

РЕФЕРАТ

на магістерську дисертацію

виконану на тему: «Модифікований метод оптимізації парсингу інформації
періодичних видань з використанням нейромереж»

Студент: Симонов Богдан Євгенович

Актуальність теми. З розвитком цифрових технологій і стрімким зростанням обсягів текстової інформації, що публікується у періодичних виданнях (новинні ресурси, журнали, блоги тощо), виникає нагальна потреба у створенні ефективних методів автоматизованого вилучення даних. Традиційні алгоритми парсингу, які базуються на фіксованих правилах і шаблонах, часто не справляються із завданням обробки неструктурованої та динамічної інформації, що ускладнює отримання точних і своєчасних результатів. Інтеграція методів глибинного навчання з використанням нейромережових моделей, зокрема архітектур LSTM і трансформерів, відкриває нові можливості для підвищення якості вилучення структурованих даних із текстів. Реалізація інтерактивного уточнення результатів через зовнішній сервіс ChatGPT API забезпечує додатковий рівень адаптивності та дозволяє коригувати недоліки початкового вилучення даних. Таким чином, розробка модифікованого нейромережевого методу оптимізації парсингу є своєчасним та актуальним завданням як з наукової, так і з практичної точки зору.

Об'єкт дослідження. Об'єктом дослідження є процеси та алгоритми автоматизованого вилучення інформації з текстових документів періодичних видань.

Предмет дослідження. Предметом дослідження є нейромережові моделі та алгоритми оптимізації парсингу, а також технології інтегрованого аналізу текстових даних з використанням зовнішніх API для уточнення результатів.

Мета роботи. Метою дисертації є розробка ефективного модифікованого нейромережевого методу оптимізації парсингу інформації, що дозволяє автоматично вилучати структуровані дані з періодичних видань із підвищеною точністю та адаптивністю до змін у форматах вхідних текстів. Для досягнення цієї

мети передбачено розробку програмного забезпечення, здатного комбінувати традиційні методи вилучення даних із сучасними підходами глибокого навчання та інтеграцією додаткового уточнення через ChatGPT API.

Наукова новизна. Запропоновано інтегрований підхід, який поєднує переваги класичних методів парсингу з адаптивними можливостями нейромережових архітектур. Особливістю пропонованого методу є застосування модифікованої LSTM-моделі для аналізу тексту з урахуванням контексту та семантичних залежностей, а також інтеграція зовнішнього сервісу ChatGPT API, що дозволяє в режимі реального часу уточнювати результати вилучення даних. Запровадження механізмів зворотного зв'язку та адаптивного налаштування гіперпараметрів сприяє підвищенню якості вилучення інформації, що підтверджує наукову новизну розробленого підходу.

Практична цінність. Практична цінність дослідження підтверджується можливістю застосування запропонованого методу в інформаційно-аналітичних системах, медіа-платформах та бізнес-додатках, де швидкість і точність вилучення даних є критично важливими. Автоматизація процесу парсингу дозволяє зменшити витрати часу на ручну обробку інформації, мінімізувати вплив людського фактора та забезпечити об'єктивний аналіз текстових даних. Результати дослідження можуть бути використані для створення систем, що здійснюють оперативний моніторинг новин, аналіз інформаційних потоків і підтримку прийняття рішень у сфері управління інформацією.

Апробація роботи. Апробація дослідження проводилася на двох етапах:

1. Романкевич В. О., Поліщук О. П., Симонов Б. Є. Модифікований нейромережовий метод оптимізації парсингу інформації періодичного видання // Прикладна математика та комп'ютинг ПМК' 2024 : збірник тез доповідей Сімнадцятої конференції магістрантів та аспірантів (20–22 листопада 2024 р. – Київ, Україна). – Київ, 2024. – С. 418–423.

2. Симонов Б. Є., Романкевич В. О., Поліщук О. П. Модифікований нейромережовий метод оптимізації парсингу інформації періодичних видань// XI Міжнародна науково-технічна Internet-конференція «Сучасні методи, інформаційне,

програмне та технічне забезпечення систем керування організаційно-технічними та технологічними комплексами»: збірник матеріалів. 27 листопада 2024 р., Київ. – Київ: Національний університет харчових технологій, 2024. – С. 105–106

Структура та обсяг роботи. Дисертація складається зі вступу, чотирьох розділів, висновків, списку використаних джерел та додатків. Загальний обсяг роботи становить 128 сторінок.

Розділ 1 присвячено теоретико-методологічним основам аналізу текстової інформації та штучних нейронних мереж: розглянуто концепції парсингу даних, типові методи видобування інформації, а також фундаментальні принципи побудови та навчання нейронних мереж різних архітектур; проаналізовано специфіку текстів періодичних видань і основні задачі їх автоматизованого аналізу.

Розділ 2 міститиме розробку модифікованого нейромережевого алгоритму оптимізації парсингу періодичних видань: постановку наукової задачі та визначення вимог, аналіз сучасних підходів та ідентифікацію недоліків існуючих моделей, розробку концептуальної моделі нейромережевого алгоритму, а також опис методології оптимізації та критеріїв валідації.

Розділ 3 присвячено програмній реалізації алгоритму з використанням мови Python: розглянуто архітектурні рішення програмного забезпечення, технологічне забезпечення реалізації алгоритму, детальний опис алгоритмічних компонентів, а також тестування, налагодження та оптимізацію програмного коду.

Розділ 4 містить емпіричну оцінку ефективності модифікованого алгоритму та інтерпретацію результатів дослідження: опис організації експериментальної бази даних та методології дослідження, емпіричну валідацію модифікованого алгоритму, порівняльну характеристику з існуючими методами, а також обговорення результатів та перспективи подальших досліджень.

У роботі представлено 5 таблиць, 11 рисунків та 4 додатки, що ілюструють як теоретичні аспекти, так і практичну реалізацію розробленого алгоритму.

Ключові слова: модифікований нейромережевий метод, парсинг інформації, періодичні видання, глибинне навчання, ChatGPT API, оптимізація, автоматизація аналізу текстових даних.

ABSTRACT

for the Master's Thesis

titled “Modified method for optimizing the parsing of periodical publication
information using neural networks ”

Student: Bogdan Symonov

Relevance of the Topic. With the rapid advancement of digital technologies and the exponential increase in the volume of textual information published in periodicals (such as news portals, journals, blogs, etc.), there is an urgent need to develop efficient methods for automated data extraction. Traditional parsing algorithms that rely on fixed rules and templates often fail to process unstructured and dynamically changing information accurately and in a timely manner. The integration of deep learning methods using neural network models—particularly architectures such as LSTM and Transformers—opens up new possibilities for enhancing the quality of extracting structured data from texts. Furthermore, the implementation of interactive result refinement via an external service like the ChatGPT API offers an additional layer of adaptability by correcting the initial data extraction shortcomings. Thus, the development of a modified neural network method for parsing optimization is a timely and pertinent task from both scientific and practical standpoints.

Object of the Research. The object of the research is the processes and algorithms for automated extraction of information from text documents in periodical publications.

Subject of the Research. The subject of the research encompasses neural network models and parsing optimization algorithms, as well as the technologies for integrated text data analysis using external APIs to refine results.

Aim of the Work. The aim of the thesis is to develop an effective modified neural network method for parsing optimization that automatically extracts structured data from periodical publications with enhanced accuracy and adaptability to changes in the format of the input texts. To achieve this, the research involves the development of software that combines traditional data extraction techniques with modern deep learning approaches and integrates additional refinement via the ChatGPT API.

Scientific Novelty. An integrated approach has been developed that combines the advantages of classical parsing methods with the adaptive capabilities of neural network architectures. A distinctive feature of the proposed method is the application of a modified LSTM model for text analysis that takes contextual and semantic dependencies into account, as well as the integration of an external service (ChatGPT API) which enables real-time refinement of the extracted data. The introduction of feedback mechanisms and adaptive hyperparameter tuning gradually improves the extraction quality, thereby confirming the scientific novelty of the proposed approach.

Practical Significance. The practical significance of the research is evidenced by the potential to apply the developed method within information-analytical systems, media platforms, and business applications where the speed and accuracy of data extraction are critically important. The automation of the parsing process reduces the time required for manual data processing, minimizes the influence of human subjectivity, and ensures an objective analysis of textual data. The study's results can be used to create systems for real-time news monitoring, information flow analysis, and decision support in the field of information management.

Validation of the Work. The study was validated in two stages:

1. Romankevich V.O., Polischuk O.P., Symonov B.Y. "Modified Neural Network Method for Optimizing the Parsing of Information from Periodical Publications" // Applied Mathematics and Computing, PMK 2024: Proceedings of the Seventeenth Conference of Master's and Doctoral Students (20–22 November 2024 – Kyiv, Ukraine). – Kyiv, 2024. – pp. 418–423.

2. Symonov B. Y., Romankevych V. O., Polishchuk O. P. Modified Neural Network Method for Optimizing the Parsing of Periodical Publication Information // XI International Scientific and Technical Internet Conference "Modern Methods, Information, Software and Technical Support of Control Systems for Organizational-Technical and Technological Complexes": collection of materials. November 27, 2024, Kyiv. – Kyiv: National University of Food Technologies, 2024. – P. 105–106.

Structure and scope of work. The dissertation consists of an introduction, four chapters, conclusions, a list of sources used and appendices. The total volume of the work is 128 pages.

Section 1 is devoted to the theoretical and methodological foundations of text information analysis and artificial neural networks: the concepts of data parsing, typical methods of information extraction, as well as the fundamental principles of building and training neural networks of various architectures are considered; the specificity of periodical texts and the main tasks of their automated analysis are analyzed.

Section 2 will contain the development of a modified neural network algorithm for optimizing the parsing of periodicals: the formulation of the scientific problem and the definition of requirements, the analysis of modern approaches and the identification of shortcomings of existing models, the development of a conceptual model of the neural network algorithm, as well as a description of the optimization methodology and validation criteria.

Section 3 is devoted to the software implementation of the algorithm using the Python language: the architectural solutions of the software, the technological support of the algorithm implementation, a detailed description of the algorithmic components, as well as testing, debugging and optimization of the program code are considered.

Section 4 contains an empirical assessment of the effectiveness of the modified algorithm and the interpretation of the research results: a description of the organization of the experimental database and the research methodology, empirical validation of the modified algorithm, a comparative characteristic with existing methods, as well as a discussion of the results and prospects for further research.

The paper presents 5 tables, 11 figures and 4 appendices, illustrating both the theoretical aspects and the practical implementation of the developed algorithm.

Keywords: modified neural network method, information parsing, periodical publications, deep learning, ChatGPT API, optimization, automated text analysis