МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ПОЛІСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет інформаційних
технологій, обліку та фінансів
Кафедра комп'ютерних технологій
і моделювання систем

Кваліфікаційна робота
на правах рукопису

Лиман Юлія Олександрівна
(прізвище, ім'я, по батькові здобувача освіти)

УДК 004.032.26:004.912:811.161.2

КВАЛІФІКАЦІЙНА РОБОТА

WEB-орієнтована технологія розпізнавання текстового контенту інформаційних ресурсів українською мовою
(тема роботи)

122 «Комп'ютерні науки»
(шифр і назва спеціальності)

Подається на здобуття освітнього ступеня бакалавр

Кваліфікаційна робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.

_____ Ю.О. Лиман
(підпис, ініціали та прізвище здобувача вищої освіти)

Науковий керівник
Молодецька Катерина Валеріївна
докторка технічних наук,
професорка

Житомир – 2024

Висновок кафедри _____
за результатами попереднього захисту: _____

Протокол засідання кафедри _____
№ _____ від «_____» _____ 20____ р.

Завідувач кафедри _____

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

«_____» _____ 20____ р.

Результати захисту кваліфікаційної роботи

Здобувач вищої освіти _____ захистив (ла)
(прізвище, ім'я, по батькові)

кваліфікаційну роботу з оцінкою:

сума балів за 100-бальною шкалою _____
за шкалою ECTS _____
за національною шкалою _____

Секретар ЕК

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

АНОТАЦІЯ

Лиман Ю.О. WEB-орієнтована технологія розпізнавання текстового контенту інформаційних ресурсів українською мовою.

Кваліфікаційна робота на здобуття освітнього ступеня бакалавра за спеціальністю 122 – Комп’ютерні науки. – Поліський національний університет, Житомир, 2024.

Зміст анотації

Кваліфікаційна робота присвячена розробці веб-орієнтованої технології розпізнавання текстового контенту інформаційних ресурсів українською мовою. Запропонована система надасть змогу автоматичного аналізу та коригування текстів, дозволяючи покращити стиль письма та підвищити зрозумілість викладеного матеріалу, а також виявить опосередковані ознаки деструктивного інформаційного впливу на користувачів.

Ключові слова: аналіз, водність, емоційне забарвлення, заспамленість, ключові слова, розпізнавання, текстовий контент, українська мова.

SUMMARY

Lyman J.O. WEB-based Technology for Text Content Recognition of Information Resources in Ukrainian.

Qualification work for the degree of Bachelor in Computer Science. – Polissky National University, Zhytomyr, 2024.

Content of the summary

The qualification work is devoted to the development of a web-based Technology for Text Content Recognition of Information Resources in Ukrainian. The proposed system will enable automatic analysis and correction of texts, improving the writing style and increasing the comprehensibility of the material, and also revealing indirect signs of destructive informational influence on users.

Keywords: analysis, water, emotional coloring, spam, key words, recognition, text content, Ukrainian language.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

API (Application Programming Interface) – інтерфейс прикладного програмування;

ML (Machine Learning) – машинне навчання;

NLU (Natural Language Understanding) – розуміння природної мови;

UML (Unified Modeling Language) – уніфікована мова моделювання.

ВСТУП

З популяризацією цифрових технологій доступ до інформації стає набагато простішим та швидшим. Особливо в сучасних умовах стрімко зростає кількість вебсайтів, форумів, новинних порталів, електронної документації, електронних книг та інших джерел на українській мові. Однак виникає питання щодо якості та достовірності інформації та її впливу на сучасне суспільство. Поширення фейкового та маніпулятивного контенту призводить до формування деструктивних ідей та поглядів, деформує навички розуміння, аналізу, оцінки та критичного сприймання інформації, знижує рівень загальної освіченості та обізнаності про ситуацію у світі. Також, нерідко маніпулятивний контент містить малозмістовні слова та велику кількість повторень для відволікання уваги читача від прихованої дезінформації. У зв’язку з цим, виникає потреба в ресурсах, які допоможуть виявляти деструктивний контент та сприятимуть мінімізації ризиків негативного впливу на інформаційну свідомість суспільства.

Метою кваліфікаційної роботи є створення веб-орієнтованої технології, яка допоможе аналізувати текстовий контент та виявляти опосередковані ознаки маніпулятивного впливу на споживачів контенту українською мовою.

Для досягнення поставленої мети було сформульовано наступні завдання:

- дослідження методів розпізнавання текстового контенту**;**
- огляд існуючих рішень;
- побудова UML-діаграм для опису функціоналу;
- практична реалізація веб-орієнтованої технології розпізнавання тексту інформаційних ресурсів українською мовою;
- тестування реалізованого програмного продукту.

Об’єктом дослідження є процес розпізнавання текстового контенту інформаційних ресурсів українською мовою.

Предметом дослідження є методи та інструменти розпізнавання текстового контенту українською мовою на наявність маніпулятивності.

Для теоретичного аналізу використовувались такі методи дослідження як аналіз, порівняння, абстрагування, для практичного дослідження – моделювання, програмування та математичний апарат нейронних мереж.

За темою кваліфікаційної роботи було опубліковано наукові публікації:

- Лиман Ю.О. Ідентифікація ознак маніпулятивного контенту Інтернет-ресурсів. Шляхи вирішення. Міжфакультетська науково-практична Інтернет-конференція «Безпека, технології, інновації: нові горизонти», Житомир : Поліський національний університет, 2023 р. С. 24-25.
- Лиман Ю.О. Критерії визначення наявності маніпулятивного контенту інформаційних ресурсів. Всеукраїнська науково-практична конференція здобувачів вищої освіти і молодих вчених «Інформаційні технології та моделювання систем». Житомир : Поліський національний університет, 2024 р. С. 47-48.

Практичне значення застосунку полягає у полегшенні виявлення непрямого маніпулятивного контенту та покращенні якості текстів. Застосунок орієнтований на потреби як звичайних користувачів, так і фахівців, що спеціалізуються в областях, які вимагають глибокого розуміння текстового контенту, наприклад, копірайтери, інформаційні аналітики, журналісти, соціологи, науковці тощо.

Структура кваліфікаційної роботи: вступ, 3 розділи, висновки, список використаних джерел (20 джерел), додатки (6 стор.), 2 таблиці та 11 рисунків. Обсяг роботи: 36 сторінок, 21 – основного тексту.

Розділ 1. ТЕОРЕТИЧНИЙ АНАЛІЗ ОСОБЛИВОСТЕЙ РОЗПІЗНАВАННЯ ТЕКСТОВОГО КОНТЕНТУ

1.1 Аналіз методів розпізнавання текстового контенту

На сьогодні існує велика кількість методів, які розпізнають текстовий контент: символьний, граматичний, емоційний тощо.

Для опису символьного розпізнавання підходить праця французького нейробіолога Станісласа Деана, в якій зазначається, що людський мозок використовує набір базових простих форм для декодування візуальних конструкцій (лінії, краї, кути) [3]. За його словами, серед різноманіття атрибутів, які сприймає людина, провідне значення має форма. Аргументація даного тезису полягає в наступному: людина може так само ефективно розпізнавати велику кількість об'єктів за їхніми простими контурними зображеннями, як і за детальними кольоровими фотографіями, які налічують додаткові атрибути, такі як величина, колір, текстура [4]. Наприклад, ми можемо легко розпізнати силуети на зображенні, навіть якщо це лише чорні контури на білому фоні. Після цього, опис об'єкта порівнюється із іншими описами, які зберігаються в зоровій пам'яті, і обирається найкраща йому відповідність. Наприклад, розпізнати конкретний об'єкт як літеру «А» – означає зазначити, що його форма більше відповідає формі літери «А», ніж формам інших літер [5].

Граматичний аналіз поділяють на морфологічний та лексичний. Морфологічний аналіз – це аналіз словникової (початкової) форми слова, визначення частини мови та граматичних характеристик, такі як відміна, дієвідміна, рід, число, відмінок тощо. Лексичний аналіз тексту виявляє його семантику і структуру. Даний аналіз включає знаходження ключових слів, стоп-слів, кількості повторів кожного слова, словесні кліше, що дозволяють маніпулювати сенсом тексту тощо [6]. Для підведення підсумків граматичного аналізу використовуються параметри водності та заспамленості. Водність

відповідає за перенасиченість тексту малозначущими словами, а заспамленість – ключовими. Такі параметри дозволяють визначати загальну інформативність та рівень читабельність контенту.

Емоційний аналіз, або емоційне забарвлення мови – складний аспект комунікації, який визначається шляхом сприйняття та аналізу різних елементів мовлення. Відмінність від усного спілкування полягає в тому, що письмове вираження емоцій не включає жести, інтонацію та міміку, але залежить від творчого використання текстових засобів. Наприклад, деякі слова безпосередньо передають емоції, тоді як інші можуть використовуватися для їхнього неявного вираження в залежності від контексту [7]. Не менш важливим є врахування пунктуації, оскільки ступінь емоційності може передаватися за допомогою знаків емоційності, таких як окличні та питальні. Також, люди зазвичай враховують стилістичні засоби, такі як повтори, вигуки, та візуальні елементи (смайлики).

На основі проаналізованих методів розпізнавання текстового контенту була побудована IDEF0-модель, наведена в Додатку А.

1.2 Аналіз існуючих рішень розпізнавання текстового контенту

З відомих на сьогодні реалізацій, які виконують схожі задачі, можна виділити Leegle, IBM Watson Natural Language Understanding (NLU), MeaningCloud, MonkeyLearn, Voyant Tools та Analyze My Writing (AMW).

Leegle [8] – це інструмент, призначений для перевірки текстового контенту та виявлення його впливу на підсвідомість людини шляхом виділення спеціальних лінгвістичних конструкцій, слів та оборотів, що сприяють некритичному сприйняттю інформації. Виявляються прийоми маніпулятивних навіювань, гіпнозу, зомбування, прихованого контролю думок, емоційного впливу, кібербулінгу тощо. Недоліки інструменту: необхідність попередньої обробки тексту та відсутність можливості зберігання результатів аналізу.

IBM Watson NLU [9] – використовує моделі машинного навчання для роботи з великими обсягами даних. Сервіс аналізує семантичні особливості

введення тексту, включаючи категорії, поняття, емоції, сутності, ключові слова, метадані, відносини, семантичні ролі та почуття. Недоліком сервісу є обмежена точність аналізу неангломовного контенту та не завжди коректне визначення настрою тексту.

MeaningCloud [10] – це вебсервіс, який фокусується на автоматичному вилученні інформації з різних неструктурованих джерел, таких як статті, документи, вебсайти тощо. Сервіс надає ряд функцій текстової аналітики, включаючи аналіз структури тексту (заголовки, теми, автори), визначення семантики (позитивний, нейтральний чи негативний настрій, суб'єктивність, об'єктивність, іронія), витяг іменованих сутностей (людей, організацій), аналіз контексту (відображення взаємозв'язків між темами та підтемами тексту) і сумаризація тексту (виділення найважливіших речень). Недоліки MeaningCloud: не всі мови мають високу точність аналізу, зниження точності через використання сленгових виразів, залежність від якості вхідних даних, відсутність візуалізації результатів.

MonkeyLearn [11] – вебсервіс для автоматизованого аналізу текстів і обробки природної мови з використанням ML. Пропонує виявлення емоційного тону (sentiment analysis), визначення тематики, виділення тегів, сутностей, ключових слів тощо. Також, дозволяє створювати моделі під власні потреби та миттєво візуалізувати результати роботи з можливістю комбінування та фільтрування графіків за кількома вхідними даними. Основними недоліками сервісу є висока вартість для малих підприємств і фрілансерів, та складність навігації.

Voyant Tools [12] – вебсайт, який спеціалізується на текстовій аналітиці: виділення ключових слів, відстеження частоти вживання термінів та виразів, візуалізація структури тексту у вигляді графіків та хмари слів. Перевага даного сервісу – експорт результатів аналізу. Використовуючи Voyant Tools, користувачі можуть легко переносити згенеровані візуалізації та дані, що допомагає збагачувати вебсторінки, блоги, форуми та будь-які інші онлайн-проєкти. Недоліком сервісу є насичений інтерфейс, що складається з панелей, кожна з яких

відповідає за власне завдання. Розміщення великої кількості інформації на екрані може викликати труднощі у користувачів. Тим не менше, біля кожної панелі є іконка зі знаком питання, при натисканні на яку відкривається сторінка з документацією для відповідного інструменту. Також, у роботі вебсайту часто виникають збої, тому рекомендується завантажити офлайн-версію.

AMW [13] призначений для детального аналізу текстового вмісту, який включає визначення кількості символів, слів, речень, здійснення перевірки наявності граматичних помилок та дає оцінку читабельності за п’ятьма індексами (Gunning fog, Flesch-Kincaid, SMOG, Coleman-Liau, Automated readability). Головною перевагою AMW є швидкість та точність. Користувачу потрібно лише вставити текст у відповідне поле та натиснути на кнопку «Analyze Text!». Незважаючи на це, AMW обмежений розпізнаванням лише англійської мови. Вебсайт не оновлювався з 2018 року, тому ймовірність впровадження нових функцій дуже низька.

У таблиці 1.1 наведений додатковий порівняльний аналіз обраних застосунків.

Таблиця 1.1 – Порівняльна характеристика застосунків

Параметр	Leegle	IBM Watson NLU	Meaning Cloud	Monkey Learn	Voyant Tools	AMW
Форми аналізу	текст	текст, посилання	текст, посилання	текст, посилання, зображення	текст, посилання	текст
Розпізнавання тексту українською мовою	+	-	-	-	+	-
Лексичний аналіз	-	-	-	-	+	+
Пошук ключових слів	+	+	+	+	+	+
Визначення водності, заспамленості	-	-	-	-	-	-
Статистика слів	-	+	+	+	+	+
Аналіз емоційності	+	+	+	+	-	-
Візуалізація результатів	+	-	-	+	+	+
Завантаження звітності	-	+	+	+	+	-

Проведене дослідження показує, що більшість наявних рішень не підтримують аналіз української мови. Основна частина інструментів спрямована на моніторинг відношення клієнтів до брендів, товарів та послуг підприємств,

орієнтуючись на їхні коментарі. Інші системи проводять аналіз тексту, але не виявляють прихованих маніпуляцій. Тож необхідність дослідження обраної предметної області визначається тим, що поки не існує доступних інтелектуальних систем, які б одночасно аналізували україномовний текст на коректність та адекватність, і виявляли опосередковані ознаки маніпулятивного впливу на споживачів.

Висновки до першого розділу

В результаті аналізу предметної області визначено та охарактеризовано методи розпізнавання текстового контенту. Побудовано функціональну модель розпізнавання текстового контенту на основі методології моделювання IDEF0.

Досліджено існуючі застосунки, які використовуються для рішення схожих задач у рамках розпізнавання текстового контенту: Leegle, IBM Watson NLU, MeaningCloud, MonkeyLearn, Voyant Tools та AMW. Визначено необхідність створення власного застосунку, який аналізує тексти українською мовою з метою покращення їхньої читабельності, а також виявляє наявність ознак маніпулятивного впливу на суспільство.

РОЗДІЛ 2. ПРОЄКТУВАННЯ ВЕБ-ОРІЄНТОВАНОЇ ТЕХНОЛОГІЇ РОЗПІЗНАВАННЯ ТЕКСТОВОГО КОНТЕНТУ ІНФОРМАЦІЙНИХ РЕСУРСІВ УКРАЇНСЬКОЮ МОВОЮ

2.1 Моделювання веб-орієнтованої технології розпізнавання текстового контенту інформаційних ресурсів українською мовою

Для опису функціоналу веб-орієнтованої технології розпізнавання текстового контенту інформаційних ресурсів українською мовою були створені діаграми в нотації UML, які забезпечують детальний огляд організації її роботи. UML – універсальна мова з відкритим стандартом, яка використовує графічні символи для формування абстрактної моделі системи [14].

Діаграма прецедентів (use case diagram) проєктує систему з точки зору кінцевого користувача, тобто демонструє її функціонування в навколишньому середовищі (рис. 2.1).

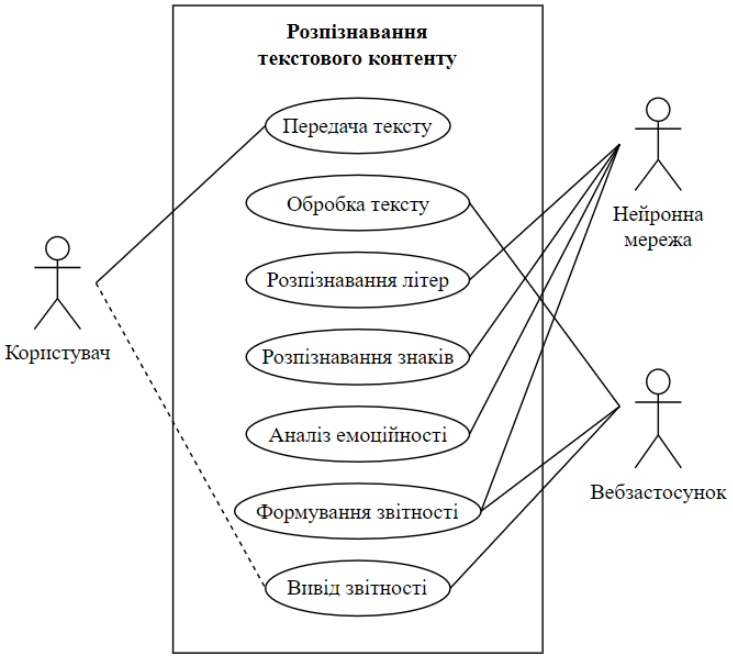

Рисунок 2.1 – Діаграма прецедентів

На діаграмі показані взаємодії між акторами та прецедентами. Актор – це роль, яка приймає активну участь у роботі системи. В веб-орієнтованій технології

розпізнавання текстового контенту акторами виступають користувачі, нейронна мережа та вебзастосунок. Прецеденти описують поведінку системи залежно від дій акторів. Кожен прецедент представляє певний функціональний елемент, який викликається конкретним актором.

Діаграма послідовності (sequence diagram) відображає послідовність взаємодій між об'єктами з часовою шкалою, яка починається вгорі та поступово спускається вниз (рис. 2.2). Кожен об'єкт представляється стовпцем, а взаємодії між ними відображаються стрілками, що вказують напрямок обміну повідомленнями [15]. Діаграма дозволяє детально та коротко відобразити динамічну поведінку системи та виділити її можливості.

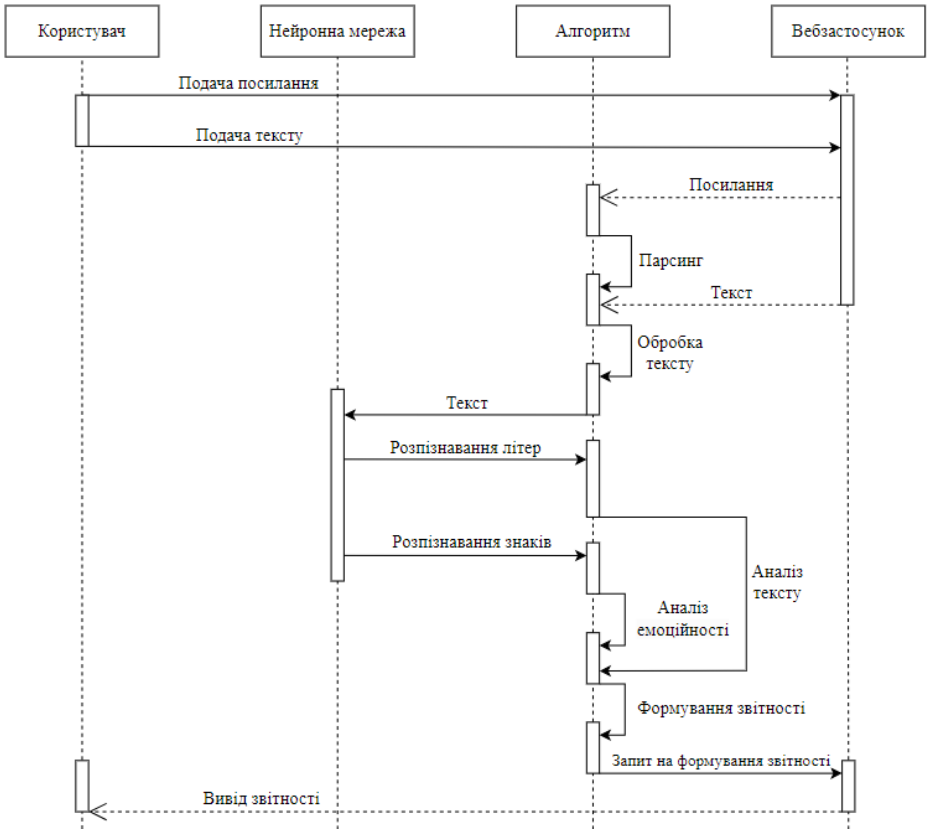

Рисунок 2.2 – Діаграма послідовності

Після запуску вебзастосунку, користувач може ввести текст для подальшого аналізу двома способами: передати посилання на вебсайт, з якого буде взятий текст, або вставити вже скопійований. Після цього проходить попередня обробка, наприклад, перетворення всіх літер у нижній регістр, заміна деяких символів (тире замінюється дефісом, нерозривний пробіл – звичайним тощо). Подальший

процес аналізу виконується двома нейронними мережами: одна відповідає за літери, а інша – за знаки. Після завершення аналізу емоційності формується звіт.

2.2. Розроблення математичної моделі

Як було зазначено в п.п. 2.1, в основі інтелектуальної системи лежать дві нейронні мережі, одна з яких розпізнає українську символіку, а інша – знаки. Архітектура мереж схожа (Додаток Б.1), відрізняється лише кількість входів та виходів. Нейронна мережа розпізнавання українських літер має 16 входів, 33 виходи, нейронна мережа розпізнавання знаків – 8 входів, 10 виходів.

Опис алгоритму.

Нехай:

N – кількість прецедентів;

L – кількість нейронів в мережі;

k_r – кількість нейронів в шарі r, де r = 1,2, ……L;

k_L – кількість вихідних нейронів;

$k_0 = l$ – розмір входу;

$x(i) = (x_1(i),x_2(i),……x_{k0}(i))$ – вхідний вектор ознак;

$y(i) = (y_1(i),y_2(i),…….y_{kL}(i))$ – вихідний вектор, який повинен бути вірно класифікований.

1. **Початкове наближення.** Випадково обираються ваги невеликих значень W_{jk}^r, де $r= 1,2,.....L$, $j=1,2,……k_r$, $k=0,1,2,……..k_{r-1}$.

2. **Прямий прохід.** Для кожного вектора прецедента $x(i)$, обчислюються всі аргументи функції активації j-го нейрона r-го шару $V_j^r(i)$:

$$y_j^r(i) = f(V_j^r(i)),$$

де $j=1,2,……k_r$, $r=1,2,…….L$.

Обчислюється сума квадратів помилок між фактичними і бажаними виходами мережі ($J(W)$):

Цикл по $i=1,2,…….N$ (по прецедентам):

Обчислити:

$$y_k^0(i) = x_k(i), \quad k = 1,2,...,k_0$$

$$y_0^0(i) = 1$$

Цикл по $r=1,2,\ldots\ldots L$ (по шарам):

Цикл по $j=1,2,\ldots\ldots k_r$ (по нейронам в шарі):

$$V_j^r(i) = \sum_{k=0}^{k_{r-1}} W_{jk}^r y_k^{r-1}(i)$$

$$y_j^r(i) = f\left(V_j^r(i)\right)$$

Кінець циклу по j.

Кінець циклу по r.

Кінець циклу по i.

$$J(W) = \sum_{i=1}^{N} \frac{1}{2}\left(y_j^L(i) - y_j(i)\right)^2$$

3. **Обернений прохід.** Для кожного значення $i=1,2\ldots\ldots N$ та $j=1,2,\ldots\ldots k_L$ обчислюється $\frac{d\varepsilon(i)}{dV_j^L(i)}$. Потім, послідовно обчислюється $\frac{d\varepsilon(i)}{dV_j^r(i)}$ для всіх $r=(L-1),\ldots\ldots 1$ та $j=1,2,\ldots\ldots k$:

Цикл по $i=1,2,\ \ldots\ldots k_r$ (по нейронам в шарі):

Обчислити:

$$e_j(i) = y_j^L(i) - y_j(i)$$

$$\delta_j^L(i) = e_j(i) \cdot f'\left(V_j^{r-1}(i)\right)$$

Цикл по $r=L,L-1,\ldots\ldots 2$ (по шарам):

Цикл по $j=1,2,\ldots\ldots k_r$ (по нейронам в шарі):

$$e_j^{r-1}(i) = \sum_{k=1}^{k_r} \delta_k^r(i) \cdot W_{kj}^r$$

$$\delta_j^{r-1}(i) = e_j^{r-1}(i) \cdot f'\left(V_j^{r-1}(i)\right)$$

Кінець циклу по j.

Кінець циклу по r.

Кінець циклу по i.

4. **Перерахунок ваг.** Для всіх $r=1,2,\ldots\ldots L$ і $j=1,2\ldots\ldots k_r$,

$W_j^r(new) = W_j^r(old) + \Delta W_j^r$, де $\Delta W_j^r = -\mu \sum_{i=1}^{N} \frac{\partial \varepsilon(i)}{\partial V_j^r} y^{r-1}(i)$.

На рисунку Б.2 (Додаток Б) показано динаміку зменшення похибки вихідного шару.

Математичні розрахунки параметрів системи продемонстровані в таблиці 2.1.

Таблиця 2.1 – Характеристика параметрів системи та їхнє визначення

Параметр	Характеристика	Визначення
Водність	Кількість малозмістовних слів, які не додають цінної інформації, перевантажують текст. Прийнятний стандарт: 55%.	$P_1 = \frac{SW}{W} * 100$, де SW – кількість малозмістовних та стоп-слів, W – кількість всіх слів.
Частотність	Кількість повторень окремих слів, враховуючи зміну відмінка, роду та числа.	$P_2 = \frac{W_i}{W} * 100$, де W_i – кількість конкретного слова, W – кількість всіх слів.
Заспамленість	Зловживання повторами ключових слів. Прийнятний стандарт для одного слова: 3%.	$P_3 = P_{2KW_0} + \sum_{i=KW_1}^{9} (P_{2KW_i} - N))$, де $P_{2KWi} > N$, P_{2KW} – частота вживаності кожного ключового слова, N – прийнятний стандарт одного слова.
Рівень емоційного забарвлення	Насиченість тексту емоціями, визначення чи є виражена думка позитивною, негативною або нейтральною.	$P_4 = \frac{(E+Q)}{CS} * 100$, де E – кількість окличних речень, Q – кількість питальних речень, CS – кількість всіх речень.

Висновок до другого розділу

Розроблено модель веб-орієнтованої технології розпізнавання текстового контенту інформаційних ресурсів українською мовою з використанням об’єктно-орієнтованого підходу та мови UML. Створено UML-діаграми прецедентів (use case diagram) та послідовності (sequence diagram) для отримання чіткого представлення про процес роботи застосунку. Розроблено математичну модель структури нейронних мереж та наведено формули розрахунку основних параметрів застосунку.

Розділ 3. РЕАЛІЗАЦІЯ ВЕБ-ОРІЄНТОВАНОЇ ТЕХНОЛОГІЇ РОЗПІЗНАВАННЯ ТЕКСТОВОГО КОНТЕНТУ ІНФОРМАЦІЙНИХ РЕСУРСІВ УКРАЇНСЬКОЮ МОВОЮ

3.1 Розробка інтерфейсу веб-орієнтованої технології

Для реалізації веб-орієнтованої технології було обрано високорівневий вебфреймворк Python з відкритим кодом, який стимулює швидкий розвиток та чистий, прагматичний дизайн, що відповідає архітектурному зразку «Model-Template-Views» (MTV) – Django [16]. На його основі будуються як невеликі односторінкові вебсайти, так і великі вебпортали, тому що розробники забезпечені великим набором можливостей. Фреймворк дозволяє комфортно працювати з базами даних, URL-маршрутизацією, аутентифікацією користувачів, має вбудований адміністративний інтерфейс для управління даними, забезпечує можливість легко розширювати функціональність за допомогою сторонніх бібліотек та плагінів. Django має вбудовані засоби для захисту від більшості типів CSRF-атак (міжсайтова підробка запиту) і XSS-атак (міжсайтовий скриптинг) [17].

Для створення структури вебсторінок системи та стилізації їхніх елементів використовувались мови-розмітки HTML, CSS та JavaScript.

Основна панель вебсайту складається з чотирьох розділів: «Головна», «Ресурси», «Аналіз» і «Контакти». Сторінки мають різну структуру.

На титульній сторінці розміщуються загальні відомості про інформаційну систему (рис. 3.1).

Розділ «Ресурси» знаходиться на тій же сторінці, що й головна, але для зручності створено посилання за допомогою <a href="\#features">. У гіперпосиланні символ «#» вказує на ідентифікатор HTML-елемента, до якого буде перенесено вікно. Таким чином, користувач перейде до вказаного розділу на

головній сторінці. У розділі надається коротка характеристика можливостей вебсайту (рис. 3.2).

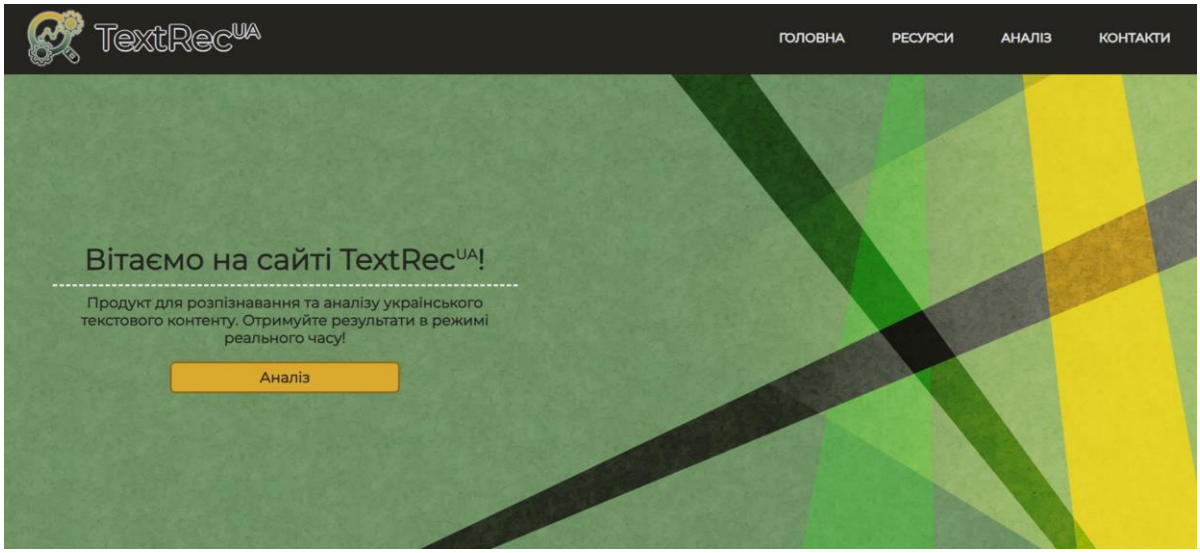

Рис. 3.1 – Інтерфейс розділу «Головна»

Рис. 3.2 – Інтерфейс розділу «Ресурси»

Розділ «Контакти» реалізований у вигляді спливаючого вікна (Додаток В). Натиснувши на посилання, з'явиться спливаюче вікно з контактною інформацією, натиснувши ще раз – вікно зникне.

При переході у розділ «Аналіз», перед користувачем з'являться два поля, одне з яких призначене для введення посилання на вебсайт, контент якого він бажає проаналізувати, а друге – для введення будь-якого тексту (рис. 3.3). Нижче

знаходяться чотири блоки: «Парсинг», «Розпізнаний текст», «Аналіз тексту» та «Статистика слів». В кожному з них будуть розміщені власні проаналізовані дані.

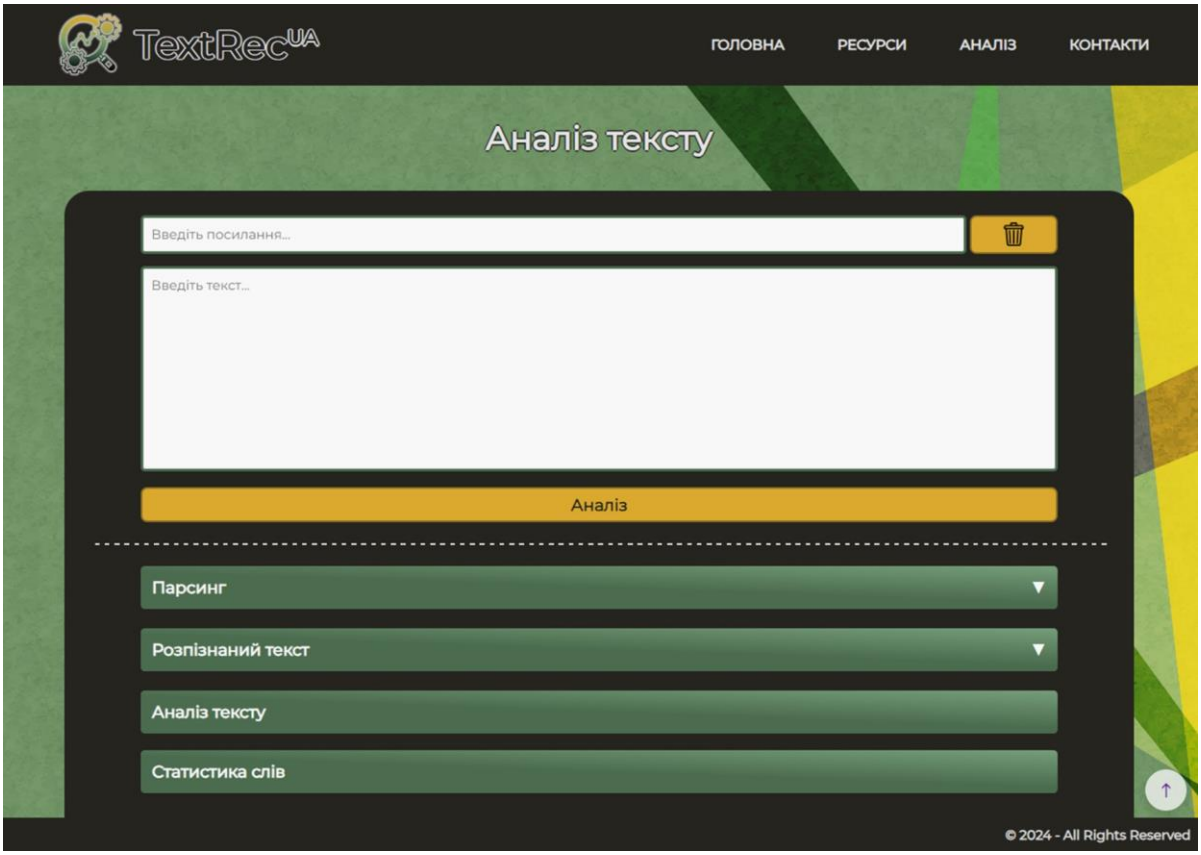

Рис. 3.3 – Інтерфейс розділу «Аналіз тексту»

В правому нижньому кутку екрана знаходиться стрілка, натиснувши на яку, сторінка автоматично прокрутиться вгору.

3.2 Реалізація функцій веб-орієнтованої технології розпізнавання текстового контенту інформаційних ресурсів українською мовою

Основні функції застосунку знаходяться у розділі «Аналіз». Їхні назви співпадають з назвами блоків на рисунку 3.3.

1. Парсинг.

Спочатку створюється заголовок User-Agent для імітації реального користувача та відправляється GET-запит на вказану URL-адресу для отримання HTML-контенту веб-сторінки за допомогою бібліотеки requests. Наступний крок

виконує бібліотека BeautifulSoup, яка дозволяє витягувати інформацію з отриманого HTML-коду, надаючи різні методи для навігації і пошуку в його структурі. Для подальшої обробки виділяється тег <body>. З нього видаляються батьківські і дочірні елементи з класами та ідентифікаторами, в яких зазвичай міститься інформація, що не відноситься до основного контенту вебсторінки, наприклад navigation, menu, cookie, comments, banner тощо. Після додаткового очищення, обирається тег з найбільшою довжиною тексту.

2. Розпізнавання тексту відбувається за допомогою двох нейронних мереж, які описуються у п.п. 2.2. Метод вибору відповідної нейронної мережі наведено в Додатку Г.

3. Аналіз тексту.

Процес аналізу складається з кількох етапів:

- підрахунок кількості символів (з пробілами, без пробілів);
- підрахунок кількості слів (формується список усіх слів тексту, а кількість визначається функцією len());
- знаходження стоп-слів:

цей етап включає виділення слів по довжині (слова, довжиною менші за 2 включно), та за частиною мови. Для останнього використовується морфологічна бібліотека pymorphy3 та словник української мови pymorphy3-dict-ua. Дана бібліотека використовує грамеми для позначення частин мови: ["NPRO", "CONJ", "PREP", "PRCL", "INTJ", "ADVB", None (не визначено)]. Розшифрування значень грамем наводиться у Додатку Д;

- обчислення параметрів водності, заспамленості та рівня емоційної забарвленості (див. табл. 2.1);
- визначення полярності тексту:

на даному етапі визначається вплив тексту (позитивний, нейтральний або негативний), використовуючи клас SentimentIntensityAnalyzer модулю vader бібліотеки nltk. Оскільки словник аналізатора англійський, було об'єднано два українських словники [18-19] та проведено ручне доповнення. Словник має

структуру «слово-значення», де позитивному слову присвоюється «-1», а позитивному – «1». Лістинг коду наведено в Додатку Ж;

- визначення ключових слів:

в ролі ключових слів виступають перші десять слів словника dict_words_count, де ключ – слово, значення – кількість вживаності в тексті.

Для кращого сприйняття інформації створений додатковий блок з іконками, використовуючи Iconfinder API [20]. Також, іконки будуть корисними для користувачів, які створюють контент, оскільки вони можуть використовувати їх для візуалізації своїх матеріалів.

4. Статистика слів.

Формується двовимірний масив з інформацією про слова: номер за порядком, слово, похідні слова, які зустрічаються у тексті (включаючи саме слово), загальна кількість повторів та частота появи, виражена у відсотках. Масив відсортований за значенням кількості повторень у порядку спадання.

3.3 Керівництво користувачу

В розділ для аналізу тексту можна потрапити двома способами: натиснувши на розділ «Аналіз» на верхньому меню або натиснувши на одноіменну кнопку головної сторінки.

В цьому розділі розташовані два поля для введення тексту та посилання. Передбачені сповіщення про помилки при введенні даних, якщо користувач:

- заповнить всі поля (рис. 3.4):

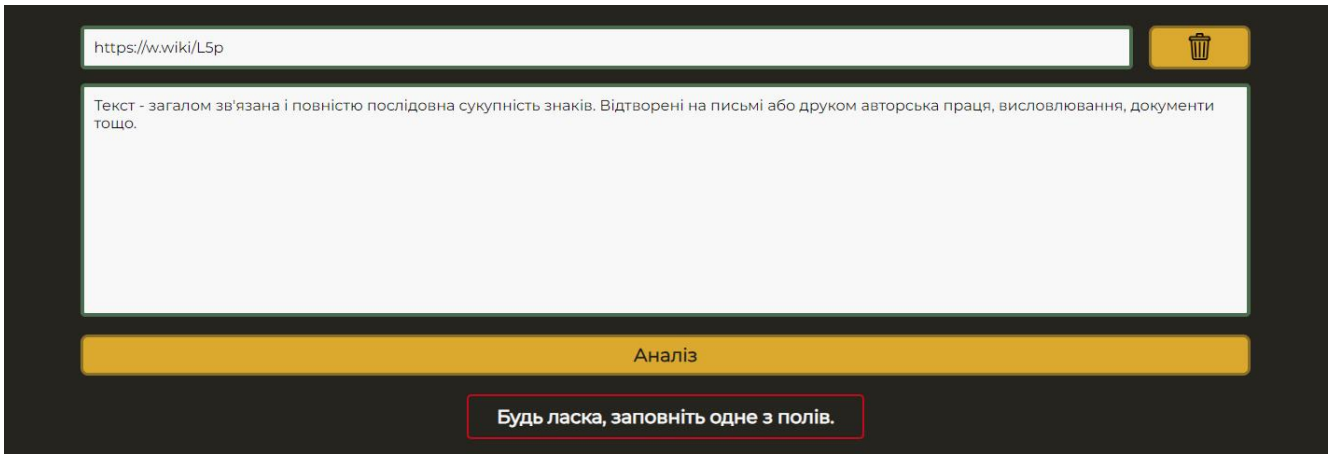

Рис. 3.4 – Сповіщення про помилку

- залишить поля пустими (рис. 3.5):

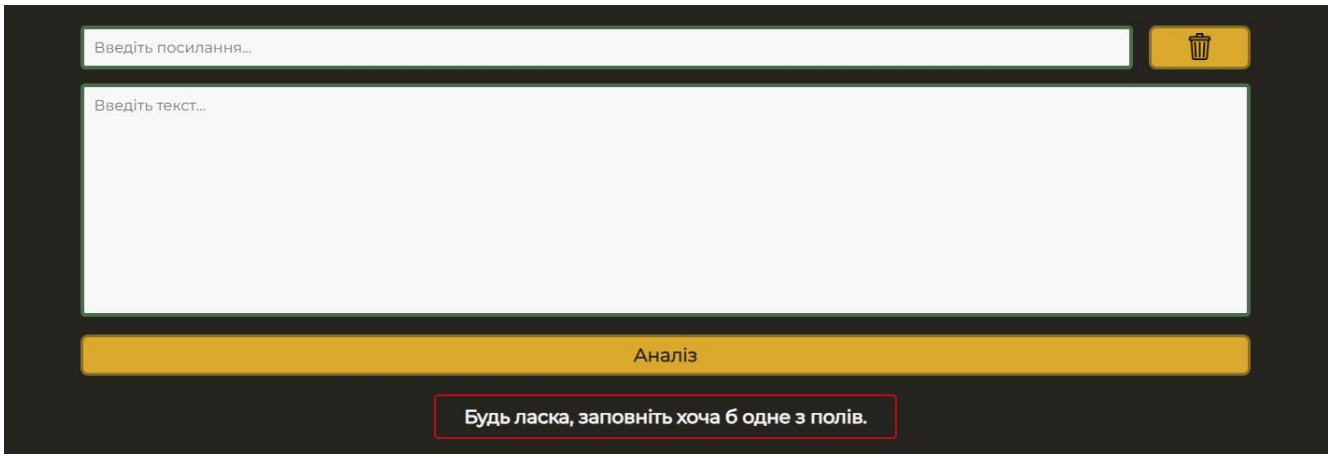

Рис. 3.5 – Сповіщення про помилку

- некоректно введе дані (рис. 3.6):

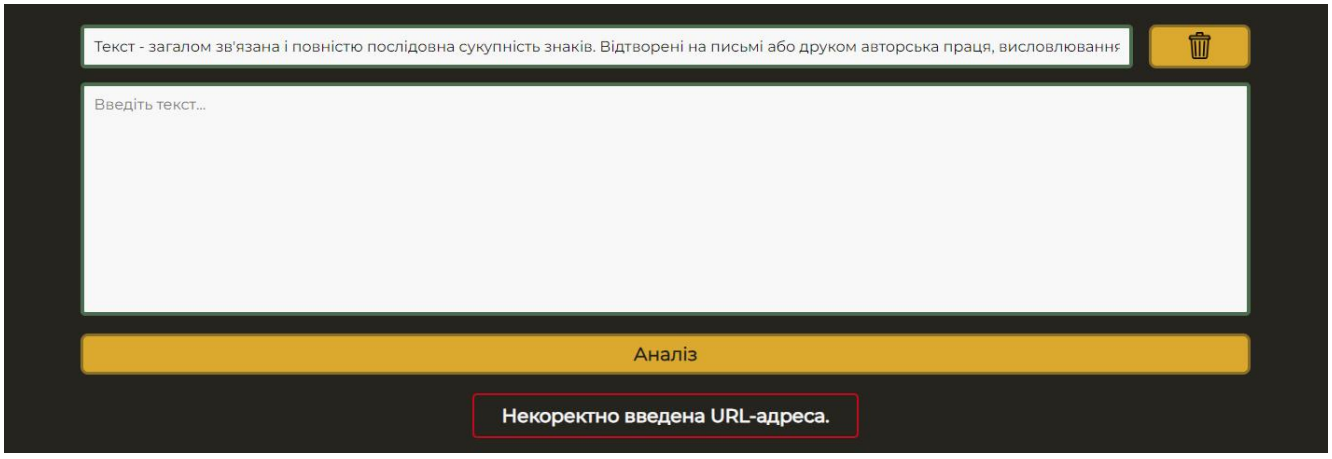

Рис. 3.6 – Сповіщення про помилку

Спочатку блоки «Парсинг» та «Розпізнаний текст» заблоковані і будуть розблоковані після натискання кнопки «Аналіз». Також, блок «Парсинг» залишиться заблокованим, якщо користувач ввів власний текст, а не посилання.

На рисунку 3.7 показано вигляд розгорнутих блоків «Парсинг» та «Розпізнаний текст».

Рисунок 3.7 – Вигляд блоків «Парсинг» та «Розпізнаний текст»

Біля текстового поля першого блоку розташовані дві кнопки. Перша копіює текст, друга – завантажує. Користувачі можуть обирати між форматами TXT та DOCX. У наступному блоці показується розпізнаний текст, але взаємодіяти з ним не можна.

Нижче розташований аналіз тексту, а саме кількість символів, слів та стоп-слів, параметри водності та заспамленості, рівень емоційного забарвлення тексту та його полярність. Далі представлено підблок з ключовими словами та іконками для кращої візуалізації та асоціативного сприйняття тексту (рис. 3.8).

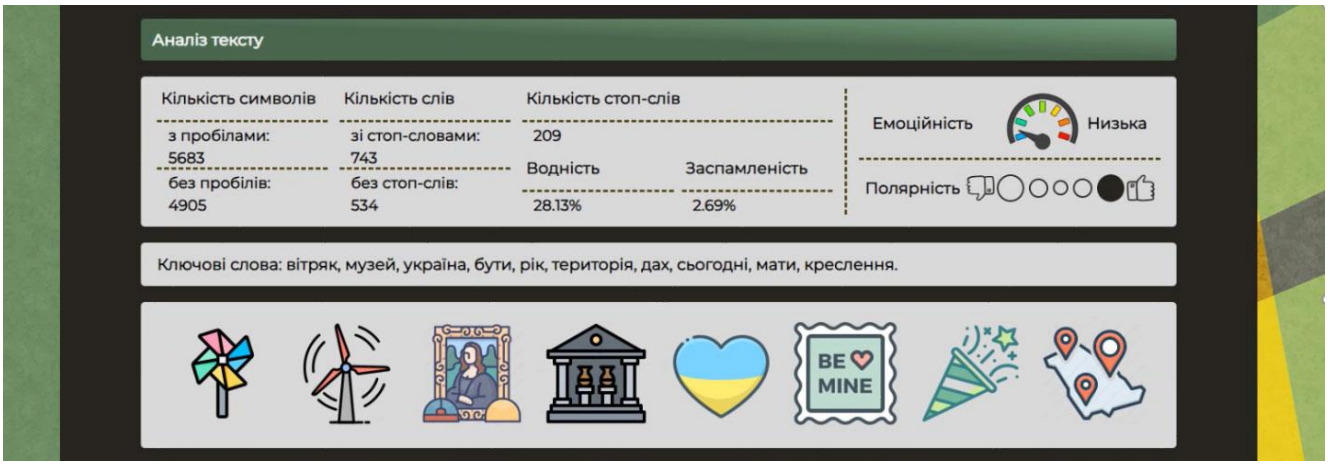

Рисунок 3.8 – Вигляд блоку «Аналіз тексту»

«Статистика слів» включає п’ять стовпців, перший з яких відображає нумерацію слів, другий – показує початкову форму слова, третій – всі форми слова, наявні в тексті, четвертий – кількість повторень, п’ятий – частота появи в тексті (рис. 3.9).

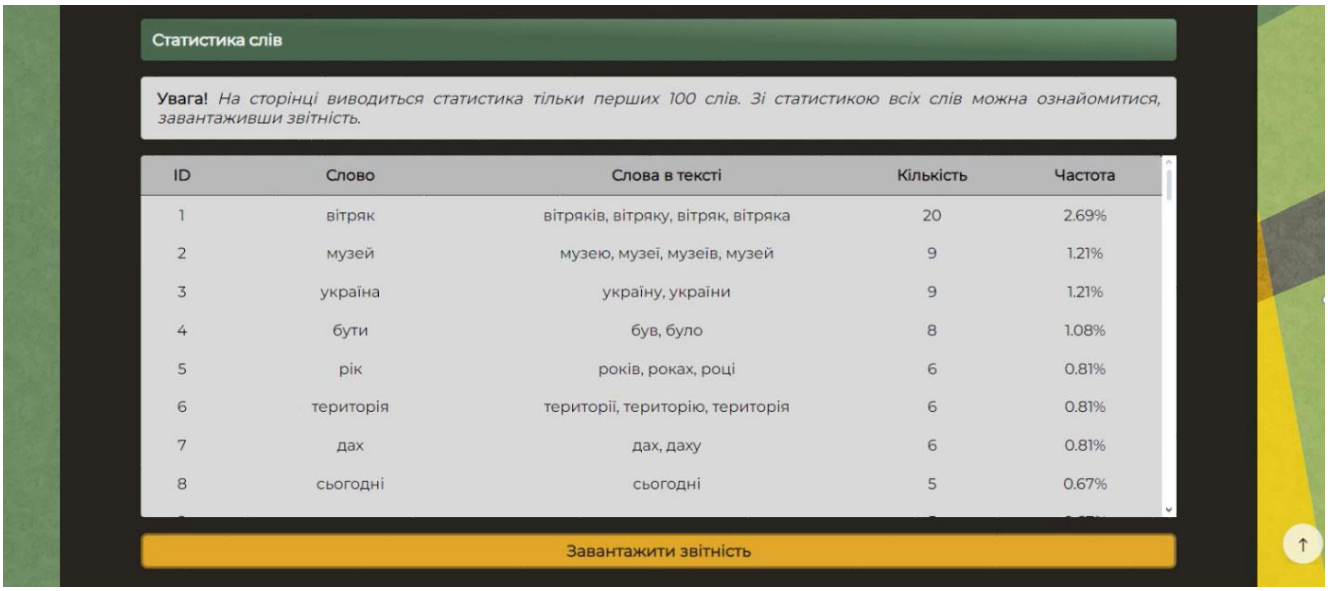

Рисунок 3.9 – Вигляд блоку «Статистика слів»

На вебсторінці виводяться перші 100 слів. З повною статистикою можна ознайомитись, завантаживши звітність.

Висновок до третього розділу

В даному розділі розроблено вебзастосунок розпізнавання текстового контенту. Описано інтерфейс кожного розділу, доповнено відповідними рисунками. Наведено реалізацію основних параметрів застосунку. Надано керівництво користувачу, що включає інструкції, перелік виняткових ситуацій та результати роботи.

ВИСНОВКИ

Метою кваліфікаційної роботи є створення системи, яка допоможе аналізувати текстовий контент та виявляти приховані маніпулятивні ознаки.

Для досягнення поставленої мети було виконано такі завдання:

- проведено аналіз методів розпізнавання текстового контенту та розглянуто існуючі рішення, визначено їхні переваги та недоліки, які були враховані в подальшій розробці;
- здійснено моделювання системи, побудовано UML-діаграми прецедентів та послідовності для демонстрації взаємодії її основних компонентів, розроблено модель нейронної мережі для розпізнавання текстового контенту;
- створено інтуїтивно-зрозумілий інтерфейс та функціональні можливості для аналізу тексту, надано детальне керівництво користувачу та показано результати роботи системи, підтверджуючи її працездатність.

В результаті кваліфікаційної роботи створено веб-орієнтовану технологію розпізнавання текстового контенту інформаційних ресурсів українською мовою, яка здатна покращувати якість текстів користувачів, а саме підвищувати рівень читабельності й інформативності, та визначати який вплив текст має на свідомість суспільства.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Лиман Ю.О. Ідентифікація ознак маніпулятивного контенту Інтернет-ресурсів. Шляхи вирішення. Міжфакультетська науково-практична Інтернет-конференція «Безпека, технології, інновації: нові горизонти», Житомир : Поліський національний університет, 2023 р. С. 24-25.

2. Лиман Ю.О. Критерії визначення наявності маніпулятивного контенту інформаційних ресурсів. Всеукраїнська науково-практична конференція здобувачів вищої освіти і молодих вчених «Інформаційні технології та моделювання систем». Житомир : Поліський національний університет, 2024 р. С. 47-48.

3. Dehaene, S. (2005) Evolution of Human Cortical Circuits for Reading and Arithmetic: The «Neuronal Recycling» Hypothesis. In: Dehaene, S., Duhamel, J.R., Hauser, M. and Rizzolatti, G., Eds., From Monkey Brain to Human Brain, MIT Press, Cambridge, 133-157.

4. Biederman, I., & Ju, G. (1988). Surface versus edge-base determinants of visual recognition. Cognitive Psychology, 20(1), 38-64.

5. Atkinson, R. L., Atkinson, R. C., Smith, E. E., Bem, D. J., & Nolen-Hoeksema, S. (2000). Hilgard's Introduction to Psychology: History, Theory, Research, and Applications (13th ed.).

6. Лінгвістичні основи компетентнісного підходу до навчання лексикології в основній школі [Електронний ресурс]. – 2015. – Режим доступу до ресурсу: https://www.ukrlogos.in.ua/10.11232-2663-4139.04.02.html.

7. Clore, G. L., Ortony, A., & Foss, M. A. (1987). The psychological foundations of the affective lexicon. Journal of Personality and Social Psychology, 53(4), 751-766.

8. Leegle: First Digital Shield Against Brainwashing [Електронний ресурс]. – 2016. – Режим доступу до ресурсу: https://www.indiegogo.com/projects/leegle-first-digital-shield-against-brainwashing#/.

9. IBM Watson Natural Language Understanding [Електронний ресурс]. – 1986. – Режим доступу до ресурсу: https://www.ibm.com/products/natural-language-understanding.

10. Text Analytics – MeaningCloud text mining solutions [Електронний ресурс]. – 2014. – Режим доступу до ресурсу: https://www.meaningcloud.com.

11. MonkeyLearn – Text Analytics [Електронний ресурс]. – 2013. – Режим доступу до ресурсу: https://monkeylearn.com.

12. Voyant Tools [Електронний ресурс]. – 2016. – Режим доступу до ресурсу: https://voyant-tools.org/docs/#!/guide/about.

13. Analyze My Writing [Електронний ресурс]. – 2015. – Режим доступу до ресурсу: https://www.analyzemywriting.com/index.html.

14. Unified Modeling Language [Електронний ресурс]. – 2023. – Режим доступу до ресурсу: https://uk.wikipedia.org/wiki/Unified_Modeling_Language.

15. Rumbaugh, J., Jacobson, I., Booch, G. (2004). The Unified Modeling Language Reference Manual, (2nd Edition). Addison-Wesley, 102-106.

16. Django [Електронний ресурс]. – 2005. – Режим доступу до ресурсу: https://www.djangoproject.com.

17. Documentation: Security in Django [Електронний ресурс]. – 2005. – Режим доступу до ресурсу: https://docs.djangoproject.com/en/5.0/topics/security/.

18. Український тональний словник [Електронний ресурс]. – 2016. – Режим доступу до ресурсу: https://github.com/lang-uk/tone-dict-uk.

19. Ukrainian-Sentiment-Analysis [Електронний ресурс]. – 2020. – Режим доступу до ресурсу: https://github.com/skupriienko/Ukrainian-Sentiment-Analysis.

20. Icons API [Електронний ресурс]. – 2004. – Режим доступу до ресурсу: https://www.iconfinder.com/api.

ДОДАТКИ

ДОДАТОК А

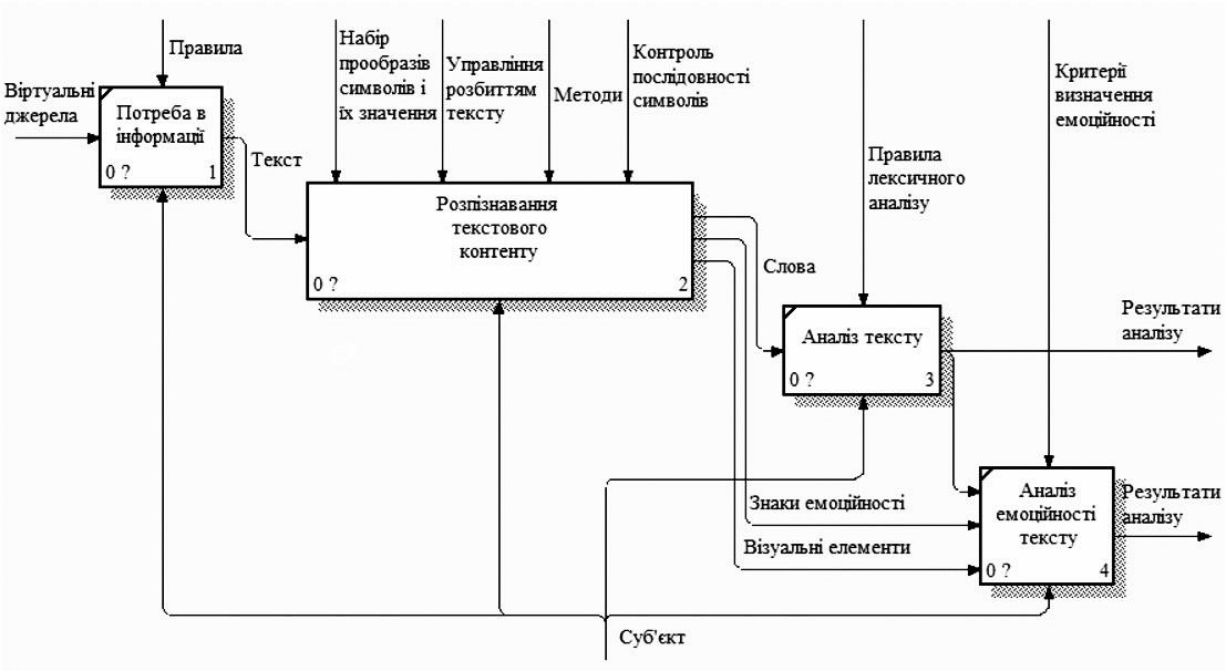

Рис. А.1 – IDEF0-модель «Розпізнавання текстового контенту»

ДОДАТОК Б

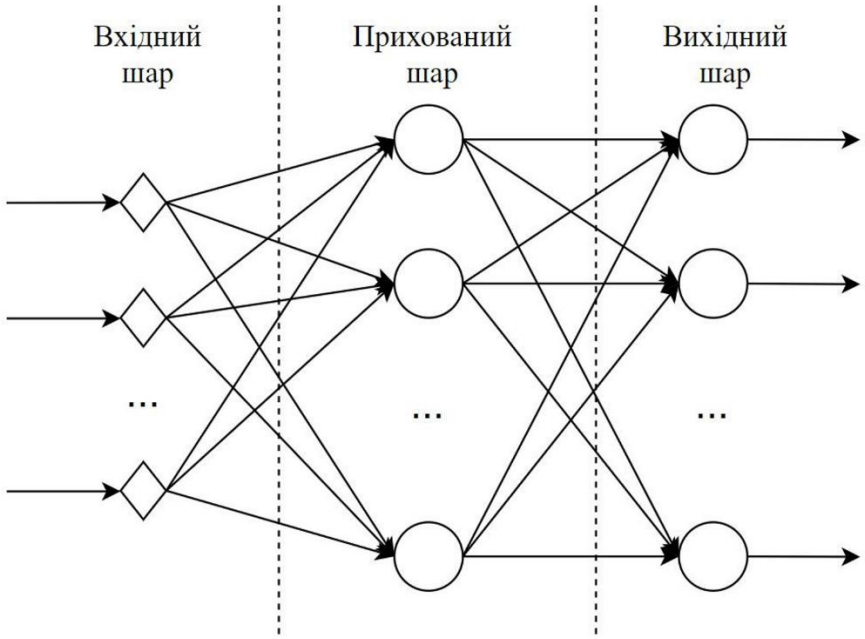

Рис. Б.1 – Архітектура нейронної мережі

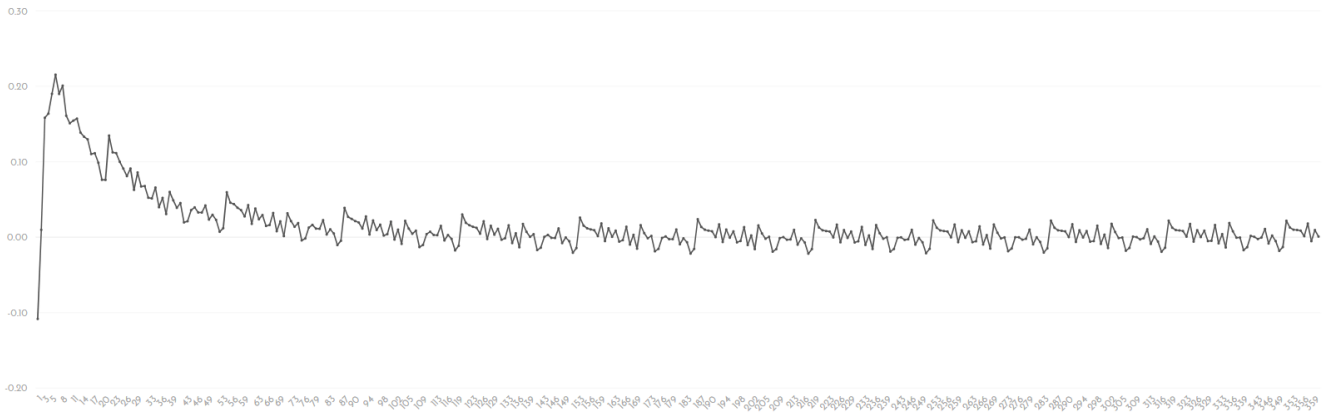

Рис. Б.2 – Графік динаміки зменшення похибки вихідного шару

ДОДАТОК В

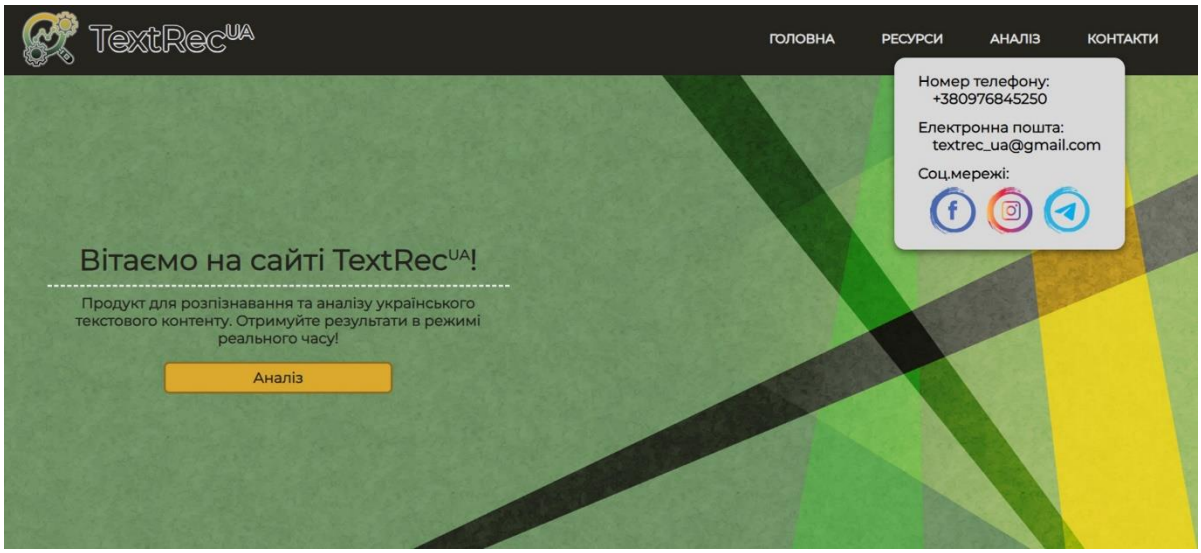

Рис. В.1 – Інтерфейс розділу «Контакти»

ДОДАТОК Г

Г.1 Лістинг програмного коду вибору нейронної мережі

```
1.  def identification_exitMatrix(self, array: list, arg_Ldict: dict, arg_Sdict: dict,
    final_arr: list):
2.          for id_, el in enumerate(array):
3.              if len(el) == 16:
4.                  x = self.nN_Letters.identification_matrix(array, id_, arg_Ldict)
5.              else:
6.                  x = self.nN_Symbols.identification_matrix(array, id_, arg_Sdict)
7.              final_arr.append(x)
```

ДОДАТОК Д

Таблиця Д.1 – Частини мови, що розпізнає бібліотека pymorphy3

Грамема	Значення	Приклад
NPRO	займенник-іменник	він, вона, воно
CONJ	сполучник	та, і, але
PREP	прийменник	в, для, понад
PRCL	частка	лише, навіть, саме
INTJ	вигуки	ой, ох, агов
ADVB	прислівник	надто, занадто, чисельно

ДОДАТОК Ж

Ж.1 Лістинг програмного коду визначення полярності тексту

```
def polarity_analyze(self):
        with open("polarity-dict-uk.json", "r", encoding="utf-8") as file:
            sentiment_dict = json.load(file)
        SIA = SentimentIntensityAnalyzer()
        new_lexicon = {}
        for element in sentiment_dict:
            new_lexicon[element['word']] = element['pos_neg']
        SIA.lexicon = new_lexicon
        self.sentiment = SIA.polarity_scores(self.for_sentimental_str)['compound']
```