

ХЕРСОНСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
(повне найменування вищого навчального закладу)
ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА ДИЗАЙНУ
(повне найменування інституту, назва факультету (відділення))
КАФЕДРА ПРОГРАМНИХ ЗАСОБІВ І ТЕХНОЛОГІЙ
(повна назва кафедри (предметної, циклової комісії))

Пояснювальна записка

до випускної
роботи магістра
(освітній рівень)

на тему: «Розробка системи обробки й аналізу текстової інформації на
основі можливостей LLM»

Виконав: студент групи 6ПР1
спеціальності
121 - «Інженерія програмного забезпечення»
(шифр і назва спеціальності)

Пушин В. О.
(прізвище та ініціали)

Керівник к.т.н., доцент Хохлов В.А.
(прізвище та ініціали)

Рецензент _____
(прізвище та ініціали)

Херсонський національний технічний університет

(повне найменування закладу вищої освіти)

Факультет, відділення Інформаційних технологій та дизайну
Кафедра Програмних засобів і технологій
Освітній рівень бакалавр
Спеціальність 121 – Інженерія програмного забезпечення
(шифр і назва)

ЗАТВЕРДЖУЮ

Завідувач кафедри
Програмних засобів і технологій
к.т.н., доцент Огнєва О.Є.
“ ___ ” _____ 2025 р.

ЗАВДАННЯ

НА ВИПУСКНУ РОБОТУ СТУДЕНТА

Пушина Владислава Олеговича

(прізвище, ім'я, по батькові)

1. Тема роботи «Розробка системи обробки й аналізу текстової інформації на основі можливостей LLM»

керівник роботи _____,
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджена наказом вищого навчального закладу від 15.09.2025 № 416-с

2. Строк подання студентом роботи 04.12.2025

3. Вихідні дані до роботи постановка завдання

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити):

1) Аналіз предметної області.

2) Проектування програмного продукту.

3) Розробка програмного продукту.

4) Тестування та оптимізація програми.

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень)

1) Блок-схеми алгоритмів.

2) _____

3) _____

4) _____

5) _____

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

7. Дата видачі завдання _____

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів виконання роботи	Термін виконання етапів роботи	Примітки
1.	Отримання завдання	15.09.2025	Виконано
2.	Підбір літератури	25.09.2025	Виконано
3.	Аналіз предметної області	29.09.2025	Виконано
4.	Розробка та обґрунтування завдання	3.10.2025	Виконано
5.	Розробка концептуальної моделі	15.10.2025	Виконано
6.	Розробка алгоритму	19.10.2025	Виконано
7.	Проектування програми	27.10.2025	Виконано
8.	Розробка інтерфейсу програми	15.11.2025	Виконано
9.	Тестування програми	28.11.2025	Виконано
10.	Оформлення пояснювальної записки	02.12.2025	Виконано
11.	Захист кваліфікаційної роботи	22.12.2025	Виконано

Студент

(підпис)

Пушин В.О.
(прізвище та ініціали)

Керівник роботи _____
(підпис)

Хохлов В.А.
(прізвище та ініціали)

РЕФЕРАТ

Пояснювальна записка до випускної роботи магістра: 197 с., 16 рис., 4 табл., 2 додатки, 40 джерел.

Об'єкт проектування – інтелектуальна система аналізу текстових даних у сфері електронної комерції.

Предмет проектування – методи, моделі та програмні засоби автоматизованої обробки текстової інформації із застосуванням великих мовних моделей (LLM).

Мета проектування – розробити програмний продукт, здатний автоматично збирати, обробляти, аналізувати та інтерпретувати великі обсяги текстових даних (відгуків, описів, службових метрик Amazon), формувати структуровані рекомендації та оптимізаційні звіти, використовуючи сучасні можливості LLM.

Метод проектування – комплексний аналіз предметної області, математичне моделювання текстових даних, побудова інформаційних потоків системи, розробка алгоритмів веб-скреїпінгу, NLP-обробки, інтеграції з API та LLM, проектування архітектури клієнт–серверного застосунку, тестування та оптимізація програмного забезпечення.

Ключові слова:

LLM, обробка текстів, NLP, Amazon, VoC-аналітика, Return Rate, NCX Rate, web-scraping, Playwright, семантичний аналіз, штучний інтелект, рекомендаційні системи.

АНОТАЦІЯ

Магістерська кваліфікаційна робота присвячена розробці інтелектуальної системи автоматизованої обробки та аналізу текстових даних із застосуванням великих мовних моделей (LLM). Робота містить: вступ, чотири розділи, висновки, список використаних джерел та додатки.

У першому розділі здійснено дослідження предметної області та огляд сучасних методів NLP, статистичних підходів, глибинних нейронних мереж та трансформерів. Проаналізовано особливості текстових даних у електронній комерції, складність їх семантичної інтерпретації та проблеми класичних підходів у порівнянні з LLM. Розглянуто VoC-аналітику, механізми збору даних з Amazon та сучасні комерційні інструменти, визначено їхні обмеження.

У другому розділі сформовано математичну та інформаційну модель системи. Визначено архітектуру модулів, потоки даних ETL, моделі представлення текстів (TF-IDF, embeddings, LLM embeddings), алгоритми виявлення ключових проблем, формування метрик Return Rate / NCX Rate та методи структурованої інтеграції LLM через tool calling. Подано формальний опис алгоритмів кластеризації, узагальнення відгуків та формування рекомендацій.

У третьому розділі спроектовано програмний продукт, що реалізує розроблену модель. Створено архітектуру бекенду (Flask), модулі скрейпінгу Amazon за допомогою Playwright, підсистему взаємодії з Aseller API, модуль аналізу текстів, інструмент генерації таблиць Rufus, а також веб-інтерфейс для перегляду звітів. Описано функціональні та нефункціональні вимоги, включно зі стійкістю до блокувань, масштабованістю та толерантністю до помилок.

У четвертому розділі наведено реалізацію програмних модулів, приклади коду, логіку багатопоточного аналізу та автоматичного моніторингу зміни метрик. Проведено тестування парсингу, аналізу VoC, інтеграції LLM і генерації рекомендацій. Продемонстровано ефективність системи у виявленні

проблем товарів, формуванні аналітичних висновків та оптимізаційних пропозицій.

Розроблений програмний продукт є інтелектуальною аналітичною системою нового покоління, здатною автоматизувати до 90% процесів аналізу текстових даних на великих платформах електронної комерції, зменшувати кількість повернень та підвищувати якість товарних лістингів.

ABSTRACT

The master's qualification work is dedicated to the development of an intelligent system for automated text data processing and analysis using Large Language Models (LLM). The work includes an introduction, four chapters, conclusions, references, and appendices.

The first chapter examines the subject area and provides an overview of modern NLP methods, including statistical approaches, deep neural networks, and transformer-based architectures. It analyzes the specifics of text data in e-commerce, the challenges of semantic interpretation, VoC analytics, data collection from Amazon, and limitations of existing commercial analytics tools.

The second chapter presents the mathematical and information model of the system. It defines the architecture, ETL data flows, models for text representation (TF-IDF, embeddings, LLM embeddings), algorithms for identifying key product issues, calculating Return Rate and NCX Rate, and integrating LLM via tool calling for structured responses and recommendations.

The third chapter focuses on software product design. It describes the backend architecture using Flask, Amazon scraping modules using Playwright, interaction with Aseller API, text analysis modules, Rufus table generation tools, and the web interface for displaying analytical results. Functional and non-functional requirements are formalized.

The fourth chapter describes implementation details, including code examples, multi-threaded analysis workflow, automatic monitoring of metric changes, and testing procedures. The system has demonstrated high efficiency in identifying product issues and generating structured optimization recommendations using LLM.

The developed system serves as a powerful analytical tool capable of automating extensive text analysis tasks in e-commerce, reducing return rates, and improving product listing quality.

ЗМІСТ

ВСТУП	12
Розділ 1. ДОСЛІДЖЕННЯ ТА АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ.....	13
1.1. Опис та загальна характеристика предметної області.....	13
1.1.1. Особливості текстових даних у цифрових середовищах.....	13
1.1.2. Роль автоматизованої обробки текстів у сучасних інформаційних системах.....	14
1.1.3. Інформаційний шум та проблема неструктурованості.....	14
1.1.4. Значення семантичного аналізу для бізнес-процесів.....	15
1.2. Огляд сучасних підходів до обробки текстової інформації.....	16
1.2.1. Статистичні методи NLP.....	16
1.2.2. Методи машинного навчання.....	17
1.2.3. Нейронні мережі та трансформери.....	18
1.2.4. Великі мовні моделі: історія розвитку та ключові принципи.....	19
1.2.5. Обмеження та проблеми традиційних NLP-підходів.....	21
1.3. LLM у сфері електронної комерції.....	21
1.3.1. Особливості текстів у e-commerce.....	21
1.3.2. Аналітика VoC (Voice of Customer) як ключовий інструмент бізнесу.....	21
1.3.3. Типові задачі e-commerce, які вирішують LLM.....	22
1.3.4. LLM у прогнозуванні споживчої поведінки.....	23
1.3.5. Переваги LLM над класичними моделями.....	23
1.4. Збір текстових даних у великих платформах.....	23
1.4.1. Веб-скрейпінг: принципи та значення.....	23
1.4.2. Особливості збору даних з Amazon.....	24
1.4.3. Виклики та обмеження.....	24
1.4.4. Browser automation (Playwright).....	25
1.4.5. Використання проксі та техніки уникнення блокувань.....	25
1.4.6. API-підходи до отримання текстових та службових даних.....	26
1.5. Аналіз існуючих рішень та інструментів.....	26
1.5.1. Комерційні платформи аналізу даних.....	26
1.5.2. Інструменти аналітики відгуків.....	27
1.5.3. Системи оптимізації контенту.....	27
1.5.4. Порівняльний аналіз характеристик.....	27
1.5.5. Виявлені обмеження.....	27
Висновки до розділу 1.....	28
Розділ 2. МАТЕМАТИЧНА ТА ІНФОРМАЦІЙНА МОДЕЛЬ СИСТЕМИ АНАЛІЗУ ТЕКСТОВИХ ДАНИХ.....	31

2.1. Загальна концепція моделі.....	31
2.1.1. Архітектурна схема системи.....	32
2.1.2. Потоки даних: парсинг збереження аналіз генерація інтерфейс.....	33
2.2. Моделювання текстів та їх семантики.....	36
2.2.1. Представлення текстів.....	36
2.2.2. Статистичні та нейромережеві моделі представлення текстів.....	36
2.2.3. Визначення ключових проблем (return reasons, complaints).....	37
2.3. Інтеграція LLM у систему.....	38
2.3.1. Формування промптів (prompt engineering).....	38
2.3.2. Функціональні виклики / tool calling.....	38
2.3.3. Моделювання відповідей LLM та їх структурування.....	39
2.4. Математична модель аналізу товарів.....	39
2.4.1. Формули розрахунку NCX Rate, Return Rate.....	39
2.4.2. Агрегація з API Aseller.....	40
2.4.3. Алгоритм визначення проблемних ASIN-ів.....	40
2.5. Модель автоматичного формування рекомендацій.....	40
2.5.1. Алгоритм узагальнення відгуків.....	40
2.5.2. Модель розпізнавання ключових проблем за текстами.....	41
2.5.3. Автоматичне створення рекомендацій.....	41
2.5.4. Генерація А/В-тестів.....	41
Висновки до розділу 2.....	41
Розділ 3. ПРОЄКТУВАННЯ ПРОГРАМНОГО ПРОДУКТУ.....	44
3.1. Вимоги до системи.....	44
3.1.1. Функціональні вимоги.....	44
3.1.2. Нефункціональні вимоги.....	47
3.2. Архітектура програмного забезпечення.....	50
3.2.1. Загальна структура проєкту.....	50
3.2.2. Архітектура бекенду.....	50
3.2.3. Архітектура фронтенду.....	51
3.2.4. Взаємодія analis.py ↔ rufus.py ↔ HTML.....	52
3.2.5. Інтеграція з OpenAI API.....	52
3.3. Методи та засоби вирішення проблеми.....	53
3.3.1. Вибір та обґрунтування методів вирішення проблеми.....	53
3.3.2. Вибір та обґрунтування засобів вирішення проблеми.....	55
3.4. Проєктування інтерфейсу користувача.....	57
3.4.1. Принципи UI.....	57
3.4.2. Опис основних екранів.....	57

	10
Висновки до розділу 3.....	59
Розділ 4. РОЗРОБКА ТА ТЕСТУВАННЯ ПРОГРАМНОГО ПРОДУКТУ.....	61
4.1. Процес розробки.....	61
4.1.1. Розробка архітектури проєкту.....	61
4.1.2. Розробка системи парсингу Amazon.....	62
4.1.3. Модуль отримання службових даних через API Aseller.....	64
4.1.4. Реалізація аналізу відгуків.....	65
4.1.5. Модуль LLM-аналізу.....	65
4.1.6. Модуль генерації таблиць Rufus.....	66
4.1.7. Розробка веб-інтерфейсу.....	66
4.2. Розробка програмного забезпечення з прикладами коду.....	67
4.2.1. Основна функція run_analysis().....	67
4.2.2. Функція parse_amazon_product().....	68
4.2.3. Функція ask_rufus_prompts_sequence().....	68
4.2.4. Функція monitor_rates().....	68
4.3. Тестування програмного продукту.....	68
4.3.1. Тестування парсингу.....	68
4.3.2. Тестування моніторингу.....	70
4.3.3. Тестування інтеграції з OpenAI.....	70
4.4. Аналіз та оптимізація.....	70
4.4.1. Оптимізація парсингу.....	70
4.4.2. Оптимізація черг.....	70
4.4.3. Оптимізація LLM-викликів.....	70
4.4.4. Оптимізація фронтенду.....	71
4.5. Аналіз отриманих результатів.....	71
4.5.1. Верифікація функціональних компонентів.....	71
4.5.2. Валідація результатів та відповідність поставленим цілям.....	72
4.5.3. Оцінка продуктивності та масштабованості.....	74
4.5.4. Підсумкова оцінка результатів.....	75
Висновки до розділу 4.....	75
ВИСНОВКИ.....	78
ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	81
Додаток А. UML діаграми.....	84
Додаток Б. Тест-кейси.....	85
Додаток В. Вихідний код програмного продукту.....	86

ВСТУП

У сучасну епоху цифрової трансформації обсяги текстових даних зростають експоненційно, охоплюючи найрізноманітніші сфери — від електронної комерції й соціальних мереж до сервісних платформ і корпоративних бізнес-процесів. Текстова інформація стала ключовим джерелом знань про користувацький досвід, тенденції ринку, оцінки продуктів та поведінкові патерни споживачів. Проте її неструктурованість, семантична неоднозначність та високий рівень інформаційного шуму істотно ускладнюють можливість ручного аналізу. Зростання обсягів таких даних потребує інтелектуальних систем, здатних працювати з великими корпусами текстів швидко, точно та контекстуально.

Традиційні методи Natural Language Processing (NLP), хоч і були ефективними в попередні десятиліття, сьогодні вже не справляються з задачами високої складності: глибинною семантичною інтерпретацією, виявленням причинно-наслідкових зв'язків, розумінням емоційних аспектів, узагальненню великих масивів текстів. Поява трансформерної архітектури та подальший розвиток великих мовних моделей (LLM) — таких як GPT, PaLM, LLaMA — відкрили нову еру інтелектуальних систем, що можуть аналізувати та генерувати тексти з точністю, наближеною до людської. Це зробило можливим автоматизований аналіз складних наборів даних у масштабах, які раніше були недосяжними.

Особливо актуальним застосування LLM є в електронній комерції, де щодня генеруються тисячі відгуків, запитань, претензій та службових показників, що відображають якість товарів, поведінку покупців і проблеми лістингів. На маркетплейсах, таких як Amazon, ключові бізнес-показники — Return Rate, NCX Rate, задоволеність клієнтів — безпосередньо залежать від здатності компаній правильно інтерпретувати текстові сигнали, що надходять від користувачів. Саме тому автоматизація VoC-аналітики та глибинного

текстового аналізу стає критично важливою для прийняття рішень щодо оптимізації контенту, підвищення продажів і зменшення повернень.

У межах цієї роботи розроблено інтелектуальну систему, яка поєднує веб-скрейпінг Amazon, використання службових API, методи NLP та можливості LLM для створення повноцінного аналітичного інструменту нового покоління. Система виконує збір, структурування, семантичний аналіз, кластеризацію відгуків, побудову таблиць атрибутів товару, а також формує структуровані рекомендації щодо покращення лістинга та зменшення відсотка повернень.

Актуальність теми визначається швидким розвитком LLM, необхідністю автоматизації трудомістких аналітичних процесів та потребою бізнесу у високоточних інструментах для роботи з великими обсягами текстової інформації. Результати цієї роботи демонструють практичну цінність інтеграції LLM у системи електронної комерції та створюють фундамент для подальших досліджень у галузі інтелектуального аналізу даних.