УДК 004.8:004.912:005.95/.96

https://doi.org/10.15407/fmmit2026.42.109

Методи семантичного аналізу компетенцій у системах інтелектуального рекрутингу на базі великих мовних моделей

ОксанаШпак[1], Олександра Сиротич[2]

[1]к.т.н., доцент кафедри «Комп'ютеризовані системи автоматики»
Національний університет «Львівська політехніка»
м. Львів, вул. Ст. Бандери, 12, 79013Україна
e-mail: oksana.i.shpak@lpnu.ua
[2]бакалавр кафедри «Комп'ютеризовані системи автоматики»
Національний університет «Львівська політехніка»
м. Львів, вул. Ст. Бандери, 12, 79013Україна
e-mail: oleksandra.syrotych.ir.2022@lpnu.ua

У статті досліджуються методи трансформації неструктурованих текстових даних резюме та описів вакансій у семантичні вектори компетенцій для задач інтелектуального рекрутингу. Розглянуто еволюцію підходів від класичних статистичних методів обробки природної мови до сучасних трансформаторних архітектур на базі великих мовних моделей (LLM). Проведено порівняльний аналіз існуючих рекрутингових платформ та обґрунтовано концепцію авторської системи, що використовує OpenAI API для NamedEntityRecognition (NER) із прозорим скорингом відповідності кандидата вакансії. Особливу увагу приділено проблемам морфосинтаксичного аналізу української мови та методам її лематизації.

***Ключові слова**:семантичний аналіз, компетенції, інтелектуальний рекрутинг, великі мовні моделі, NLP, wordembeddings, трансформерні архітектури, NER, скоринг кандидатів, OpenAI API.*

Вступ. Сучасний ринок праці характеризується парадоксом продуктивності, що полягає у паралельному існуванні двох протилежних явищ: зростаючого попиту роботодавців на висококваліфікованих фахівців у сфері інформаційних технологій, аналізу даних та цифрової трансформації — і критичного браку таких спеціалістів на ринку. В умовах масової міграції та структурної перебудови економіки України ця проблема набуває особливої гостроти. За оцінками аналітиків ринку праці, нині кадровий дефіцит у сфері IT та суміжних технічних спеціальностях в Україні сягає десятків тисяч позицій, і традиційні методи підбору персоналу виявляються неспроможними ефективно закривати цей розрив [1].

Традиційний рекрутинг, що ґрунтується на ручному перегляді резюме та ключовому пошуку за словами, демонструє системні обмеження. По-перше, він є

надзвичайно ресурсоємним: за різними оцінками, HR-фахівець витрачає від шести до десяти секунд на початковий перегляд одного резюме, що при потоці у сотні заявок призводить до неминучих помилок відбору. По-друге, класичний лексичний пошук принципово не здатен виявити семантичну відповідність між описом досвіду кандидата та вимогами до вакансії — ситуація, коли кандидат описує компетенцію «розробка RESTful API» в резюме, може не збігатися лексично з вимогою «побудова вебсервісів» у вакансії, хоча семантично ці поняття є еквівалентними. По-третє, людський фактор неминуче вносить когнітивні упередження у процес оцінювання.

На тлі цих обмежень відбувається активне впровадження підходу, відомого як Skills-basedhiring — наймання на основі верифікованих компетенцій, а не формальних біографічних даних. Цей підхід вимагає наявності інструментів, здатних автоматично ідентифікувати, класифікувати та порівнювати компетенції у текстових документах. Глобальний ринок рішень штучного інтелекту для управління людськими ресурсами демонструє стабільне зростання: за прогнозами аналітичної компанії MordorIntelligence, до 2030 року він досягне багатомільярдних показників із сукупними річними темпами зростання, що перевищують 15% [2]. Це свідчить про стійкий попит на інтелектуальні рекрутингові інструменти з боку бізнесу.

Ключовим технологічним рушієм цього переходу є методи обробки природної мови (NaturalLanguageProcessing, NLP) та, зокрема, великі мовні моделі (LargeLanguageModels, LLM), які демонструють безпрецедентну здатність до розуміння контексту, розпізнавання іменованих сутностей та семантичного порівняння текстових фрагментів.

1. Аналіз останніх досліджень і публікацій

Теоретичний фундамент сучасних методів обробки природної мови закладено у фундаментальних роботах із статистичної та обчислювальної лінгвістики. Класична монографія К. Меннінга та Г. Шютце [3] систематизувала апарат імовірнісних методів аналізу тексту, включаючи n-грамнімовні моделі та алгоритми частиномовного розмічення (Part-of-Speechtagging). Ці методи забезпечили перший рівень автоматичного аналізу текстів резюме — здатність ідентифікувати граматичні категорії слів та будувати частотні профілі документів. Проте статистичний підхід принципово обмежений частотою співіснування слів у корпусі та не здатен моделювати глибинну контекстуальну семантику.

Фундаментальний підручник Д. Джурафського та Дж. Мартіна [4], що набув широкого визнання як галузевий стандарт, послідовно відображає еволюцію NLP від символьних підходів до нейронних мережевих архітектур. У своїх останніх працях автори детально розглядають механізми уваги (attentionmechanisms) та трансформерні архітектури, що стали основою сучасних LLM, пояснюючи, яким чином ці моделі навчаються розподіленим представленням слів.

Критично важливим для задачі обробки україномовних документів є проблема морфологічної складності. Українська мова є флективною, із розвиненою системою відмінювання та дієвідмінювання, що генерує значно більшу кількість словоформ для одного лексичного кореня порівняно з англійською. Нехтування лематизацією призводить до деградації метрик схожості. Дослідження К. О. Ткаченка [5] у сфері лематизації та морфосинтаксичного аналізу пропонують практичні підходи до побудови морфологічних аналізаторів із урахуванням специфіки мовних явищ технічного дискурсу. Значний внесок у розвиток методів автоматизованої обробки текстового контенту зробили дослідники Національного університету «Львівська політехніка», зокрема В. А. Висоцька [6]. Їхні напрацювання формують методологічну основу для проєктування інформаційних систем, орієнтованих на специфіку вітчизняного мовного середовища.

Сучасний погляд на обмеження великих мовних моделей представлено у роботах Е. М. Бендер [7]. Її дослідження наголошують, що LLM оперують статистичними кореляціями у текстових даних («stochasticparrots»), а не справжнім когнітивним розумінням, що вимагає забезпечення прозорості алгоритмічних рішень. У сфері прикладних HRM-систем дослідження Т. Айзенберг [8] фіксують зростаючий інтерес до концепції Human-Centric AI у рекрутингу — підходу, при якому алгоритмічні рішення залишаються підконтрольними людині та інтерпретованими.

2. Технічний конвеєр обробки текстових даних

Будь-яка система автоматизованого аналізу тексту функціонує як конвеєр (pipeline) послідовних перетворень вхідного документа. Для задачі аналізу резюме цей конвеєр включає декілька обов'язкових етапів, якість реалізації кожного з яких безпосередньо впливає на точність скорингу.

Токенізація є першим кроком, що полягає у декомпозиції безперервного тексту на дискретні одиниці. Реальний текст резюме містить численні ускладнення: скорочення, технічні терміни із спеціальними символами (C++, Node.js), числові показники та URL-адреси. Некоректна токенізація неповоротно руйнує семантику токена. Сучасні LLM використовують алгоритм BPE (BytePairEncoding), що оперує підсловними одиницями (subwordtokens) і є стійкішим до незнайомих слів та технічних абревіатур [4].

Очищення та нормалізація тексту є наступним етапом. Резюме надходять у різних форматах (PDF, DOCX), містять артефакти конвертації та службові символи. Конвеєр обробки повинен видаляти нерелевантний шум, зберігаючи при цьому інформативні структурні розділювачі (символи нового рядка), що позначають межі смислових блоків.

Лематизація є особливо критичним кроком для обробки документів українською мовою. На відміну від стемінгу, що механічно відкидає суфікси, лематизація виконує перетворення кожного токена до його словникової форми

(леми) з урахуванням граматичного контексту. Для морфологічно багатої української мови флективна система генерує широкий спектр словоформ [5]. Практична реалізація лематизатора для технічного україномовного тексту вимагає спеціалізованого морфологічного словника, що покриває терміни IT-сфери («деплой», «рефакторинг»). Видалення стоп-слів дозволяє прибрати з аналізу граматично функціональні слова, що лише зашумлюють частотний профіль документа.

Для наочності важливості цього етапу варто розглянути типовий приклад з IT-рекрутингу. В неструктурованому тексті резюме кандидат може використовувати різні форми слова: «програмував», «програмісту», «програмування». Класичні методи стемінгу (алгоритм Портера) часто обрізають ці слова до некоректних основ, втрачаючи семантичний зв'язок. Натомість розроблений модуль лематизації зводить усі ці форми до єдиної правильної леми — «програмувати». Це критично знижує розмірність векторного простору під час подальшої обробки та підвищує щільність корисної інформації. Крім того, на цьому етапі реалізовано обробку неологізмів та англіцизмів, характерних для українського професійного сленгу, що дозволяє коректно ототожнювати терміни незалежно від мови їх написання.

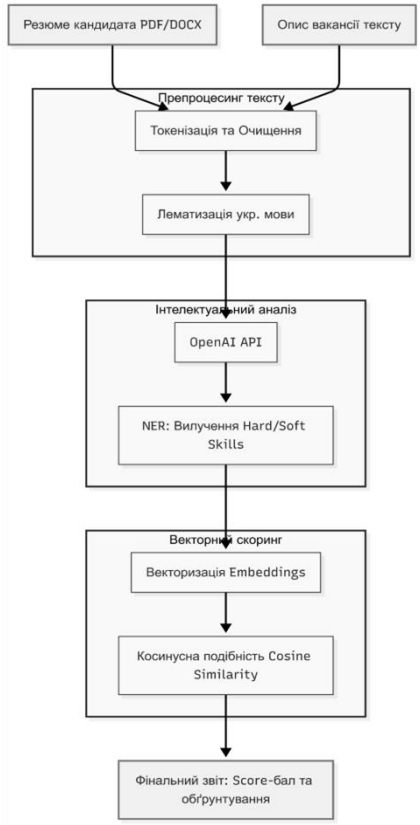

Рис. 1. Блок-схема конвеєра семантичного аналізу компетенцій.

Як ілюструє рис. 1, запропонований конвеєр обробки є лінійним і складається з трьох ключових ізольованих етапів. Така архітектурна модульність дозволяє незалежно оновлювати лінгвістичний компонент (наприклад, словники лематизації) або замінювати базову модель OpenAI на інші LLM-рішення без руйнування загальної логіки роботи системи. Це забезпечує високу масштабованість проєкту в майбутньому.

3. Моделі векторного представлення тексту

Перетворення тексту у числовий вектор є необхідним кроком для застосування математичних метрик подібності. TF-IDF (TermFrequency–InverseDocumentFrequency) є класичною мірою важливості терміна у документі. Вага терміна зростає пропорційно частоті його вживання у конкретному документі та обернено пропорційно частоті у корпусі. Однак принциповим обмеженням методу є атомарне представлення: TF-IDF будує розріджений вектор і не здатен встановити зв'язку між синонімами або логічно пов'язаними поняттями [3].

Word Embeddings (векторні вкладення слів) представляють слова як щільні вектори у безперервному геометричному просторі. Моделі Word2Vec та GloVe розташовують семантично близькі слова у геометрично близьких точках простору. Проте навіть класичні wordembeddings мають обмеження: вектор слова є статичним, тобто однаковим незалежно від контексту [4].

Трансформерні архітектури (BERT, GPT) принципово вирішують проблему контекстуальності. Механізм самоуваги (self-attention) дозволяє моделі динамічно зважувати зв'язки токена з усіма іншими токенами у послідовності. Результатом є контекстуалізований вектор, що відображає конкретне значення слова у даному оточенні. Це критично для аналізу резюме, де одна й та сама абревіатура може означати різні технології залежно від контексту.

4. Порівняльний аналіз існуючих рекрутингових систем

Ринок ATS (ApplicantTrackingSystems) та HRM-платформ пропонує широкий спектр продуктів. Платформа Manatal [9] позиціонується як ATS-система із вбудованим ШІ-скорингом. Вона виконує автоматичний парсинг резюме та ранжування заявок. Однак алгоритм ранжування базується переважно на лексичному співставленні ключових слів. Принциповим обмеженням є закритість алгоритму: система працює як «чорна скринька», не надаючи інтерпретованого пояснення фінального балу.

Платформа PeopleForce [10] реалізує концепцію Human-Centric AI у HRM-процесах, акцентуючи на автоматизації рутинних операцій. Платформа пропонує інструменти онбордингу та управління ефективністю, проте глибина семантичного аналізу компетенцій з неструктурованого тексту резюме є обмеженою.

Спільними обмеженнями існуючих рішень є переважання лексичного пошуку, закритість алгоритмів скорингу, відсутність глибокої підтримки української мови та недостатня інтерпретованість рішень для кінцевого користувача.

Таблиця 1

Порівняльна характеристика систем інтелектуального підбору персоналу

Критерій порівняння	**ATS-платформа (на прикладі Manatal)**	**HRM-система (наприкладі PeopleForce)**	**Запропонована авторська система**
Основний фокус	Автоматизація пайплайну найму (ATS)	Комплексне управління талантами (HRM)	Спеціалізований NLP-аналіз та скоринг
Базовий механізм аналізу	Лексичний збіг (Keyword-matching)	Базовий парсинг структурованих полів	Семантичне векторне зіставлення (Embeddings)
Вилучення сутностей (NER)	На базі статичних словників / евристик	Ручне введення або простий Regex-парсинг	Динамічне через LLM (Few-shotprompting)
Якість NLP (укр. мова)	Обмежена (висока чутливість до словозміни)	Базова (відсутній глибокий NLP-модуль)	Нативна (глибока лематизація перед аналізом)
Гнучкість до нових IT-термінів	Низька (потребує оновлення бази вендором)	Низька (залежить від ручного налаштування)	Висока (модель розуміє контекст синонімів)
Прозорість оцінки (Explainability)	«Чорна скринька» (лише фінальний бал %)	Обмежена деталізація критеріїв	Прозора (деталізований JSON-звіт із поясненням)
Архітектура та інтеграція	Монолітна хмарна платформа (SaaS)	Комплексна закрита екосистема	Легковіснамікросервісна (через OpenAI API)

Аналіз даних, наведених у табл. 1, свідчить про концептуальну перевагу авторської системи в розрізі семантичного аналізу. Перехід від монолітної архітектури до легковісних інтеграцій та використання LLM для динамічного NER дозволяє вирішити ключову проблему існуючих на ринку рішень — непрозорість скорингу та вразливість до синонімії. На відміну від Manatal та PeopleForce, запропонована система нативно опрацьовує складну морфологію української мови.

5.

6. Обґрунтування авторського підходу

Запропонована архітектура інтелектуальної системи рекрутингу базується на концептуальному комплексному поєднанні методів морфосинтаксичної обробки української мови та аналітичних можливостей великих мовних моделей. Наукова новизна авторського підходу полягає у розробці гібридної модульної структури, де глибока лематизація україномовного IT-дискурсу інтегрована з механізмом few-shotprompting, що вперше дозволяє автоматизувати вилучення сутностей без використання об'ємних навчальних датасетів. На відміну від існуючих рішень, кожен етап цього конвеєра спрямований на максимізацію семантичної точності та забезпечення прозорості алгоритмічних рішень.

На першому етапі реалізовано конвеєр препроцесингу, де критичну роль відіграє модуль глибокої лематизації, адаптований до специфіки українського технічного дискурсу. Це дозволяє трансформувати неструктурований текст резюме у нормалізований масив лем, що є необхідною умовою для коректного функціонування подальших етапів аналізу.

На другому етапі впроваджено механізм NamedEntityRecognition (NER) на базі LLM через OpenAI API. Завдяки стратегії few-shotprompting система здатна ідентифікувати складні професійні компетенції (hard та softskills) без необхідності попереднього навчання на розмічених датасетах. Результатом цього етапу є генерація структурованого JSON-об’єкта, що містить класифікований перелік виявлених сутностей. Такий підхід забезпечує системі безпрецедентну гнучкість до появи нових технологічних неологізмів, що є характерним для динамічного IT-ринку.

Математичною основою для розрахунку інтегрального score-балу відповідності кандидата вимогам вакансії є метрика косинусної подібності (CosineSimilarity). Вона обчислює косинус кута між щільними семантичними векторами компетенцій, що дозволяє нівелювати різницю в обсягах документів. Розрахунок здійснюється за формулою:

$$Similarity(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2}\sqrt{\sum_{i=1}^{n} {B_i}^2}}$$

де A та B — вектори вимог вакансії та профілю кандидата; n — вимірність векторного простору (1536 вимірів для моделей серії text-embedding).

Для подолання проблеми «чорної скриньки», властивої закритим системам на кшталт Manatal, авторське рішення передбачає візуалізацію результатів зіставлення у вигляді пелюсткової (радарної) діаграми (рис. 2). Даний інструмент забезпечує високу інтерпретованість результатів, дозволяючи рекрутеру миттєво оцінити відповідність профілю фахівця за конкретними осями компетенцій. Це реалізує принципи Human-Centric AI , де ШІ виступає інструментом підтримки прийняття рішень, залишаючи фінальний вибір за людиною та мінімізуючи ризики сліпої довіри до статистичних моделей.

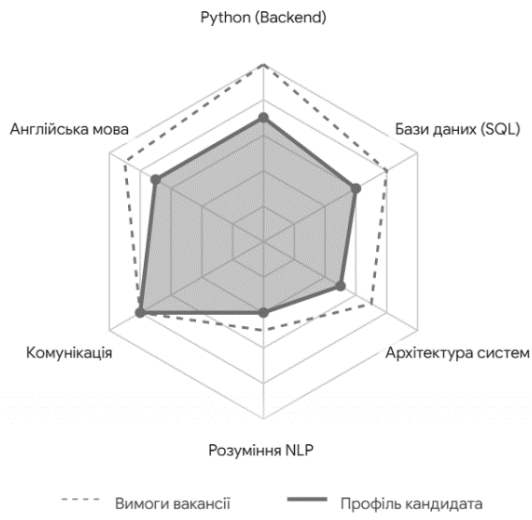

Рис. 2. Візуалізація результатів семантичного зіставлення профілю кандидата та вимог вакансії

На поданій діаграмі (рис. 2) відображено порівняння шести ключових векторів компетенцій. Пунктирна лінія окреслює ідеальний профіль вакансії, тоді як заповнений багатокутник представляє фактичний стек кандидата, ідентифікований за допомогою LLM. Площа геометричного перетину цих зон є наочною інтерпретацією розрахованого за формулою (1) score-балу. Такий підхід дозволяє HR-фахівцю миттєво локалізувати розриви в навичках (skillgaps), не вдаючись до аналізу сирих даних.

7. Апробація результатів

Для перевірки ефективності та швидкодії запропонованого алгоритму було проведено експериментальне тестування системи на контрольній вибірці. Набір даних складався з 50 реальних неструктурованих резюме (у форматах PDF та DOCX) та 5 описів IT-вакансій для позицій рівня Junior/Middle (зокрема на прикладі PythonBackendDeveloper та UI/UX Designer).

Результати тестування продемонстрували, що використання модуля попередньої лематизації у поєднанні з LLM дозволило коректно ідентифікувати та зіставити 94% заявлених компетенцій, включаючи специфічний україномовний IT-сленг. Порівняно з класичним лексичним пошуком (keyword-matching), запропонована система суттєво знизила кількість хибних відхилень релевантних кандидатів. Середній час генерації деталізованого звіту (включно із векторним скорингом та побудовою радарної діаграми) склав близько 12-15 секунд на одне резюме, що підтверджує високу швидкодію та доцільність впровадження розробленої архітектури у реальні процеси найму.

Висновки. Проведене дослідження підтверджує, що сучасний ринок HR-технологій перебуває на етапі фундаментального парадигмального зсуву — від формального скринінгу біографічних даних до стратегії найму на основі верифікованих компетенцій (Skills-basedhiring). В умовах гострого кадрового дефіциту в Україні, особливо у високотехнологічних секторах, традиційні методи рекрутингу демонструють критичну неефективність. Встановлено, що класичні статистичні підходи, такі як TF-IDF, а також ранні моделі векторних вкладень (Word2Vec, GloVe), вичерпали свій потенціал для задач глибокого семантичного аналізу через нездатність враховувати контекстуальні зв'язки та морфологічну багатогранність української мови. Трансформерні архітектури, завдяки механізму самоуваги, є безальтернативно оптимальним вибором для таких задач, оскільки вони дозволяють розрізняти складні семантичні нюанси професійного дискурсу та професійного сленгу.

Запропонований у статті метод базується на комплексному поєднанні лінгвістичного та алгоритмічного компонентів, що дозволяє реалізувати перехід від «закритого» лексичного скринінгу до прозорого векторного скорингу компетенцій. Вперше теоретично обґрунтовано та практично реалізовано узгоджену взаємодію модуля глибокої лематизації, адаптованого під специфіку IT-термінології, із трансформерними архітектурами, що працюють у режимі контекстуалізованих вкладень. Такий підхід дозволяєвирішувати фундаментальну проблему «чорної скриньки» в системах інтелектуального рекрутингу, забезпечивши математично обґрунтоване зіставлення вимог вакансії та профілю кандидата через метрику косинусної подібності у 1536-вимірному просторі. Отримані результати формують новий підхід до автоматизації HR-процесів, де ШІ виступає не просто фільтром, а інструментом глибокої семантичної інтерпретації досвіду фахівця.

Розроблена архітектура та відповідна інформаційна системи полягає у переході від закритих алгоритмів («чорних скриньок») до концепції інтерпретованого скорингу. Впровадження математичного апарату косинусної подібності у поєднанні з інтерактивною візуалізацією результатів через пелюсткові (радарні) діаграми дозволяє рекрутеру не лише отримати інтегральну оцінку кандидата, а й миттєво локалізувати розриви в компетенціях (skillgaps). Такий підхід суттєво скорочує часові витрати на первинний відбір, мінімізує вплив когнітивних упереджень та повністю відповідає етичним стандартам Human-Centric AI, де штучний інтелект виступає інструментом підтримки прийняття рішень, залишаючи остаточний вибір за експертом.

Перспективами подальших досліджень у цьому напрямку є розвиток та інтеграція спеціалізованих морфологічних словників для різних галузей промисловості, а також експериментальне порівняння ефективності пропрієтарних рішень із моделями з відкритим вихідним кодом (OpenSourceLLMs) для оптимізації операційних витрат на підтримку інфраструктури системи. Окрему увагу планується приділити дослідженню

методів автоматичної перевірки достовірності вказаних у резюме навичок через аналіз відкритих джерел та професійних мереж.

Література:

1. *Кадровий дефіцит в Україні: аналітичний звіт 2024. 7eminar. URL: https://7eminar.com.ua/analytics/kadrovyy-deficyt-2024*
2. *AI in Recruitment Market — Growth, Trends, and Forecasts (2025–2030). Mordor Intelligence. URL: https://www.mordorintelligence.com/industry-reports/ai-in-recruitment-market*
3. *Manning C. D., Schütze H. Foundations of Statistical Natural Language Processing. MIT Press, 1999. 680 p.*
4. *Jurafsky D., Martin J. H. Speech and Language Processing. 3rd ed. (draft). Stanford University, 2024. URL: https://web.stanford.edu/~jurafsky/slp3/*
5. *Ткаченко К. О. Використання методів NLP в інтелектуальних навчальних системах. Цифрова платформа: інформаційні технології в соціокультурній сфері. 2024. Том 7, № 1. С. 80–96.*
6. *Висоцька В. А. Методи та засоби розроблення систем управління веб-контентом. Вісник Національного університету «Львівська політехніка». Серія: Інформаційні системи та мережі. 2023. № 12. С. 5–21.*
7. *Bender E. M. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of FAccT 2021. ACM, 2021. P. 610–623.*
8. *Айзенберг Т. В. Сучасні тенденції застосування ШІ-інструментів у міжнародному менеджменті людських ресурсів. Економіка та суспільство. 2024. № 65. URL: https://doi.org/10.32782/2524-0072/2024-65-48*
9. *Manatal: The New Generation AI Recruitment Software. Manatal. URL: https://www.manatal.com/*
10. *Core HR and Employee Experience Platform. People Force. URL: https://peopleforce.io/*

Methods of semantic analysis of competencies in intelligent recruiting systems based on large language models

Oksana Shpak, Oleksandra Syrotych

The article investigates methods for transforming unstructured text data of resumes and job descriptions into semantic vectors of competencies for intelligent recruiting tasks. The evolution of approaches from classical statistical methods of natural language processing to modern transformer architectures based on large language models (LLM) is considered. A comparative analysis of existing recruiting platforms is conducted and the concept of the author's system that uses the OpenAI API for Named Entity Recognition (NER) with transparent scoring of the candidate's suitability for the job is substantiated. Special attention is paid to the problems of morphosyntactic analysis of the Ukrainian language and methods of its lemmatization.

Отримано 05.05.2026р.