УДК 37.09.32 : 681.3

О. М. Ткаченко
Б. С. Дема

ВИКОРИСТАННЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ДЛЯ АВТОМАТИЗОВАНОГО ВИЛУЧЕННЯ ДАНИХ ІЗ НЕСТРУКТУРОВАНИХ РЕЗЮМЕ

Вінницькій національний технічний університет

Анотація

Розглянуто підхід до автоматизованого аналізу та вилучення ключових сутностей із неструктурованих резюме з використанням великих мовних моделей (LLM). Показано, що традиційні методи парсингу на основі регулярних виразів поступаються інтелектуальним моделям у здатності розуміти контекст та складні професійні описи. Описано процес перетворення тексту резюме у структурований JSON-формат, що включає ідентифікацію навичок, досвіду та кваліфікації.

Ключові слова: великі мовні моделі, парсинг резюме, вилучення сутностей, обробка природної мови, структуризація даних, штучний інтелект у рекрутингу.

Abstract

A practical approach to automated extraction of key entities from unstructured resumes using Large Language Models (LLMs) is considered. It is shown that traditional parsing methods based on regular expressions are inferior to intelligent models in their ability to understand context and complex professional descriptions. The process of converting resume text into a structured JSON format is described, including the identification of skills, experience, and qualifications.

Keywords: Large Language Models, resume parsing, Named Entity Recognition, natural language processing, data structuring, AI in recruiting.

Вступ

Ефективна робота сучасних систем підбору персоналу залежить не лише від алгоритмів ранжування, а й від якості вхідних даних. Оскільки кандидати надають резюме у довільних форматах (PDF, DOCX) та з різною структурою викладу, виникає критична потреба у точному парсингу – перетворенні вільного тексту в структуровану базу даних [1].

Традиційні системи часто помиляються при аналізі нестандартних макетів або специфічних галузевих термінів. Використання великих мовних моделей дозволяє ідентифікувати сутності (Named Entity Recognition) на основі глибокого розуміння семантики тексту, що забезпечує набагато вищу точність у порівнянні з методами, що базуються на жорстких правилах [2].

Результати дослідження

Практичний процес інтелектуального парсингу резюме з використанням LLM починається з етапу декомпозиції документа на окремі текстові блоки. На відміну від класичних методів, де кожне слово аналізується ізольовано, великі мовні моделі здатні виявляти приховані зв’язки між описаними обов’язами та реальними компетенціями. Це дозволяє системі не просто знаходити ключові слова, а й правильно інтерпретувати ієрархію професійного досвіду, розрізняючи основні технології та допоміжні інструменти.

Ключовим аспектом дослідження є застосування спеціалізованих інструкцій (Prompts) для моделі, які змушують її ігнорувати нерелевантну інформацію та зосереджуватися на вилученні конкретних сутностей: дати роботи, назви посад, перелік Hard Skills та рівень володіння мовами. Використання архітектури трансформерів дозволяє моделі коректно обробляти складні випадки, такі як перерви у стажі або суміщення посад, що раніше вимагало значного втручання людини. Отриманий у результаті обробки структурований масив даних дозволяє миттєво порівнювати профілі сотень кандидатів за заданими параметрами.

Важливим етапом є реалізація механізму Few-Shot Learning, де моделі надається кілька прикладів ідеально розпарсених резюме для задання необхідного тону та формату виводу. Це забезпечує стабільність результатів незалежно від мови оригіналу чи графічного оформлення документа (наявність колонок, таблиць або інфографіки). Крім того, інтегрований ШІ-модуль виконує автоматичну валідацію вилучених даних, звіряючи назви компаній та технологій із внутрішніми професійними словниками, що мінімізує ризик занесення некоректної інформації до бази даних. Такий

багаторівневий підхід до обробки тексту гарантує, що рекрутер отримає максимально чисті та структуровані дані, готові для подальшого аналітичного опрацювання та ранжування.

Далі вилучені дані перетворюються у структурований формат, такий як JSON, що дозволяє легко інтегрувати їх у CRM-системи або бази даних для подальшого аналізу. Перевагою LLM є можливість автоматичного перекладу або уніфікації назв технологій (наприклад, розуміння того, що "FastAPI" та "Django" належать до категорії "Python Frameworks"), що значно спрощує роботу рекрутера [3]. На відміну від систем попередніх поколінь, інтелектуальний парсинг на основі LLM здатний виділяти навіть "м'які навички" (Soft Skills), аналізуючи опис реалізованих проєктів. Приклад процесу інтелектуального підбору та обробки даних за допомогою ШІ на рис. 1. Це створює надійну інформаційну базу для подальших етапів оцінювання кандидатів.

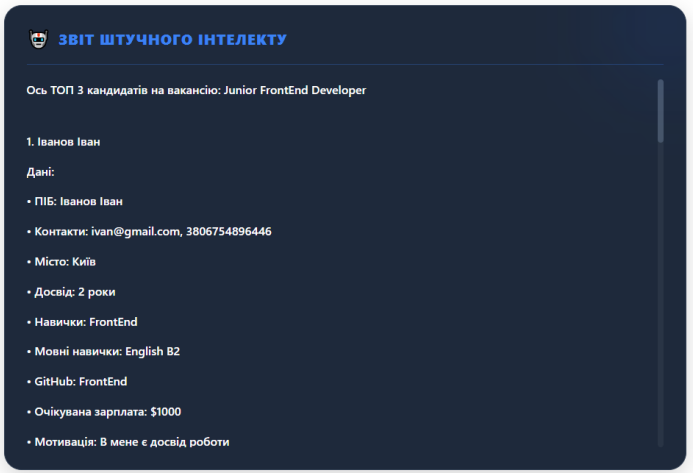

Рис. 1. Процес інтелектуального підбору та обробки даних за допомогою ШІ

Висновки

Впровадження великих мовних моделей для обробки неструктурованих резюме дозволяє суттєво оптимізувати адміністративні процеси рекрутингу. Запропонований підхід гарантує високу точність вилучення професійних сутностей та зручну структуризацію даних у межах єдиного інтелектуального інструменту. У перспективі використання LLM для парсингу сприятиме підвищенню загальної конкурентоспроможності компаній завдяки швидкій та якісній цифровізації вхідних потоків кандидатів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Zhu F. et al**.** Large Language Models for Information Extraction: A Survey // IEEE Access. – 2024. – Vol. 12. – P. 21045–21060. – DOI: 10.1109/ACCESS.2024.3361245.
2. Kaur R., Kautish S**.** Intelligent Resume Parsing and Information Extraction Using Deep Learning and NLP // International Journal of Intelligent Systems and Applications in Engineering. – 2023. – Vol. 11, No. 4. – P. 158–169.
3. Zhang L. et al. LLM-Recruiter: Empowering Recruitment with Large Language Models // arXiv. – 2024. – DOI: 10.48550/arXiv.2402.10543.

Ткаченко Олександр Миколайович – к.т.н., доцент кафедри програмного забезпечення, Вінницький національний технічний університет, e-mail: alextk1960@gmail.com

Дема Богдан Сергійович – студент групи 3ПІ-22б, Факультет інформаційних технологій та комп’ютерної інженерії, Вінницький національний технічний університет, м. Вінниця, e-mail: bogdan.dema24@gmail.com

Tkachenko Oleksandr Mykolaiovych – Candidate of Technical Sciences, Associate Professor of the Department of Software, Vinnytsia National Technical University, Vinnytsia.

Dema Bohdan Serhiiovych – student of the group 3PI-22b, Faculty of Information Technology and Computer Engineering, Vinnytsia National Technical University, Vinnytsia.