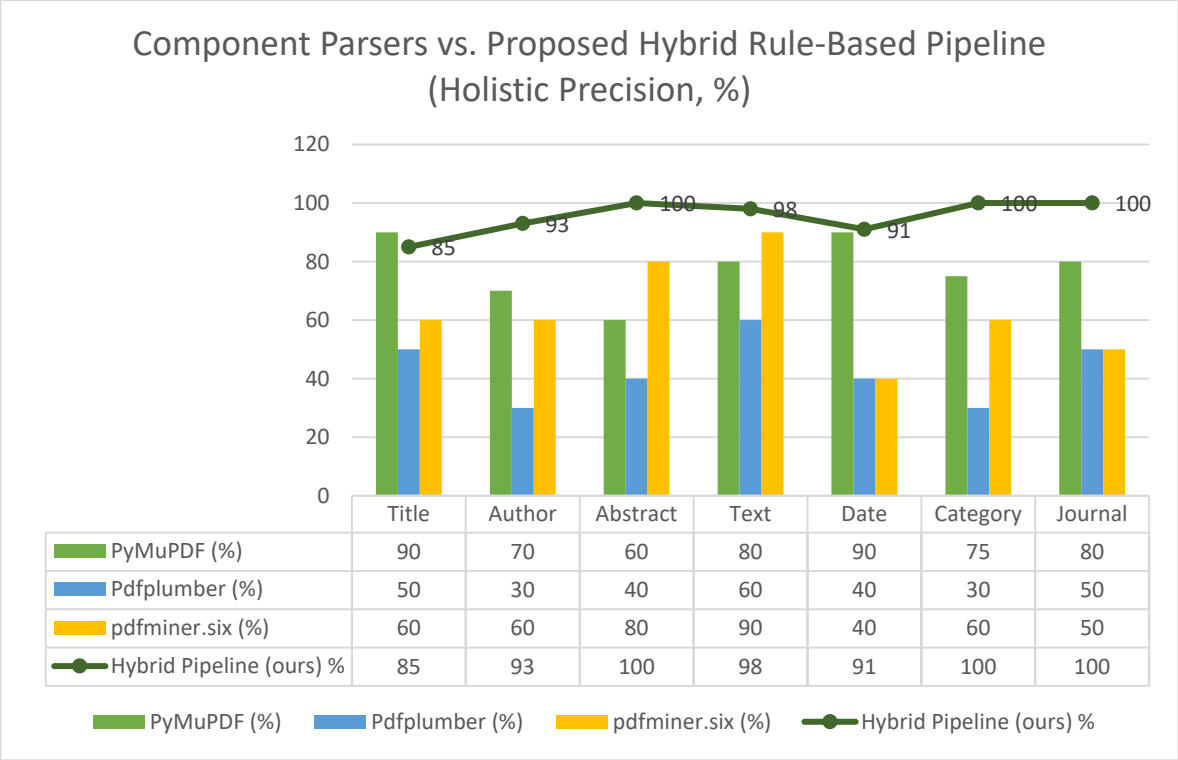

Figure 5 – Results reporting per-tag Holistic Precision for the three parsers (bars) – PyMuPDF, pdfplumber, and pdfminer.six – and the proposed hybrid rule-based pipeline (line).

These results align with prior findings [21, 23] on the advantages of rule-based approaches for documents with predictable visual structure in low-resource settings. The Egemen Qazaqstan newspaper, with its consistent multi-column layout and recurring placement of key elements, is a suitable testbed for such heuristics.

Nonetheless, potential bias introduced by heuristic rules should be considered. Because the pipeline relies on hand-crafted heuristics over visual, lexical, and structural cues (Rules 1–7; Fig. 1; Table 1), systematic biases may arise. The title rule (first block, large font, 10–180 chars) can over-split short briefs or miss multi-line/small-caps headlines; the author regex (two capitalized tokens) may under-detect three-part names, initials, hyphenated surnames, or mixed-script forms; the date pattern (dd Month yyyy) can miss numeric formats or variant month spellings; the category rule (uppercase keyword list of 208 topics, small font) may bias counts toward frequent uppercase rubrics and under-represent rare/lowercase labels; the text merger (de-hyphenation, column stitching, NFC) can over-join captions or sidebars; the abstract heuristic (first sentence / first 200 chars) may be non-representative for quotes or very short items; and the journal rule (uppercase in top blocks) can confuse mastheads or slogans with the outlet name. Multilingual inserts and dense layouts amplify these risks. In evaluation terms, such effects tend to

increase Precision errors on structured tags (boundary/format mismatches), while TSS may mask minor string differences in free-text fields. We mitigate these risks via Unicode normalization (NFC), near-duplicate removal at the (title, date) level, lexicon expansion, and layout refinement with pdfplumber, but residual skew is possible across issues and years and should be considered when interpreting per-tag results.

Some limitations follow directly from these observations. For example, relatively lower Precision for the title tag (avg. 0.78) indicates sensitivity to stylistic and linguistic variation. Newspaper headlines often feature creative phrasing, non-standard punctuation, or unusual syntax, making fixed-template extraction challenging. Although the final Holistic Precision for title reaches 0.85 due to compensation via semantic similarity, this highlights opportunities for improvement – e.g., integrating lightweight ML components to capture more flexible patterns.

Scaling the method to the full Egemen Qazaqstan corpus (2017–March 2025) confirmed robustness and practicality: 2,140 issues processed and 159,135 articles converted to JSON demonstrate suitability for building large machine-readable corpora and knowledge bases. Such resources are vital for training Kazakh LLMs and developing AI assistant for data journalism.

Finally, several operational constraints remain. Manual rule design requires occasional maintenance as layout/editorial policies evolve. Despite combining three libraries (PyMuPDF, pdfminer.six, pdfplumber), throughput may lag behind optimized ML-based pipelines in large-scale deployments. The current version also lacks automatic segmentation for long multi-page articles and pre-processing for scanned PDFs with heavy OCR noise.

Future work includes expanding the tag set (e.g., subtitles, image captions), adding active learning to partially automate rule configuration, and developing a hybrid architecture that couples rule-based logic with transformer-based models for more flexible post-processing. Evaluating the pipeline on additional multilingual news corpora will further illuminate its generalizability.

In summary, the proposed rule-based method achieves high accuracy and strong interpretability on Kazakh-language newspapers, offering a reliable foundation for structured corpora, AI applications, and knowledge-base enrichment in low-resource settings.

5. Conclusions

In this study, a hybrid rule-based pipeline was developed and experimentally validated for the automatic extraction of semantically annotated tags from Kazakh-language newspaper PDF documents. The system combines the strengths of three libraries – PyMuPDF (visual layout analysis), pdfminer.six (linear text extraction), and pdfplumber (structure refinement) – and utilizes a set of manually crafted heuristic rules to identify key metadata fields: title, author, date, abstract, text, journal, and category.

Experimental evaluation on a sample of 113 manually annotated articles demonstrated high values for Precision, TSS, and the integrated Holistic Precision metric (averaging 0.95), confirming the reliability and accuracy of the proposed method. Particularly strong results were achieved for the date, journal, and category tags (1.00), as well as high semantic similarity for the title and abstract fields. Following this validation, the method was scaled to the entire Egemen Qazaqstan corpus from 2017 to March 2025, successfully extracting and structuring 159,135 articles from 2,140 PDF files.

The research goal was achieved: an effective and scalable pipeline was developed for structured data extraction from Kazakh-language PDFs. The study also provided a clear answer to the research question – rule-based methods, when adapted to the visual and linguistic characteristics of newspaper layouts, proved highly suitable for automatic tag extraction under conditions of limited annotated data.

The proposed approach lays a strong foundation for building a machine-readable Kazakh-language corpus, which is strategically important for training large language models (LLMs) and developing an AI assistant focused on data journalism tasks.

Acknowledgments

The authors express their sincere gratitude to the doctoral students of the Faculty of Journalism for their invaluable assistance in collecting and organizing the archival PDF documents of the Egemen Qazaqstan newspaper from 2017 to March 2025.

Funding

This research was funded by the Science Committee of the Ministry of Science and Higher

Education of the Republic of Kazakhstan grant number BR24993001 “Creation of a large language model (LLM) to maintain the implementation of Kazakh language and increase the technological progress”.

Author Contributions

Conceptualization, A.O. and T.S.; Methodology, A.O.; Software, A.M.; Validation, A.O., A.M. and K.A.; Formal Analysis, T.S.; Investigation, K.A.; Resources, K.A.; Data Curation, K.A.; Writing – Original Draft Preparation, A.O.; Writing – Review & Editing, T.S. and K.A.; Visualization, A.M.; Supervision, K.A.; Project Administration, T.S.; Funding Acquisition, T.S.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al., “Language models are few-shot learners,” arXiv preprint arXiv:2005.14165, May 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

2. A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, et al., “PaLM: Scaling language modeling with pathways,” arXiv preprint arXiv:2204.02311, Apr. 2022. [Online]. Available: https://arxiv.org/abs/2204.02311

3. H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, et al., “LLaMA 2: Open foundation and fine-tuned chat models,” arXiv preprint arXiv:2307.09288, Jul. 2023. [Online]. Available: https://arxiv.org/abs/2307.09288

4. V. Barakhnin, M. Mansurova, I. Grigorieva, O. Kozhemyakina, and A. Ospan, “TableProcessor: The tool for the analysis and the interpretation of web tables to create the geo knowledge base of Kazakhstan,” in Artificial Intelligence in Models, Methods and Applications, AIES 2022, Studies in Systems, Decision and Control, vol. 457, Cham, Switzerland: Springer, 2023. [Online]. Available: https://doi.org/10.1007/978-3-031-22938-1_15

5. M. Mansurova, V. Barakhnin, A. Ospan, and R. Titkov, “Ontology-driven semantic analysis of tabular data: An iterative approach with advanced entity recognition,” Appl. Sci., vol. 13, no. 19, pp. 10918, 2023. https://doi.org/10.3390/app131910918

6. A. B. Nugumanova, K. S. Apayev, Y. M. Baiburin, M. Mansurova, and A. G. Ospan, “QURMA: A table extraction pipeline for knowledge base population,” J. Math., Mech. Comput. Sci., vol. 114, no. 2, 2022. https://doi.org/10.26577/JMMCS.2022.v114.i2.08

7. “Egemen Qazaqstan – Official Republican Socio-Political Newspaper of Kazakhstan.” Accessed: Jul. 21, 2025. [Online]. Available: https://egemen.kz/

8. A. Alamoudi, A. Alomari, S. Alwarthan, and A. Rahman, “A rule-based information extraction approach for extracting metadata from PDF books,” ICIC Express Lett., Part B: Appl., vol. 12, no. 2, pp. 121–132, Feb. 2021. [Online]. Available: https://www.researchgate.net/publication/347948003

9. E. Hetzner, “A simple method for citation metadata extraction using hidden Markov models,” in Proc. 8th ACM/IEEE-CS Joint Conf. Digital Libraries, JCDL '08, pp. 280–284, 2008. https://doi.org/10.1145/1378889.1378937

10. C. Yu, C. Zhang, and J. Wang, “Extracting body text from academic PDF documents for text mining,” arXiv preprint arXiv:2010.12647, Oct. 2020.

11. B. Shi, X. Bai, and C. Yao, “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 11, pp. 2298–2304, Nov. 2016.

12. G. Kim, T. Hong, M. Yim, J. Nam, J. Park, and J. Yim, et al., “OCR-free document understanding transformer,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Cham, Switzerland: Springer, 2022, pp. 498–517.

13. M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, et al., “TrOCR: Transformer-based optical character recognition with pre-trained models,” in Proc. AAAI Conf. Artif. Intell., vol. 37, pp. 13094–13102, 2023.

14. B. Pfitzmann, C. Auer, M. Dolfi, A. S. Nassar, and P. Staar, “DocLayNet: A large human-annotated dataset for document-layout segmentation,” in Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Mining, KDD '22, pp. 3743–3751, 2022. https://doi.org/10.1145/3534678.3539043

15. N. S. Adhikari and S. Agarwal, “A comparative study of PDF parsing tools across diverse document categories,” arXiv preprint arXiv:2410.09871, Oct. 2024. [Online]. Available: https://arxiv.org/abs/2410.09871

16. pdfminer.six, “PDF parser for Python.” [Online]. Available: https://github.com/pdfminer/pdfminer.six

17. PyMuPDF, “Python bindings for the MuPDF library.” [Online]. Available: https://pypi.org/project/PyMuPDF

18. pdfplumber, “Tool for extracting text, tables, and metadata from PDFs.” [Online]. Available: https://github.com/jsvine/pdfplumber

19. L. Blecher, G. Cucurull, T. Scialom, and R. Stojnic, “Nougat: Neural optical understanding for academic documents,” arXiv preprint arXiv:2308.13418, Aug. 2023. [Online]. Available: https://arxiv.org/abs/2308.13418

20. Microsoft, “Table Transformer (TATR): Transformer-based table detection model.” [Online]. Available: https://github.com/microsoft/table-transformer

21. J. Dagdelen, A. Dunn, S. Lee, et al., “Structured information extraction from scientific text with large language models,” Nat. Commun., vol. 15, p. 1418, 2024. https://doi.org/10.1038/s41467-024-45563-x

22.X. Yang, X. He, H. Zhang, Y. Ma, J. Bian, and Y. Wu, “Measurement of semantic textual similarity in clinical texts: Comparison of transformer-based models,” JMIR Med. Inf., vol. 8, no. 11, p. e19735, 2020. https://doi.org/10.2196/19735

23.G. Bazin, X. Tannier, F. Adda, A. Cohen, A. Redjdal, and E. Kempf, “Development of the user-friendly decision aid Rule-based Evaluation and Support Tool (REST) for optimizing the resources of an information extraction task,” arXiv preprint arXiv:2506.13177, Jun. 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2506.13177

24.N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” *Proceedings of EMNLP-IJCNLP*, 2019.

25.Sentence-Transformers. *Model card: sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2* (multilingual, 384-dim, BERT backbone, default max_seq_length 128). Hugging Face, accessed 22.08.2025.

26.Manning, C. D. (2010). *Information extraction & named entity recognition* (CS224N lecture slides). Stanford University. Available at: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1106/handouts/InfoExtract-cs224n-2010-1up.pdf

Information about authors

Assel Ospan is a senior lecturer at the Department of Artificial Intelligence and Big Data, al-Farabi Kazakh National University (Almaty, Kazakhstan, assel.ospan@kaznu.edu.kz). Her research focuses on the development of large language models for the Kazakh language, intelligent information extraction, and knowledge base construction. She actively participates in national AI research initiatives and has authored several publications on NLP and data journalism. ORCID iD: 0000-0002-1860-6997.

Kanat Auyesbay is the Dean of the Faculty of Journalism at Al-Farabi Kazakh National University (Almaty, Kazakhstan, kanat.auyesbay@kaznu.edu.kz). He is a journalist-educator who bridges the fields of media and higher education. Dr. Auesbay holds a Candidate of Philological Sciences degree (equivalent to PhD) and has extensive experience in both media production and academic leadership. As a recipient of the Bolashak International Scholarship, Kanat Auesbay completed a research and teaching internship at the University of East Anglia, UK (Norwich, 2013–2014). He served as Chairman of the State Attestation Commission at the Faculty of Journalism and Political Science of L.N. Gumilyov Eurasian National University (2023–2024). Since 2018, he has been a corresponding member of the Kazakhstan Academy of Pedagogical Sciences and a member of the Educational-Methodical Association under the Republican Educational-Methodical Council (ROƏK) for Journalism and Information (2019–2021). He has also served on the expert commission for training specialists abroad under the Bolashak program and has supervised and reviewed numerous theses and doctoral dissertations in media studies. ORCID iD: 0009-0001-3529-9888

Talshyn Sarsembayeva is a a senior lecturer at the Department of Artificial Intelligence and Big Data, al-Farabi Kazakh National University (Almaty, Kazakhstan, talshyn.sagdatbek@kaznu.edu.kz). Her work focuses on the integration of artificial intelligence and data processing tools in journalistic practice. She has contributed to projects involving the structuring of large-scale media archives and the development of AI-assisted systems for Kazakh-language content. ORCID iD: 0000-0001-7668-2640.

Aman Mussa is a research assistant at the Department of Artificial Intelligence and Big Data, al-Farabi Kazakh National University (Almaty, Kazakhstan, mussa.aman0519@gmail.com). He is engaged in the development of rule-based and hybrid NLP pipelines, with a focus on Kazakh-language PDF processing. His work supports large-scale knowledge base generation for intelligent assistants in data journalism. ORCID iD: 0009-0001-9972-7677.

Submission received: 28 July, 2025.
Revised: 24 September, 2025.
Accepted: 24 September, 2025.