

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДЕРЖАВНИЙ УНІВЕРСИТЕТ «ЖИТОМИРСЬКА ПОЛІТЕХНІКА»
ФАКУЛЬТЕТ ІНФОРМАЦІЙНО-КОМП'ЮТЕРНИХ ТЕХНОЛОГІЙ
КАФЕДРА ІНЖЕНЕРІЇ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

ПОЯСНЮВАЛЬНА ЗАПИСКА

до кваліфікаційної роботи освітнього ступеня «магістр»
за спеціальністю 121 «Інженерія програмного забезпечення»
(освітня програма «Інженерія програмного забезпечення»)

на тему:

«Система прогнозування складності англomовних текстів»

Виконала студентка групи ІПЗм-24-2
ТУРЧИН Анастасія Юріївна

Керівник роботи:
САВІЦЬКИЙ Роман Святославович

Рецензент:
ЄФРЕМОВ Юрій Миколайович

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДЕРЖАВНИЙ УНІВЕРСИТЕТ «ЖИТОМИРСЬКА ПОЛІТЕХНІКА»
ФАКУЛЬТЕТ ІНФОРМАЦІЙНО-КОМП'ЮТЕРНИХ ТЕХНОЛОГІЙ
КАФЕДРА ІНЖЕНЕРІЇ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

«ЗАТВЕРДЖУЮ»

зав. кафедри інженерії
програмного забезпечення
Тетяна ВАКАЛЮК
«18» вересня 2025 р.

ЗАВДАННЯ
на кваліфікаційну роботу

Здобувач вищої освіти: **ТУРЧИН Анастасія Юріївна**

Керівник роботи: **САВІЦЬКИЙ Роман Святославович**

Тема роботи: **«Система прогнозування складності англомовних текстів»**,
затверджена Наказом закладу вищої освіти від **«18» вересня 2025 р. № 491/с**

Вихідні дані для роботи: **процеси автоматичного аналізу англомовних текстів, а також методи та програмні засоби прогнозування рівня складності англомовних текстів відповідно до шкали CEFR.**

Консультанти з магістерської кваліфікаційної роботи із зазначенням розділів, що їх стосуються:

Розділ	Консультант	Завдання видав	Завдання прийняв
1	Савіцький Р.С.	22.09.2025р.	22.09.2025р.
2	Савіцький Р.С.	22.09.2025р.	22.09.2025р.
3	Савіцький Р.С.	22.09.2025р.	22.09.2025р.

РЕФЕРАТ

Магістерська атестаційна робота, що представлена на розгляд, присвячена темі “Система прогнозування складності англомовних текстів”. Робота включає вступ, три основні розділи, список використаних джерел із 31 найменування, 21 рисунка, 5 лістингів та 2 таблиць.

Ціль роботи полягає у розробці програмної системи автоматичного визначення рівня складності англомовних текстів відповідно до шкали CEFR на основі лінгвістичних та статистичних характеристик. Для досягнення поставленої мети було сформовано корпус текстів, виконано їх попередню обробку, розраховано комплекс лексичних, синтаксичних, морфологічних, семантичних метрик та індексів читабельності, а також розроблено модель машинного навчання на основі алгоритму Random Forest.

Науковий результат полягає в аналізі впливу лінгвістичних та статистичних метрик на якість автоматичного визначення рівня складності текстів та у формуванні обґрунтованої структури ознак для задачі класифікації за шкалою CEFR. Практичним результатом є розроблення веборієнтованої програмної системи, що реалізує повний цикл обробки текстових даних, забезпечує автоматичне визначення рівня складності тексту та формує узагальнену статистику використання сервісу.

Особливу увагу у роботі приділено застосуванню методів обробки природної мови, інженерії ознак та машинного навчання для задачі автоматичної класифікації рівнів складності текстів. Структура роботи включає огляд шкали CEFR, методів автоматичного аналізу текстів, підходів до побудови моделей машинного навчання для класифікації текстів, розробку програмної системи, а також аналіз результатів її функціонування та можливостей інтеграції в освітні процеси.

Ключові слова: автоматичне визначення складності тексту, CEFR, обробка природної мови, машинне навчання, Random Forest, лінгвістичні метрики, англомовні тексти, освітні технології, вебзастосунок.

ABSTRACT

The master's thesis submitted for review is devoted to the topic "System for Predicting the Complexity of English Texts". The thesis consists of an introduction, three main chapters, a list of references comprising 31 sources, 18 figures, and 2 tables.

The aim of the study is to develop a software system for the automatic determination of the complexity level of English texts according to the CEFR scale based on linguistic and statistical features. To achieve this goal, a text corpus was formed, preprocessing was performed, a set of lexical, syntactic, morphological, semantic metrics, as well as readability indices, were calculated, and a machine learning model based on the Random Forest algorithm was developed.

The scientific contribution of the work lies in analysing the impact of linguistic and statistical metrics on the performance of automatic text-difficulty prediction and in establishing a well-justified feature structure for CEFR-level classification. The practical outcome consists in the development of a web-oriented software system that implements the full text-processing pipeline, provides automatic determination of text difficulty level, and generates aggregated usage statistics of the service.

Special attention in the thesis is paid to the application of natural language processing methods, feature engineering, and machine learning for the task of automatic text complexity classification. The thesis structure includes an overview of the CEFR scale, methods of automatic text analysis, approaches to building machine learning models for text classification, the development of the software system, as well as an analysis of its performance and integration possibilities into educational processes.

Keywords: automatic text complexity assessment, CEFR, natural language processing, machine learning, Random Forest, linguistic metrics, English texts, educational technologies, web application.

ЗМІСТ

РЕФЕРАТ.....	3
ABSTRACT.....	4
ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ.....	6
ВСТУП.....	7
РОЗДІЛ 1. АНАЛІТИЧНИЙ ОГЛЯД ПРЕДМЕТНОЇ ОБЛАСТІ ТА МЕТОДІВ ОЦІНЮВАННЯ СКЛАДНОСТІ ТЕКСТІВ.....	9
1.1 CEFR та лінгвістичні основи оцінювання складності текстів.....	9
1.2 Методи та інструменти NLP для обробки текстів.....	11
1.3 Підходи машинного навчання до класифікації текстів за рівнем.....	14
1.4 Огляд існуючих систем автоматичного визначення рівня складності..	17
Висновки до розділу 1.....	21
РОЗДІЛ 2. РОЗРОБКА МЕТОДИКИ ОЦІНЮВАННЯ СКЛАДНОСТІ ТЕКСТУ.....	23
2.1 Формування корпусу англійських текстів і їх попередня обробка.....	23
2.2 Розрахунок лінгвістичних та статистичних метрик.....	27
2.3 Навчання моделі машинного навчання.....	32
2.4 Інтерпретація результатів моделі.....	35
Висновки до розділу 2.....	40
РОЗДІЛ 3. РОБОТА ПРОГРАМНОЇ СИСТЕМИ.....	42
3.1 Загальна організація та архітектурні рішення програмної системи.....	42
3.2 Визначення рівня CEFR.....	49
3.3 Інтеграція моделі в освітні процеси.....	56
Висновки до розділу 3.....	57
ВИСНОВКИ.....	59
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	60

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

AWL – Academic Word List

CEFR – Common European Framework of Reference for Languages

DFD – Data Flow Diagram

FKGL – Flesch-Kincaid Grade Level

JSON – JavaScript Object Notation

ML – Machine Learning

NLP – Natural Language Processing

POS – Part of Speech

REST API – Representational State Transfer Application Programming Interface

TTR – Type-Token Ratio

ВСТУП

Актуальність даної магістерської роботи зумовлена зростаючою роллю автоматизованих методів аналізу текстової інформації в сучасному освітньому середовищі. В умовах цифровізації освіти, розвитку дистанційного та змішаного навчання особливої ваги набуває проблема об'єктивного визначення рівня складності навчальних матеріалів відповідно до мовної підготовки здобувачів освіти.

Водночас активний розвиток методів обробки природної мови та машинного навчання створює передумови для автоматизації задачі оцінювання складності текстів на основі об'єктивних лінгвістичних і статистичних характеристик. Зростання обсягів англomовного освітнього контенту, що використовується в онлайн-курсах, освітніх платформах та системах самостійного навчання, підсилює потребу у створенні автоматизованих засобів прогнозування рівнів складності текстів. Це обґрунтовує доцільність розробки спеціалізованої програмної системи для автоматичного визначення рівня складності англomовних текстів за шкалою CEFR, що і визначає актуальність даної роботи.

Метою роботи є формалізація та побудова моделі автоматичного визначення складності англomовних текстів відповідно до шкали CEFR шляхом аналізу лінгвістичних та статистичних ознак і використання алгоритмів машинного навчання.

Завдання роботи. Для досягнення поставленої мети в роботі необхідно розв'язати такі основні завдання:

1. Проаналізувати шкалу CEFR та лінгвістичні характеристики для оцінювання складності текстів.
2. Дослідити сучасні методи обробки природної мови та підходи машинного навчання до класифікації текстів.
3. Сформувати корпус англomовних текстів та здійснити їх попередню обробку.
4. Побудувати модель машинного навчання для прогнозування рівнів складності текстів за шкалою CEFR.
5. Розробити веборієнтований програмну систему для автоматичного аналізу текстів
6. Провести експериментальну перевірку ефективності роботи моделі та проаналізувати отримані результати.

7. Розглянути можливості інтеграції розробленої системи в освітні процеси.

Об'єктом дослідження у магістерській роботі є процеси автоматичного аналізу англomовних текстів з використанням методів обробки природної мови та машинного навчання.

Предметом дослідження є методи та програмні засоби прогнозування рівня складності англomовних текстів відповідно до шкали CEFR на основі лінгвістичних та статистичних характеристик.

Методи дослідження, використані у магістерській роботі, охоплюють підходи обробки природної мови, статистичного аналізу текстових даних, машинного навчання із застосуванням алгоритму Random Forest, а також методи проектування програмних систем.

Наукова новизна магістерської роботи полягає в обґрунтованому доборі та адаптації лінгвістичних метрик і показників читабельності для задачі автоматичного визначення рівня складності англomовних текстів за шкалою CEFR, а також у комплексному аналізі їхнього впливу на якість класифікації. У роботі уточнено структуру ознак, що найбільш суттєво корелюють із рівнями CEFR, та продемонстровано можливість їх ефективного використання у рамках моделі машинного навчання.

Практичне значення роботи полягає у можливості використання розробленої програмної системи у навчальному процесі, зокрема для добору та адаптації навчальних матеріалів відповідно до рівня підготовки здобувачів освіти.

Публікації. За темою атестаційної магістерської роботи було опубліковано тези [1].

РОЗДІЛ 1. АНАЛІТИЧНИЙ ОГЛЯД ПРЕДМЕТНОЇ ОБЛАСТІ ТА МЕТОДІВ ОЦІНЮВАННЯ СКЛАДНОСТІ ТЕКСТІВ

1.1 CEFR та лінгвістичні основи оцінювання складності текстів

Common European Framework of Reference for Languages (CEFR) є міжнародним стандартом, розробленим Радою Європи для опису рівнів володіння іноземними мовами. Його поява була зумовлена необхідністю створення уніфікованої системи, яка б дозволяла порівнювати мовні компетенції в освітніх, академічних та професійних контекстах. CEFR охоплює шість основних рівнів — від A1 до C2, які згруповані у три категорії: початкову (A1–A2), середню (B1–B2) та просунуту (C1–C2). Кожен рівень містить опис того, що користувач мови здатний розуміти та продукувати, включно з характеристиками текстів, які він може опрацьовувати [2].

У контексті оцінювання складності текстів CEFR відіграє роль орієнтира, який дозволяє співвіднести властивості тексту з очікуваними можливостями читача певного рівня. Наприклад, для рівнів A1–A2 характерні короткі, чітко структуровані речення, обмежений словниковий запас та переважно повсякденна лексика. На рівнях B1–B2 читач здатний опрацьовувати більш різноманітні тексти, у яких трапляються складніші граматичні конструкції, ширша лексика та більша густина ключових понять. Рівні C1–C2 передбачають опрацювання високостилістичних текстів із розгорнутими аргументативними структурами, абстрактними значеннями та значним стилістичним варіюванням [3]. Таким чином, CEFR описує не лише компетенції користувача, але й дає непрямі орієнтири щодо параметрів текстів відповідної складності.

Оцінювання складності текстів ґрунтується на сукупності мовних характеристик, які відображають різні аспекти структури та змісту текстового матеріалу. Центральне місце в такій оцінці займають лінгвістичні показники, що дозволяють формалізувати властивості тексту, пов'язані з його словниковою різноманітністю, синтаксичною організацією, граматичними засобами, семантичним наповненням та загальним рівнем читабельності. Кожна група показників по-своєму впливає на сприйняття тексту й визначає обсяг когнітивних зусиль, необхідних для розуміння змісту, а отже – напряду пов'язана з рівнями володіння мовою, що описані у CEFR [4].

Лексичні показники характеризують словниковий склад тексту та ступінь його варіативності. На рівнях A1–A2 зазвичай домінує висока частка частотної, базової лексики, що відповідає повсякденним ситуаціям спілкування та обмеженій кількості тем. Зі зростанням рівня мовної компетентності збільшується частота вживання рідкісних, менш передбачуваних та більш абстрактних слів. Так само змінюється довжина слів, кількість складів та словникова різноманітність: у текстах рівнів B2–C1 спостерігається виразне збільшення частки маловживаних слів, академічної лексики й тематично спеціалізованих термінів [2]. Лексичні параметри, такі як середня довжина слова, частка слів академічного словника або співвідношення унікальних слів до загальної кількості, дають змогу кількісно оцінити рівень складності лексичного наповнення.

Синтаксичні показники відображають організацію речень і способи побудови граматичних зв'язків. Тексти початкових рівнів CEFR, як правило, характеризуються короткими, переважно простими реченнями з мінімальною кількістю підрядних конструкцій. У текстах середнього рівня (B1–B2) поступово зростає середня довжина речення, збільшується частка складносурядних конструкцій та з'являються перші систематичні прояви вкладених підрядних речень. У матеріалах рівнів C1–C2 синтаксис стає значно багатшим: речення містять декілька предикативних частин, ускладнені умовні, відносні та підрядні конструкції, які вимагають від читача глибшого опрацювання граматичної структури й здатності до відтворення складних логічних зв'язків [4]. Таким чином, синтаксичні показники дозволяють оцінити рівень структурної складності та когнітивного навантаження, що виникає при читанні.

Морфологічні характеристики доповнюють аналіз, розкриваючи деталі граматичної організації тексту. На низьких рівнях CEFR переважають прості дієслівні форми теперішнього та минулого часу, а також обмежене використання прикметників та прислівників. У текстах вищого рівня зростає різноманітність частин мови, збільшується кількість модальних дієслів, перфектних та прогресивних форм, що свідчить про здатність автора передавати складні часові, аспектуальні та модальні значення. Частковий розподіл частин мови, співвідношення службових та повнозначних слів, а також аспекти дієслівної системи дають змогу точніше оцінити граматичний профіль тексту.

Семантичні показники формують окремий рівень аналізу, пов'язаний зі змістовністю та концептуальною насиченістю тексту. Зі зростанням рівня CEFR підвищується абстрактність тем, pojawiaються лексеми з великою кількістю можливих значень, збільшується кількість концептуальних зв'язків між реченнями. Для високих рівнів характерною є тематична неоднорідність, складні логічні переходи, насиченість тексту аргументами та поясненнями. Важливими показниками є також семантична зв'язність і схожість речень: вони відображають наскільки логічним та цілісним є текст із погляду змістовних зв'язків [3].

Окрему групу складають показники читабельності, до яких належать формули Flesch Reading Ease, Flesch–Kincaid Grade Level, Gunning Fog Index та SMOG. Хоча вони були розроблені в інший історичний період і мають певні обмеження, їхня інформативність залишається високою, адже вони дозволяють швидко оцінити загальний рівень складності тексту на основі довжини речень та слів [5]. У поєднанні з лексичними та синтаксичними характеристиками ці формули формують комплексне уявлення про те, наскільки текст є легким чи важким для читача певного рівня мовної компетентності.

Системний аналіз таких лінгвістичних показників дає змогу кількісно оцінити рівень складності тексту, забезпечуючи об'єктивну основу для побудови автоматизованих систем класифікації. Оскільки кожен із показників відображає окремий аспект мовної компетентності, їхнє поєднання дозволяє моделі машинного навчання гнучко відтворювати структуру рівнів CEFR і точніше здійснювати диференціацію текстів за складністю.

1.2 Методи та інструменти NLP для обробки текстів

Сучасні методи обробки природної мови (NLP) становлять фундамент для автоматичного аналізу текстів та побудови моделей, що здатні оцінювати їхню складність. NLP поєднує у собі алгоритми лінгвістичного аналізу, математичні моделі й статистичні методи, що дають змогу перетворювати текст у формалізоване подання, придатне для обчислень [6]. У контексті визначення рівня складності текстів за шкалою CEFR методи NLP забезпечують можливість автоматично ідентифікувати лексичні, граматичні, синтаксичні та семантичні характеристики, які лежать в основі диференціації рівнів володіння мовою.

Першим етапом обробки тексту у будь-якій NLP-системі є токенизація, яка є процесом поділу тексту на окремі одиниці: слова, речення або символи [7]. Попри очевидну простоту, токенизація є важливим кроком, оскільки від коректності початкового розбиття залежить точність усіх наступних обчислень. Сучасні інструменти токенизації враховують морфологічні властивості слів, скорочення, пунктуацію та особливості різних мов, що особливо важливо для англійської через її неоднозначність та багатозначність [8].

Наступним базовим методом є лематизація – це приведення слів до їхньої початкової форми [7]. Лематизація дозволяє зменшити розмір словника тексту, узагальнити частотні характеристики та уникнути змішування варіантів однієї лексеми. У системах, що визначають складність тексту, лематизація допомагає точніше оцінювати словникове різноманіття, оскільки різні граматичні форми одного дієслова чи іменника трактуються як спільна мовна одиниця. Поряд із цим використовується частиномовне тегування, яке дозволяє автоматично визначити, до якої частини мови належить кожне слово у тексті. Це критично для аналізу граматичної організації, оцінювання частки іменників, дієслів, прикметників або службових слів, що є ключовими індикаторами рівнів CEFR.

Синтаксичний аналіз, зокрема побудова залежностей між словами, відіграє важливу роль у виявленні граматичної структури тексту. Алгоритми *dependency parsing* дають змогу визначити, які слова керують іншими, як формуються підрядні конструкції, яким є загальний синтаксичний каркас речень [8]. Для оцінювання складності текстів це дозволяє обчислювати такі показники, як глибина синтаксичного дерева, кількість предикативних частин у реченні чи частка складних речень. Моделі синтаксичного аналізу, реалізовані у бібліотеках сучасного NLP, здатні забезпечити стабільно високу точність, що необхідно для коректного відтворення синтаксичних особливостей тексту.

Семантичні методи NLP використовуються для оцінювання змістової насиченості та зв'язності тексту. Широко застосовуються векторні моделі, такі як *word embeddings* або контекстуальні векторні представлення слів. Вони дозволяють визначати семантичну близькість між словами та реченнями, що є важливим при аналізі логічної цілісності тексту та виявленні абстрактності понять [9]. Високі рівні CEFR, як правило, характеризуються більшою семантичною варіативністю, ширшим спектром тематичних зв'язків

та використанням багатозначних слів, що відображається на семантичних метриках.

Технічним підґрунтям для реалізації таких методів є програмні бібліотеки, які забезпечують ефективну роботу з текстовими даними. У поєднанні з Python та інструментами машинного навчання вони створюють гнучкий та ефективний інструментарій для побудови систем класифікації текстів.

Однією з найпоширеніших є spaCy, що поєднує високу швидкість обробки тексту з точними моделями для токенізації, тегування частин мови, лематизації та синтаксичного аналізу. Завдяки використанню попередньо навчених мовних моделей, spaCy дає змогу одразу отримувати повний лінгвістичний аналіз тексту та семантичні векторні подання [10].

Однією з найстаріших і найбільш універсальних є бібліотека NLTK, яка використовується для навчання й досліджень і містить широкий набір алгоритмів для токенізації, стемінгу, граматичного аналізу, а також доступ до лінгвістичних корпусів.

Для оцінювання читабельності часто застосовуються спеціалізовані інструменти, такі як TextStat, що реалізують класичні формули складності Flesch Reading Ease, Flesch–Kincaid Grade Level, Gunning Fog Index та SMOG [5].

У більш складних дослідженнях поширеним є використання системи Coh-Metrix [11], яка забезпечує розширений набір метрик когезії, синтаксичної складності й семантичної зв'язності. Сучасні підходи також включають використання бібліотек на кшталт HuggingFace Transformers, що дозволяють застосовувати контекстуальні моделі типу BERT або RoBERTa для глибшого семантичного аналізу текстів [12].

Таким чином, сукупність методів NLP забезпечує багаторівневий аналіз текстових даних, що охоплює від базових операцій розбиття тексту до високорівневих семантичних моделей [9]. Ці методи формують основу для обчислення лінгвістичних та статистичних метрик, на яких ґрунтується автоматичне визначення рівнів складності текстів. Інтеграція таких технологій у системи оцінювання забезпечує їхню точність, масштабованість та адаптивність до широкого спектра текстових жанрів і стилів.

1.3 Підходи машинного навчання до класифікації текстів за рівнем

Задача автоматичного визначення рівня складності англомовних текстів відповідно до шкали CEFR у термінах машинного навчання формулюється як задача багатокласової класифікації, у якій кожному тексту необхідно поставити у відповідність один із шести можливих класів (A1–C2). При цьому особливістю даної задачі є те, що шкала CEFR не є строго дискретною з погляду лінгвістичних характеристик, а має континуальний характер, де межі між сусідніми рівнями часто є розмитими [4].

На відміну від класичних задач тематичної класифікації текстів, у яких класи визначаються за змістовною належністю, у випадку CEFR-класифікації ключовим є не тема тексту, а сукупність його мовних властивостей. Тексти різних рівнів можуть описувати однакові теми, проте відрізнятися лексичною насиченістю, синтаксичною складністю, морфологічною варіативністю та семантичною абстрактністю [2]. Це ускладнює побудову чітких правил розмежування класів і зумовлює необхідність використання статистичних і навчальних підходів.

Додаткову складність становить суттєве перекриття лінгвістичних ознак між суміжними рівнями, зокрема A2 і B1 або B2 і C1. У таких випадках тексти можуть мати близькі значення окремих метрик, що призводить до підвищеної кількості помилок навіть у добре навчених моделях. Таким чином, задача визначення рівня складності тексту не зводиться до простого порогового аналізу окремих показників, а потребує врахування їхніх взаємозв'язків та нелінійних залежностей.

У контексті машинного навчання це означає необхідність побудови моделей, здатних працювати з багатовимірними векторами ознак і відтворювати складну структуру мовних характеристик, що лежать в основі рівнів CEFR. Саме такі вимоги визначають доцільність застосування алгоритмів машинного навчання для автоматичної класифікації складності текстів.

Перші підходи до автоматичної класифікації текстів у межах задач обробки природної мови ґрунтувалися на використанні класичних алгоритмів машинного навчання. До цієї групи належать логістична регресія, наївний баєсівський класифікатор та метод опорних векторів. Ці алгоритми зазвичай застосовуються у поєднанні з відносно простими представленнями тексту, такими як частотні вектори слів або n-грам [13].

Логістична регресія є лінійним класифікатором, який забезпечує інтерпретовані результати та добре масштабується для великих наборів даних. Проте її обмеженням є нездатність ефективно моделювати складні нелінійні залежності між лінгвістичними ознаками, що є характерними для задачі оцінювання складності текстів [13].

Наївний баєсівський підхід відзначається простотою реалізації та низькими обчислювальними витратами, однак його ключове припущення про незалежність ознак рідко виконується для лінгвістичних характеристик. У результаті такий підхід може втрачати точність у випадках, коли важливими є взаємодії між синтаксичними, морфологічними та семантичними показниками [11].

Метод опорних векторів здатний будувати складніші межі розділення між класами, особливо при використанні ядерних функцій. Водночас для багатокласових задач із великою кількістю різнорідних ознак його застосування ускладнюється вибором оптимальних параметрів та зростанням обчислювальної складності [11].

Загалом класичні алгоритми машинного навчання можуть бути ефективними для базових задач класифікації текстів, однак у контексті визначення рівнів CEFR їхні можливості є обмеженими через складну багатовимірну природу мовної складності.

Ансамблеві методи машинного навчання ґрунтуються на поєднанні результатів кількох базових моделей з метою підвищення загальної якості класифікації. Основна ідея таких підходів полягає у зменшенні впливу випадкових помилок окремих моделей та підвищенні стійкості результатів.

Одним із найбільш поширених ансамблевих алгоритмів є Random Forest, який базується на побудові множини дерев рішень, навчених на випадкових підмножинах даних і ознак. Такий підхід дозволяє ефективно працювати з лінгвістичними характеристиками різної природи та зменшувати ризик перенавчання. Random Forest добре пристосований до табличних представлень даних, у яких кожен текст описується набором числових метрик [13].

Важливою перевагою ансамблевих методів є їхня здатність враховувати нелінійні залежності між ознаками, що є критично важливим для задачі класифікації складності текстів. Крім того, такі моделі дозволяють оцінювати відносну важливість окремих ознак, що підвищує прозорість результатів і є особливо цінним у контексті освітніх застосувань.

До цієї ж групи належать градієнтні бустингові алгоритми, зокрема XGBoost та LightGBM, які забезпечують високу точність шляхом послідовного виправлення помилок попередніх моделей. Водночас ці підходи є більш чутливими до налаштування гіперпараметрів і можуть вимагати значних обчислювальних ресурсів, що не завжди є доцільним у дослідницьких проєктах середнього масштабу [13].

Окремий напрям у задачах автоматичної обробки текстів становлять моделі глибинного навчання, зокрема рекурентні нейронні мережі та трансформерні архітектури [12]. Вони здатні автоматично вивчати контекстні залежності між словами та реченнями без явного задання лінгвістичних ознак.

Рекурентні нейронні мережі та моделі типу LSTM дозволяють враховувати послідовну структуру тексту, що є корисним для аналізу довгих фрагментів. Трансформерні моделі, такі як BERT, забезпечують глибше контекстуальне представлення слів і демонструють високі результати у багатьох задачах обробки природної мови [14].

Разом із тим використання глибинних моделей у задачі класифікації складності текстів має низку обмежень. По-перше, вони потребують великих обсягів розмічених даних для навчання [14], що не завжди доступно для CEFR-орієнтованих корпусів. По-друге, результати таких моделей є складними для інтерпретації, що ускладнює їх використання в освітніх системах, де важливо пояснювати причини віднесення тексту до певного рівня складності.

Вибір моделі машинного навчання для автоматичного визначення рівня складності текстів має ґрунтуватися не лише на показниках точності, але й на низці практичних та методологічних критеріїв. До таких критеріїв належать інтерпретованість результатів, стійкість до шуму в даних, здатність працювати з обмеженими за обсягом корпусами та можливість інтеграції у веборієнтовані програмні системи.

Для освітніх застосувань особливого значення набуває прозорість роботи моделі, оскільки результати класифікації можуть використовуватися для добору навчальних матеріалів та прийняття педагогічних рішень. У цьому контексті ансамблеві методи, які поєднують достатньо високу точність з можливістю аналізу важливості ознак, виглядають найбільш збалансованим рішенням.

Оцінювання ефективності моделей класифікації базується на загальноприйнятих метриках, таких як точність, повнота, F1-міра та аналіз помилок. Додаткову цінність становить аналіз важливості ознак, який дозволяє виявити, які характеристики тексту найбільше впливають на прогнозований рівень. Це підвищує прозорість системи та сприяє її застосуванню у навчальних та аналітичних середовищах.

Методи машинного навчання створюють надійну та гнучку основу для побудови систем автоматичного визначення складності текстів. Вони дозволяють інтегрувати різноманітні мовні характеристики, моделювати їхні взаємозв'язки та забезпечувати точні прогнози рівня складності за CEFR. Сучасні підходи поєднують високу продуктивність з можливістю адаптації до різних типів текстів, що робить машинне навчання одним із центральних інструментів у розробці інтелектуальних систем аналізу мови [13].

1.4 Огляд існуючих систем автоматичного визначення рівня складності

Системи автоматичного визначення складності текстів посідають важливе місце у сучасних освітніх технологіях, забезпечуючи підтримку викладачів, студентів та розробників навчальних матеріалів у виборі адекватних за рівнем ресурсів [15]. Вони дозволяють швидко оцінити відповідність тексту конкретному мовному рівню, полегшуючи адаптацію контенту для навчальних платформ, тестових систем і педагогічних рекомендацій.

Перші системи автоматичного аналізу читабельності були переважно засновані на застосуванні класичних формул оцінювання, таких як Flesch Reading Ease, Flesch–Kincaid Grade Level або Gunning Fog Index. Ці формули враховують переважно довжину слів та речень, що обмежує їхню здатність коректно відображати реальну комунікативну складність тексту [5]. Вони не враховують лексичної різноманітності, граматичних структур, частоти використання рідкісних слів чи семантичних взаємозв'язків між реченнями. У зв'язку з цим формульні підходи поступово втратили актуальність для завдань, пов'язаних із визначенням рівнів CEFR, які вимагають аналізу набагато ширшого спектра характеристик.

Платформа Newsela є однією з найбільш відомих комерційних систем, що реалізують автоматизовану адаптацію текстів за рівнем складності для освітніх потреб. Основним призначенням Newsela є надання актуальних новинних матеріалів у декількох версіях, адаптованих до різного рівня

читацької підготовки, що робить її популярною серед викладачів та освітніх закладів (рис. 1.1).

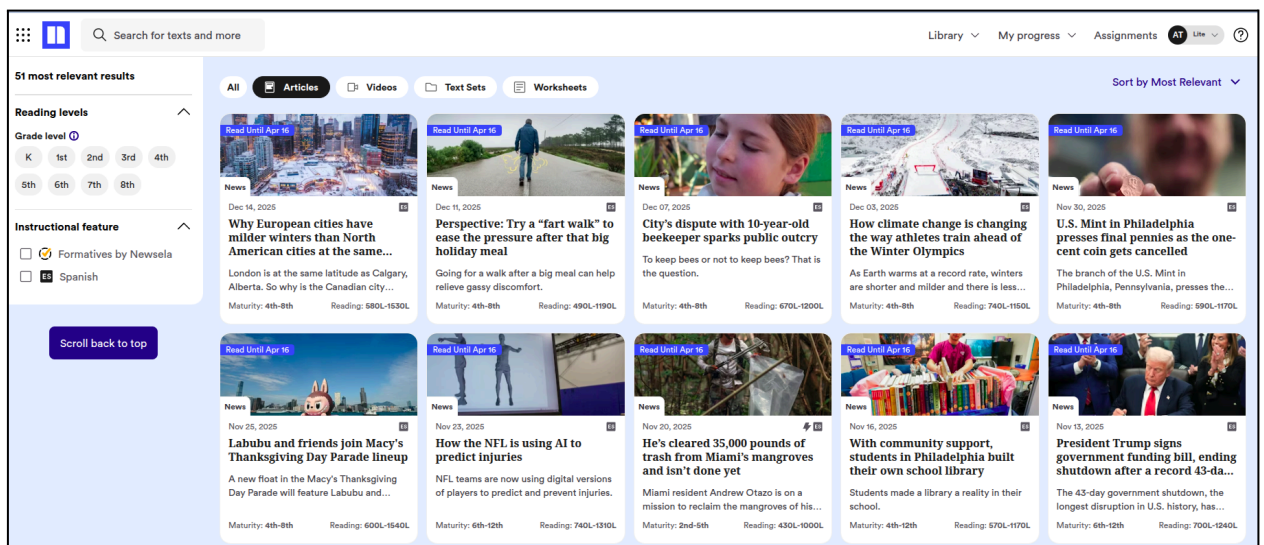


Рис. 1.1 Сторінка з текстами відповідно до рівня читацької підготовки

Функціонування Newsela ґрунтується на поєднанні алгоритмів машинного навчання та експертної лінгвістичної редакції. Вихідні тексти, як правило, створюються або відбираються редакційною командою, після чого адаптуються для кількох рівнів складності. У процесі адаптації враховуються лексичні характеристики, довжина речень, синтаксична структура та загальна читабельність тексту. Хоча точні алгоритми класифікації є закритими, відомо, що система використовує статистичні та мовні ознаки для підтримки рівневої диференціації контенту. З погляду функціональних можливостей Newsela дозволяє користувачам отримувати тексти, попередньо адаптовані до різних рівнів складності; обирати відповідну версію тексту для певної аудиторії; інтегрувати матеріали у навчальні курси та освітні платформи.

Водночас Newsela має низку суттєвих обмежень. По-перше, система не надає прозорого механізму визначення рівня складності та не дозволяє аналізувати вплив окремих лінгвістичних ознак на результат. По-друге, Newsela не працює з довільними текстами користувача, а обмежується власним корпусом матеріалів. По-третє, рівні складності не завжди явно співвідносяться зі шкалою CEFR, що ускладнює використання платформи в академічних дослідженнях.

Cambridge English Text Inspector є спеціалізованим онлайн-інструментом, розробленим під егідою Cambridge Assessment English, який безпосередньо орієнтований на аналіз текстів відповідно до стандартів CEFR (рис. 1.2). Основною метою цього інструмента є допомога викладачам,

авторам навчальних матеріалів та екзаменаторам у попередній оцінці лексичної відповідності текстів певному рівню володіння мовою.

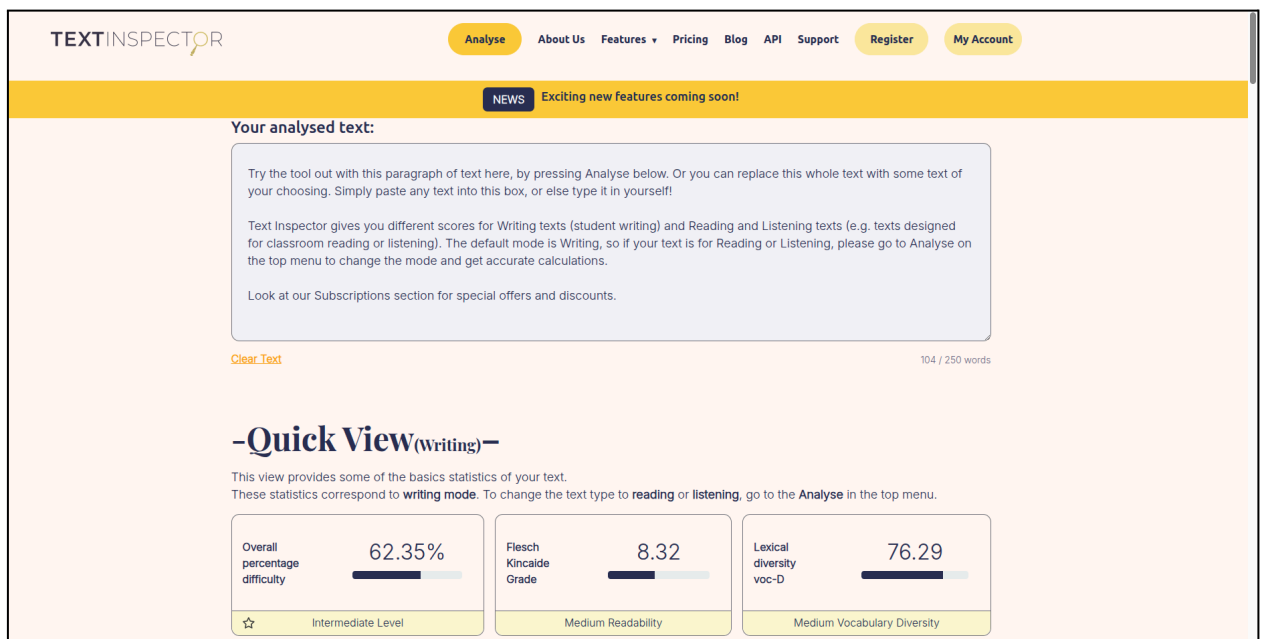


Рис. 1.2 Сторінка аналізу складності тексту

Принцип роботи Text Inspector базується на лексичному аналізі тексту з використанням великих корпусів навчальної англійської мови, зокрема Cambridge Learner Corpus. Система визначає частотність слів, їх належність до певних рівнів CEFR та співвідношення базової, середньочастотної та рідкісної лексики. У результаті користувач отримує оцінку ймовірного рівня складності тексту, а також деталізований список слів із зазначенням їх рівневої приналежності.

Серед основних можливостей Text Inspector можна виділити: аналіз довільного тексту користувача, визначення ймовірного рівня CEFR, лексичну декомпозицію тексту за рівнями, виявлення слів, що перевищують заданий рівень складності. Однак дана система має низку істотних обмежень. Найважливішим з них є обмежена увага до синтаксичних, морфологічних та семантичних характеристик тексту. Оцінювання складності фактично зводиться до аналізу словникового складу, що не дозволяє адекватно враховувати граматичну складність або логічну структуру тексту. Крім того, внутрішні алгоритми класифікації не є відкритими, що унеможливорює глибоку інтерпретацію результатів.

Система Coh-Metrix є дослідницьким інструментом, створеним для глибокого аналізу когнітивних, лінгвістичних та дискурсивних характеристик текстів. Вона широко використовується у психолінгвістиці, прикладній

лінгвістиці та дослідженнях читабельності для аналізу когезії, синтаксичної складності та семантичної зв'язності тексту (рис. 1.3).

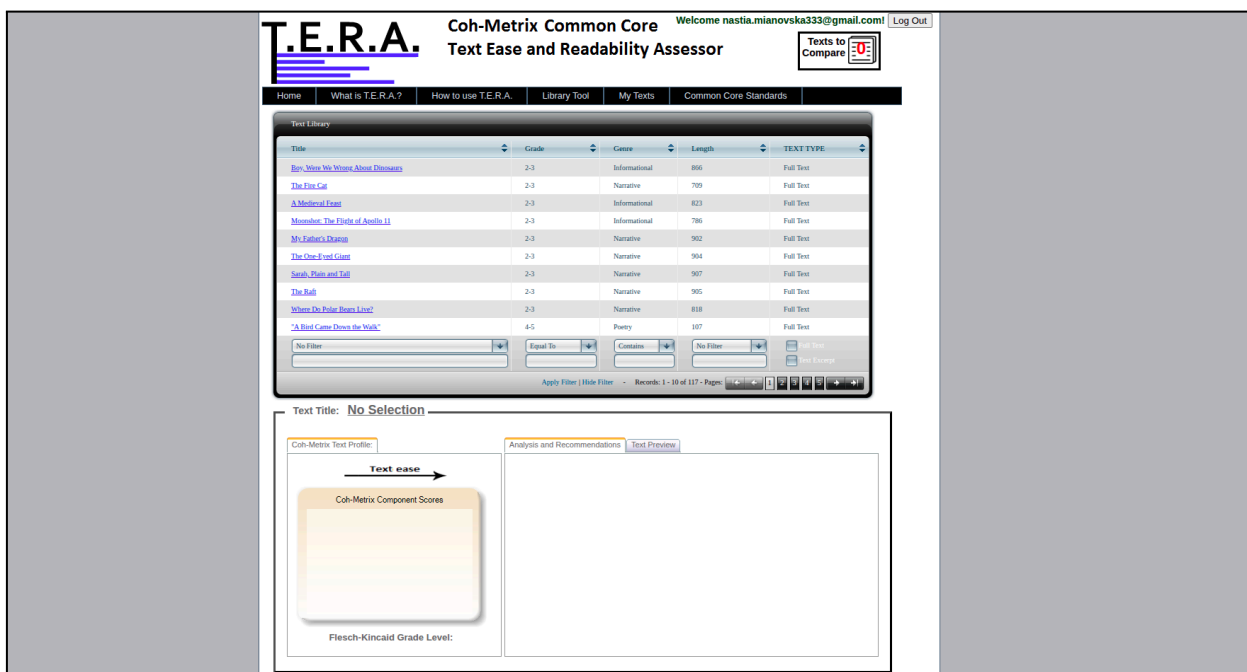


Рис. 1.3 Сторінка порівняння та аналізу текстів

Coh-Metrix функціонує на основі великого набору метрик, що охоплюють лексичну різноманітність, синтаксичну структуру, частотність слів, когезійні зв'язки між реченнями, семантичну подібність та абстрактність. Система використовує поєднання корпусних ресурсів, статистичних методів і семантичних моделей для обчислення понад сотні показників. Це дозволяє отримати надзвичайно детальний мовний профіль тексту.

Попри високу аналітичну цінність, Coh-Metrix не виконує прямої класифікації текстів за шкалою CEFR. Отримані метрики потребують подальшої інтерпретації або використання в окремих моделях машинного навчання. Крім того, система не орієнтована на кінцевого користувача-нефахівця та не забезпечує повного циклу “текст → рівень складності”.

Під назвою CEFR-Classifіer зазвичай об'єднують низку дослідницьких рішень і прототипів, спрямованих на автоматичну класифікацію текстів за рівнями CEFR із використанням машинного навчання. Такі системи будуються на корпусах розмічених текстів, зокрема EFCAMDAT, OneStopEnglish або подібних освітніх колекціях.

Архітектура CEFR-класифікаторів зазвичай включає попередню обробку тексту, обчислення лінгвістичних ознак, навчання класифікатора, оцінювання результатів за стандартними метриками. Основною перевагою таких підходів є інтерпретованість та відкритість моделей, а також можливість аналізу важливості ознак. Водночас більшість реалізацій мають характер наукових прототипів і не супроводжуються повноцінною програмною системою з вебінтерфейсом, API або механізмами інтеграції в освітні платформи.

Сучасні системи автоматичного визначення складності текстів демонструють значну різноманітність підходів, від простих формул читабельності до складних багаторівневих моделей машинного навчання. Їхня еволюція відображає розвиток інструментів обробки мови та зростання потреб освітніх платформ у точному визначенні відповідності текстів рівням CEFR [15].

Висновки до розділу 1

У першому розділі було здійснено комплексний аналіз теоретичних засад оцінювання складності англомовних текстів та сучасних підходів до автоматичного визначення рівня володіння мовою. Розглянуто роль CEFR як міжнародного стандарту, що забезпечує уніфіковану систему опису мовних компетентностей і створює концептуальну основу для розроблення алгоритмів оцінювання складності текстів.

Показано, що шість рівнів CEFR акумулюють як лексичні, так і граматичні, комунікативні та функційні характеристики мовлення, а тому є придатною структурою для алгоритмізації аналізу текстів.

Описано лінгвістичні метрики, які найчастіше застосовують під час оцінювання складності: лексичне різноманіття, частка рідкісної лексики, морфологічна структура, синтаксична складність, когнітивні індикатори та показники читабельності. Підкреслено, що ці показники відображають різні аспекти мовної складності, а їх поєднання дозволяє створити більш надійну та інтерпретовану модель.

Розглянуто ключові методи NLP та алгоритми машинного навчання, що застосовуються для класифікації текстів за рівнем складності. Крім того, відзначено розвиток нейронних моделей, що використовують семантичні представлення слів і здатні автоматично виявляти складні структурні закономірності текстів.

Проведено огляд існуючі системи автоматичного визначення рівня текстів. Порівняльний аналіз продемонстрував, що більшість таких рішень зосереджуються або на читабельності, або на окремих лінгвістичних характеристиках, а тому не завжди відображають реальну складність тексту відповідно до вимог CEFR. Це підкреслює потребу в системах, що поєднують багатовимірні метрики, сучасні NLP-інструменти та машинне навчання.

Узагальнюючи результати розділу, можна стверджувати, що завдання автоматичного визначення рівня складності текстів є міждисциплінарним і вимагає синтезу лінгвістичних знань, обчислювальних методів та статистичного аналізу. Проведений огляд дозволяє сформулювати теоретичну базу, необхідну для розроблення власної методики та програмної системи, що буде представлена у наступних розділах роботи.

РОЗДІЛ 2. РОЗРОБКА МЕТОДИКИ ОЦІНЮВАННЯ СКЛАДНОСТІ ТЕКСТУ

2.1 Формування корпусу англомовних текстів і їх попередня обробка

Для побудови системи автоматичного визначення рівня складності текстів відповідно до шкали CEFR було сформовано спеціалізований корпус англомовних текстів. Корпус слугує базою для обчислення лінгвістичних та статистичних характеристик, а також для подальшого навчання і перевірки моделі машинного навчання. Саме якість вхідних даних значною мірою визначає точність і узагальнювальну здатність побудованої моделі, тому формування корпусу розглядається як один із ключових етапів дослідження [16].

На відміну від загальних текстових колекцій, корпус у межах даного дослідження орієнтований на представлення різних рівнів мовної складності відповідно до CEFR. Це зумовлює потребу не лише у збиранні достатньої кількості текстів, але й у забезпеченні їх коректної відповідності заявленому рівню володіння мовою [17].

Тексти для корпусу було зібрано з відкритих освітніх джерел, зокрема навчальних платформ, онлайн-курсів, адаптованих підручників, а також ресурсів для вивчення англійської мови, що містять матеріали з маркованими рівнями CEFR. Корпус містить тексти рівнів A1, A2, B1, B2, C1 та C2.

Під час формування корпусу було виявлено, що кількість текстів у різних рівнях CEFR відрізняється. Зокрема, найменший обсяг даних припадає на рівень A2, тоді як найбільша кількість текстів представлена рівнем B1. Такий розподіл зумовлений специфікою доступних відкритих джерел, у яких тексти середнього рівня складності (B1) представлені значно ширше, ніж матеріали початкового рівня [16]. Попри це, сформований корпус характеризується помірним рівнем дисбалансу, що дозволяє застосовувати алгоритми машинного навчання без суттєвого ризику зміщення результатів. Для зниження впливу нерівномірності даних у процесі навчання моделі використовувалися методи внутрішнього балансування, закладені в обраному алгоритмі класифікації.

На рисунку 2.1 подано розподіл кількості текстів у сформованому корпусі відповідно до рівнів шкали CEFR. Як видно з діаграми, найбільша кількість текстових матеріалів припадає на рівень B1, натомість найменшу – тексти рівня A2. Рівні A1, C1 та C2 представлені приблизно рівномірною

кількістю текстів, що забезпечує достатню репрезентативність корпусу на початковому та просунутому етапах володіння мовою.

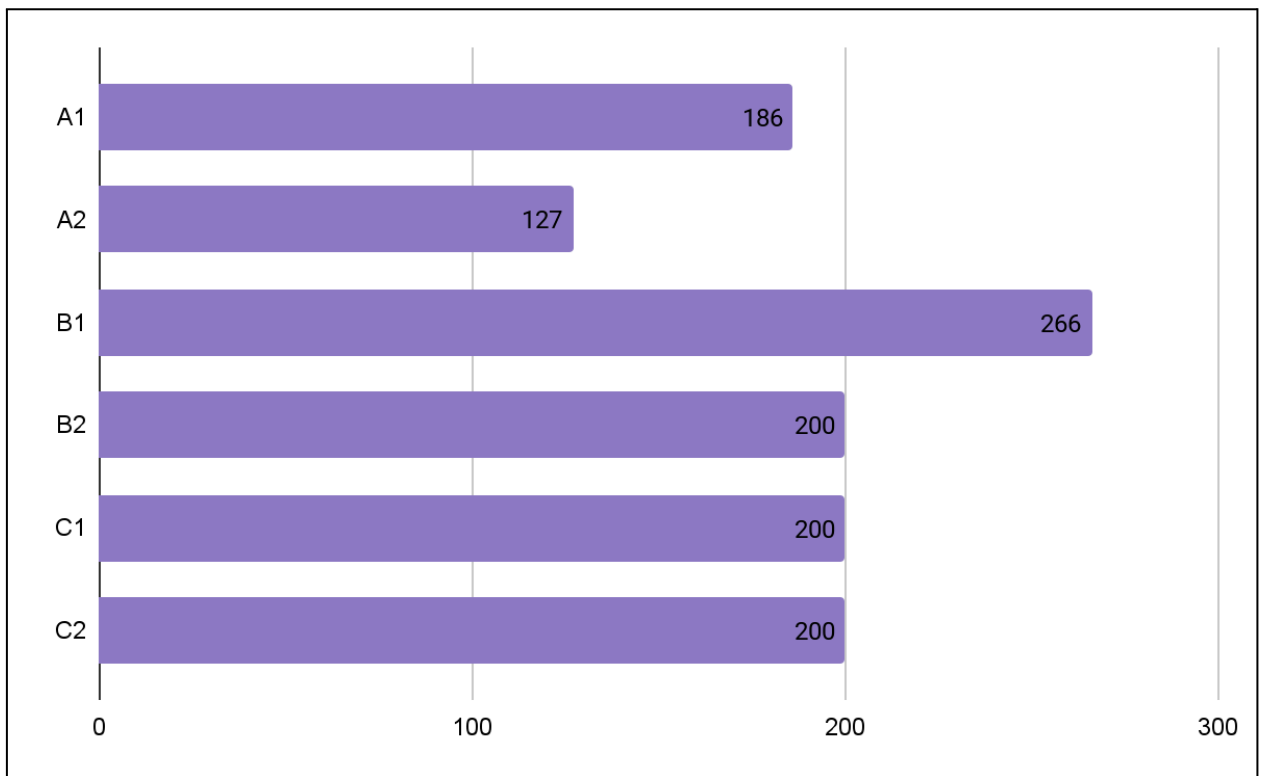


Рис 2.1 Розподіл кількості текстів у сформованому корпусі

Загалом отриманий розподіл характеризується помірним дисбалансом, який не є критичним для навчання моделі машинного навчання та був врахований під час побудови класифікатора. Представлена структура корпусу дозволяє забезпечити коректне формування ознак та подальше визначення рівня складності англомовних текстів.

Основними критеріями відбору текстів були: відповідність рівня CEFR, використання сучасної літературної англійської мови, відсутність значних фрагментів іншомовного тексту, HTML-розмітки, таблиць та спеціальних символів. Мінімальна довжина тексту встановлювалася на рівні 150 символів, оскільки коротші фрагменти не дозволяють достеменно визначити лінгвістичні характеристики. Максимальна довжина тексту обмежувалася 8000 символами з метою уніфікації обсягів даних та зниження обчислювального навантаження.

Після формування початкового корпусу текстів наступним етапом стала їх попередня обробка, яка є необхідною умовою для коректного обчислення лінгвістичних характеристик та подальшого застосування методів машинного навчання [9]. Сирі текстові дані, зібрані з відкритих джерел, зазвичай містять

значну кількість неоднорідних елементів, зокрема зайві пробіли, службові символи, фрагменти розмітки, нестандартизовану пунктуацію та випадкові числові вставки. Наявність таких елементів може суттєво спотворювати значення метрик і негативно впливати на результати класифікації [18].

На рис. 2.2 зображено схему етапів попередньої обробки англomовних текстів, що застосовувалася в межах даного дослідження.

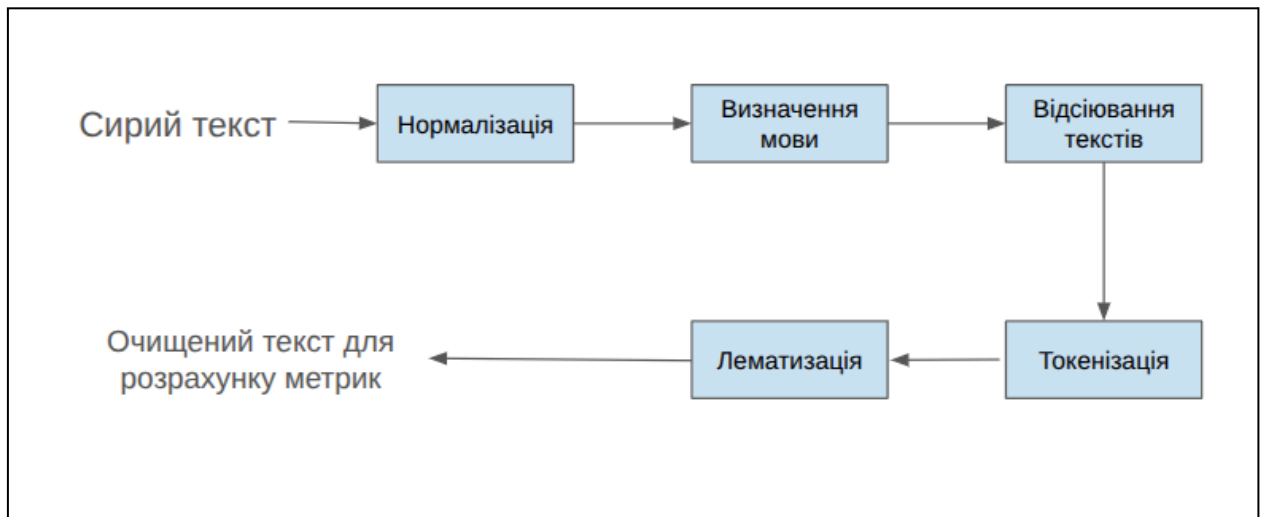


Рис. 2.2 Етапи попередньої обробки текстів

На першому етапі попередньої обробки здійснювалась нормалізація текстів. Вона включала приведення всіх символів до одного регістру, видалення надлишкових пробілів, уніфікацію розділових знаків та усунення некоректних або неінформативних символів. Такий підхід дозволив усунути формальні відмінності між текстами та забезпечити їхню однакову структуру для подальшого аналізу.

Наступним кроком була мовна фільтрація, яка застосовувалася для запобігання потраплянню до корпусу неанглomовних текстів. Для автоматичної перевірки мовної приналежності текстів у роботі використано бібліотеку *langdetect*, яка реалізує статистичний підхід до ідентифікації мови на основі частотних характеристик символів і послідовностей букв. Усі тексти, що не були ідентифіковані як англomовні, виключалися з подальшого аналізу. Це дозволило забезпечити однорідність корпусу та коректну роботу мовних моделей, орієнтованих на обробку англійської мови.

Після нормалізації та мовної фільтрації тексти піддавалися токенизації та лематизації із застосуванням інструментів бібліотеки *spaCy*. Токенізація дала змогу розбити текст на окремі лексеми, тоді як лематизація забезпечила приведення слів до початкових форм. Це, своєю чергою, стало основою для

коректного розрахунку лексичних, морфологічних та синтаксичних показників, бо дозволило уникнути впливу лексемних варіацій на значення ознак [19].

Для наочності розглянемо приклад результатів попередньої обробки тексту.

Вихідний фрагмент тексту:

“Students were reading different books and discussing their ideas. Some of them were writing notes, others were preparing presentations and explaining the topics to their classmates.”

Після нормалізації:

“students were reading different books and discussing their ideas. some of them were writing notes others were preparing presentations and explaining the topics to their classmates.”

Після токенізації:

[“students”, “were”, “reading”, “different”, “books”, “and”, “discussing”, “their”, “ideas”, “.”, “some”, “of”, “them”, “were”, “writing”, “notes”, “others”, “were”, “preparing”, “presentations”, “and”, “explaining”, “the”, “topics”, “to”, “their”, “classmates”, “.”]

Після лематизації:

[“student”, “be”, “read”, “different”, “book”, “and”, “discuss”, “their”, “idea”, “.”, “some”, “of”, “they”, “be”, “write”, “note”, “other”, “be”, “prepare”, “presentation”, “and”, “explain”, “the”, “topic”, “to”, “their”, “classmate”, “.”]

У результаті виконання всіх етапів попередньої обробки було сформовано очищений та уніфікований корпус англomовних текстів, придатний для подальшого розрахунку лінгвістичних метрик і навчання моделі машинного навчання. Застосування комплексного підходу до попередньої обробки дозволило підвищити якість вхідних даних та забезпечити стабільність і відтворюваність результатів дослідження.

2.2 Розрахунок лінгвістичних та статистичних метрик

Ключовим етапом для побудови ефективної моделі машинного навчання є кількісне подання мовних характеристик тексту у вигляді набору числових ознак. Модель класифікації не працює безпосередньо з текстом у вигляді послідовності символів [20], тому кожен текст описується вектором лінгвістичних та статистичних метрик, які відображають його лексичні, синтаксичні, морфологічні, семантичні характеристики, а також рівень читабельності.

Лексичні метрики описують словниковий склад тексту та особливості вживання лексики. Показник середньої довжини слова (`lex_avg_word_len`) обчислюється як середня кількість символів у словах, що дає змогу оцінити, наскільки текст насичений короткими або довгими лексичними одиницями. Коефіцієнт лексичної різноманітності на основі лем (`lex_ttr_lemma`) визначається як відношення кількості унікальних лем до загальної кількості слів і відображає різноманітність словника: чим вищий показник, тим більш варіативною є лексика тексту. Частка стоп-слів (`lex_share_stop`) показує, яку частину тексту становлять службові слова, що характерно для простих текстів з невисокою інформаційною щільністю. Додатково обчислюються частка числівників і символів (`lex_share_num_symbol`), що дозволяє врахувати наявність числових даних та спеціальних позначень, а також частка рідкісних або маловживаних слів (`lex_share_oov`) на основі частотних характеристик. Показник середньої кількості складів на слово (`lex_avg_syll_per_word`) дозволяє опосередковано оцінити фонетичну складність лексики, що також корелює з рівнем складності тексту.

Для практичної реалізації лексичних характеристик у системі використовується програмний алгоритм обчислення частки слів з Academic Word List (`lex_share_awl`), приклад якого наведено у відповідному фрагменті коду. Частка AWL визначається як відношення кількості лем, присутніх у словнику Academic Word List, до загальної кількості відібраних слів у тексті. У випадках, коли текст не містить лексичних одиниць або словник AWL недоступний, значення метрики коректно фіксується як нульове. Такий підхід дозволяє кількісно оцінити рівень академічної насиченості тексту та використовувати відповідний показник як одну з інформативних лексичних ознак у моделі класифікації рівнів складності за шкалою CEFR.

```
def share_awl(doc, awl_path: Optional[str] = None) -> float:
    words = _alpha_tokens(doc)
    if not words:
        return 0.0
    awl = _load_awl(awl_path)
    if not awl:
        return 0.0
    hits = sum(1 for t in words if t.lemma_.lower() in awl)
    return hits / len(words)
```

Синтаксичні метрики характеризують будову речень, типи синтаксичних структур та ступінь їх ускладнення. Середня довжина речення (`syn_avg_sentence_length`) є одним з базових показників складності: довші речення, як правило, містять більше підрядних конструкцій і складніші зв'язки між частинами. Показник середньої кількості предикативних більш довгі на речення (`syn_avg_clause_per_sentence`) відображає, скільки предикативних центрів міститься в одному реченні, що дозволяє відрізнити прості та складні синтаксичні структури. Частка пасивних конструкцій (`syn_share_passive_sentences`) дає змогу врахувати вживання пасивного стану, що є характерною рисою більш формальних та академічних текстів. Показник середньої глибини синтаксичного дерева залежностей (`syn_avg_dependency_depth`) відображає ступінь вкладеності граматичних зв'язків у реченні. Окрім цього, обчислюються частка підрядних сполучників (`syn_share_sub_conjs`) та середня кількість координаційних зв'язків на речення (`syn_avg_coord_per_sentence`), що дає можливість оцінити, наскільки активно автор використовує складнопідрядні та складносурядні конструкції.

Для ілюстрації практичної реалізації синтаксичних характеристик у системі використовується алгоритм обчислення частки складних речень (`syn_share_complex_sentences`), приклад якого наведено у відповідному фрагменті програмного коду. У межах цього підходу текст попередньо сегментується на окремі речення, після чого для кожного з них визначається кількість дієслівних предикатів на основі морфосинтаксичних тегів. Речення вважається складним у випадку, якщо воно містить більше одного дієслівного або допоміжного дієслівного центру, що свідчить про наявність складних синтаксичних зв'язків. Отримане значення нормалізується відносно загальної кількості речень у тексті, що забезпечує порівнюваність показника для текстів різної довжини. Така метрика дозволяє кількісно оцінити ступінь

синтаксичної ускладненості тексту та є інформативною ознакою для розмежування рівнів складності за шкалою CEFR.

```
def share_complex_sentences(doc: Doc) -> float: 1 usage anastasia-mia
    sentences = list(doc.sents)
    if not sentences:
        return 0.0
    complex_sentences_count = sum(1 for s in sentences if sum(1 for t in s if t.pos_ in ("VERB", "AUX")) > 1)
    return complex_sentences_count / len(sentences)
```

Морфологічні метрики зосереджені на частинах мови, граматичних формах та словотворчій складності. Для тексту обчислюється частка іменників (`morph_share_nouns`), дієслів (`morph_share_verbs`), прикметників (`morph_share_adj`), прислівників (`morph_share_adv`), займенників (`morph_share_pronouns`), власних назв (`morph_share_propn`), допоміжних дієслів (`morph_share_aux`) та модальних дієслів (`morph_share_modals`). Такі показники дозволяють описати розподіл частин мови й відрізнити, наприклад, більш описові тексти від наративних. Окремий блок метрик пов'язаний із часовими формами дієслів: частка минулого часу (`morph_tense_past_share`), теперішнього часу (`morph_tense_present_share`), а також частка перфектних, тривалих, перфектно-тривалих та майбутніх форм (`morph_share_perfect`, `morph_share_progressive`, `morph_share_perfect_progressive`, `morph_share_future`). Ці показники враховують граматичну складність дієслівного ладу, зокрема вживання складних часових форм, що є типовим для текстів високого рівня. Показник співвідношення повнозначних та службових частин мови (`morph_content_function_ratio`) відображає інформаційну насиченість тексту, а середня кількість морфем на слово (`morph_avg_morphemes_per_word`) дає змогу врахувати словотворчу складність лексики на основі наявності префіксів, суфіксів, граматичних категорій часу, числа та ступеня порівняння.

Для демонстрації практичної реалізації морфологічних показників у системі використовується алгоритм обчислення частки перфектних дієслівних форм (`morph_share_perfect`), приклад якого наведено у відповідному фрагменті програмного коду. У межах цього алгоритму аналіз здійснюється на рівні дієслівних предикатів тексту, що дозволяє врахувати лише граматично значущі часові конструкції. Перфектні форми визначаються на основі поєднання допоміжного дієслова *have* з дієприкметником минулого часу, при цьому перфектно-тривалі конструкції виключаються з розрахунку. Отримане значення нормалізується відносно загальної кількості предикатів,

що забезпечує порівнюваність показника для текстів різної довжини. Застосування цієї метрики дозволяє кількісно оцінити граматичну складність дієслівного ладу, оскільки вживання перфектних форм є характерним для текстів більш високого рівня складності за шкалою CEFR.

```
def morph_share_perfect(doc: Doc) -> float: 1 usage  🍷 anastasia-mia
    preds = _predicates(doc)
    if not preds:
        return 0.0
    perfect_count = 0
    for v in (t for t in doc if t.pos_ == "VERB"):
        has_have = _has_aux_link(v, lemmas={"have"})
        has_be = _has_aux_link(v, lemmas={"be"})
        is_vbn = (v.tag_ == "VBN")
        is_perfect_prog = has_have and has_be and (v.tag_ == "VBG")
        if has_have and is_vbn and not is_perfect_prog:
            perfect_count += 1
    return perfect_count / len(preds)
```

Семантичні метрики орієнтовані на оцінку значеннєвих характеристик слів та зв'язності тексту. Середня частота слів за шкалою Zipf (`sem_mean_zipf`) відображає, наскільки вжиті слова є поширеними в мові загалом, тоді як частка рідкісних (`sem_share_rare_zipf_lt_4`) та дуже рідкісних (`sem_share_very_rare_zipf_lt_3`) слів фіксує використання лексики низької частотності, що характерно для текстів високої складності. Показник середньої полісемії (`sem_avg_polysemy`) оцінює кількість значень, які має слово відповідно до лексикографічних ресурсів, а середня глибина ієрархії гіперонімів (`sem_avg_hyponym_depth`) характеризує рівень абстрактності та узагальненості вживаної лексики. Окремий набір метрик описує семантичну зв'язність тексту на рівні речень: середня, мінімальна та стандартне відхилення семантичної подібності між сусідніми реченнями (`sem_avg_sent_sim`, `sem_min_sent_sim`, `sem_std_sent_sim`) дають змогу оцінити, наскільки послідовно та логічно розгортається зміст. Показник розсіювання векторів слів у семантичному просторі (`sem_word_vector_dispersion`) відображає, наскільки тематично однорідним або різноплановим є текст.

Для практичної реалізації семантичних характеристик у системі використовується алгоритм обчислення показника розсіювання векторів слів у семантичному просторі (`sem_word_vector_dispersion`), фрагмент якого

наведено у відповідному прикладі програмного коду. У межах цього підходу для повнозначних слів тексту, що мають векторні подання, формується спільний центр мас у багатовимірному семантичному просторі. Далі для кожного слова визначається ступінь відхилення його векторного представлення від цього центру на основі косинусної міри подібності, після чого отримані значення агрегуються у вигляді середнього показника. Вищі значення цієї метрики свідчать про більшу семантичну різноплановість і тематичну неоднорідність тексту, тоді як нижчі значення характерні для текстів зі стабільною, цілісною тематикою. Таким чином, показник `sem_word_vector_dispersion` дозволяє кількісно оцінити семантичну складність і зв'язність тексту та використовується як інформативна ознака у моделі класифікації рівнів складності за шкалою CEFR.

```
def sem_word_vector_dispersion(doc: Doc) -> float: 1 usage  anastasia-mia
    toks = [t for t in _content_tokens(doc) if t.has_vector]
    if len(toks) < 2:
        return 0.0
    dim = len(toks[0].vector)
    centroid = [0.0] * dim
    for t in toks:
        v = t.vector
        for i in range(dim):
            centroid[i] += v[i]
    centroid = [x / len(toks) for x in centroid]

    def cos_sim(v, c):  anastasia-mia
        return _cosine(v, c)
    sims = [cos_sim(t.vector, centroid) for t in toks]

    dists = [1.0 - s for s in sims]
    return float(mean(dists))
```

Окрему групу становлять метрики читабельності, які обчислюються на основі відомих формул автоматичного оцінювання складності тексту. До них належать індекс Flesch Reading Ease (`read_flesch`), шкала Flesch–Kincaid Grade Level (`read_fkgl`), індекс Gunning Fog (`read_fog`), показник SMOG (`read_smog`) та індекс Dale–Chall (`read_dale_chall`). Ці метрики комбінують інформацію про середню довжину речень, кількість складів на слово та частку «важких» слів, що дозволяє отримати наближену оцінку рівня читабельності тексту в термінах шкільних класів або балів [21]. У рамках даного дослідження вони використовуються не як самостійна оцінка рівня, а як додаткові ознаки, які доповнюють лексичні, синтаксичні, морфологічні та семантичні показники.

Усі розглянуті метрики обчислюються для кожного тексту у вигляді набору числових значень, які потім об'єднуються у впорядкований вектор

ознак згідно з фіксованим порядком. Такий вектор є єдиним входом для моделі машинного навчання, що виконує класифікацію тексту за рівнями CEFR. Таким чином, розрахунок лінгвістичних та статистичних метрик виступає ключовою ланкою між сирим текстом та формальним представленням, необхідним для побудови ефективної моделі прогнозування складності тексту.

2.3 Навчання моделі машинного навчання

Після формування корпусу текстів та обчислення набору лінгвістичних і статистичних метрик наступним етапом дослідження стало навчання моделі машинного навчання для автоматичної класифікації рівня складності тексту відповідно до шкали CEFR. На цьому етапі кожен текст уже представлений не у вигляді послідовності слів, а у вигляді числового вектора ознак, який містить лексичні, синтаксичні, морфологічні, семантичні характеристики та показники читабельності.

Як модель класифікації у даній роботі було обрано метод Random Forest, що належить до ансамблевих алгоритмів машинного навчання. Його принцип роботи полягає у побудові великої кількості незалежних дерев рішень, кожне з яких навчається на випадковій підмножині навчальних даних та випадковому наборі ознак. Остаточний результат класифікації визначається шляхом голосування серед усіх дерев ансамблю. Такий підхід дозволяє зменшити вплив перенавчання, підвищити стійкість моделі до шумів у даних та забезпечити стабільні результати навіть за наявності корельованих ознак [20].

Вибір алгоритму Random Forest зумовлений кількома чинниками. По-перше, він добре працює з табличними даними та числовими ознаками різної природи, що повністю відповідає структурі сформованого вектора лінгвістичних метрик. По-друге, алгоритм не потребує складної нормалізації або масштабування даних, що спрощує етап підготовки навчальної вибірки. По-третє, Random Forest забезпечує достатньо високу точність класифікації та дозволяє оцінювати важливість окремих ознак, що є корисним для подальшої інтерпретації результатів [22].

У лістингу поданому нижче наведено фрагмент програмної реалізації процесу навчання моделі машинного навчання, призначеної для автоматичного визначення рівня складності англійських текстів відповідно до шкали CEFR. На першому етапі формується матриця ознак X , що містить

обчислені лінгвістичні та статистичні метрики текстів, та вектор цільових значень Y, який відповідає рівням складності за шкалою CEFR. Наступними кроками є розподіл вибірки на навчальну та тестову підмножини та навчання моделі на основі обраного алгоритму RandomClassifier. Після навчання моделі на навчальній вибірці здійснюється прогнозування рівнів складності для тестових даних, а якість класифікації оцінюється за допомогою метрики загальної точності. Наведений фрагмент ілюструє практичну реалізацію методики, описаної в даному розділі, та слугує прикладом застосування алгоритмів машинного навчання у задачі прогнозування складності текстів.

```
X = df[FEATURE_COLS]
y = df[TARGET_COL]

X_train, X_test, y_train, y_test = train_test_split(
    *arrays: X,
    y,
    test_size=0.2,
    random_state=42,
    stratify=y
)

model = RandomForestClassifier(
    n_estimators=300,
    max_depth=None,
    random_state=42,
    n_jobs=-1
)

model.fit(X_train, y_train)

y_pred = model.predict(X_test)
acc = accuracy_score(y_test, y_pred)
```

Уся вибірка обсягом 1176 зразків була поділена на навчальну та тестову підвибірки у співвідношенні 80/20. До навчальної вибірки було віднесено 940 текстів, які використовувалися для побудови моделі машинного навчання. Тестова вибірка складала 236 текстів і застосовувалася для об'єктивної оцінки якості роботи моделі (таблиця 2.1). Кожен елемент вибірки являє собою пару «вектор ознак – цільовий клас», де цільовим класом є рівень CEFR, що відповідає конкретному тексту. Такий підхід до формування вибірок дозволяє уникнути витоку інформації між етапами навчання та тестування моделі. Розподіл даних у співвідношенні 80/20 є загальноприйнятою практикою в задачах машинного навчання та забезпечує

збалансоване поєднання достатнього обсягу даних для навчання і репрезентативної оцінки узагальнювальної здатності моделі.

Таблиця 2.1

Розподіл вибірки на навчальну та тестову

Тип вибірки	К-сть зразків	Частка від загальної вибірки, %
Загальна вибірка	1176	100
Навчальна вибірка	940	80
Тестова вибірка	236	20

Процес навчання полягав у підборі оптимальних параметрів ансамблю дерев рішень та мінімізації помилки класифікації. Під час навчання модель автоматично виявляла найбільш інформативні ознаки, які мають найбільший вплив на результат прогнозування. У результаті модель навчилася встановлювати зв'язок між числовими характеристиками тексту та відповідним рівнем складності за шкалою CEFR [23].

У таблиці 2.2 наведено основні гіперпараметри навченої моделі. Кількість дерев встановлено на рівні 300, що дозволяє досягти стабільних результатів без суттєвого зростання обчислювальної складності. Максимальна глибина дерев не обмежувалась з метою надання моделі можливості повноцінно вивчати складні нелінійні залежності між ознаками. Для забезпечення відтворюваності результатів використано фіксоване значення `random_state`. Параметр `n_jobs = -1` дозволяє задіяти всі доступні обчислювальні ресурси процесора, що істотно пришвидшує процес навчання. Як критерій якості розбиття використано індекс Джині, який є стандартним для задач багатокласової класифікації.

Обрані значення гіперпараметрів відповідають компромісу між здатністю моделі до узагальнення та її обчислювальною ефективністю. Використання бутстреп-вибірок сприяє зменшенню перенавчання шляхом навчання окремих дерев на різних підмножинах даних, що є однією з ключових переваг ансамблевих методів. У поєднанні з великою кількістю дерев та відсутністю обмежень на глибину це дозволяє моделі Random Forest ефективно враховувати складні взаємодії між лінгвістичними ознаками

різних типів. Таким чином, обрана конфігурація гіперпараметрів є обґрунтованою для задачі класифікації текстів за рівнями складності CEFR.

Таблиця 2.2

Гіперпараметри моделі Random Forest

Гіперпараметр	Значення	Пояснення
n_estimators	300	Кількість дерев у лісі
max_depth	None	Максимальна глибина дерев
random_state	42	Фіксація випадковості для відтворюваності
n_jobs	-1	Використання всіх доступних ядер процесора
criterion	gini	Критерій якості розбиття
bootstrap	True	Використання бутстреп-вибірок

Після завершення навчання модель була збережена у вигляді серіалізованого файлу для подальшого використання в програмній системі. Це дозволяє не виконувати повторне навчання при кожному запуску застосунку, а лише завантажувати готову модель у пам'ять та використовувати її для прогнозування рівня складності нових текстів у реальному часі.

2.4 Інтерпретація результатів моделі

Після завершення етапу навчання моделі машинного навчання необхідним є проведення її ґрунтовного аналізу з метою оцінювання ефективності, стабільності та практичної придатності отриманих результатів. Навіть за високих показників точності важливо з'ясувати, які саме рівні класифікуються найкраще, де виникають помилки та між якими класами спостерігається найбільша плутанина. Такий аналіз дозволяє не лише оцінити загальну якість моделі, а й визначити її сильні й слабкі сторони.

Оцінювання якості побудованої моделі здійснювалося на тестовій вибірці, яка складала 20% від загального обсягу корпусу. Для аналізу були використані стандартні метрики якості багатокласової класифікації: точність (precision), повноту (recall), F1-міру та загальну точність (accuracy), а також матрицю помилок. Такий підхід забезпечує комплексний аналіз результатів і дозволяє оцінити не лише загальний рівень роботи моделі, а й якість розпізнавання кожної категорії складності [24].

На рис. 2.3 наведено результати оцінювання моделі за основними метриками для кожного рівня складності. Варто додати, що класи від 1 до 6 відповідають рівням A1 - C2 відповідно.

Для рівня A1 модель демонструє високі значення precision та recall ($\approx 0,95$), що свідчить про дуже точне та стабільне розпізнавання простих початкових текстів. Аналогічно високі результати спостерігаються і для рівнів B1, B2 та C2, де F1-міра перевищує 0,87, що вказує на добру узгодженість між повнотою і точністю.

Найменші значення точності зафіксовано для рівнів A2 та C1, де спостерігається більша кількість помилок при класифікації. Це пояснюється тим, що тексти цих рівнів мають проміжну складність і частіше мають спільні лінгвістичні характеристики з сусідніми рівнями A1/B1 та B2/C2 відповідно.

Загальна точність моделі становить 0,898, тобто близько 89,8 % усіх текстів були класифіковані правильно. Значення macro average та weighted average також наближаються до 0,89–0,90, що підтверджує стабільність роботи моделі для всіх класів.

	precision	recall	f1-score	support
1	0.9459459459459459	0.9459459459459459	0.9459459459459459	37.0
2	0.8076923076923077	0.8076923076923077	0.8076923076923077	26.0
3	0.9259259259259259	0.9433962264150944	0.9345794392523364	53.0
4	0.9285714285714286	0.975	0.951219512195122	40.0
5	0.8648648648648649	0.8	0.8311688311688312	40.0
6	0.875	0.875	0.875	40.0
accuracy	0.8983050847457628	0.8983050847457628	0.8983050847457628	0.89830508
macro avg	0.8913334121667456	0.8911724133422246	0.8909343393757573	236.0
weighted avg	0.8975064653030755	0.8983050847457628	0.8975773051480169	236.0

Рис. 2.3 Результати classification report

На рис. 2.4 представлено матрицю помилок для шести рівнів складності. Головна діагональ матриці містить значення правильно класифікованих текстів для кожного рівня.

Для рівня A1 правильно визначено 35 текстів, тоді як лише 2 тексти були помилково віднесені до A2. Для рівня A2 спостерігається незначне зміщення у бік рівня B1, що свідчить про близькість складності цих рівнів. Найбільша кількість правильних передбачень зафіксована для рівня B1 — 50 зразків, що підтверджує стабільність розпізнавання середнього рівня складності.

Для рівнів C1 та C2 також спостерігається часткова взаємна плутанина по 5 випадків, що є типовим явищем для високих рівнів, де тексти можуть поєднувати складні синтаксичні та семантичні конструкції з різним ступенем абстракції.

Водночас слід зазначити, що грубі помилки класифікації практично відсутні, тобто тексти простого рівня A1 помилково не належать до рівня C2, і навпаки. Це свідчить про коректну роботу моделі та її здатність адекватно відокремлювати принципово різні за складністю мовні рівні.

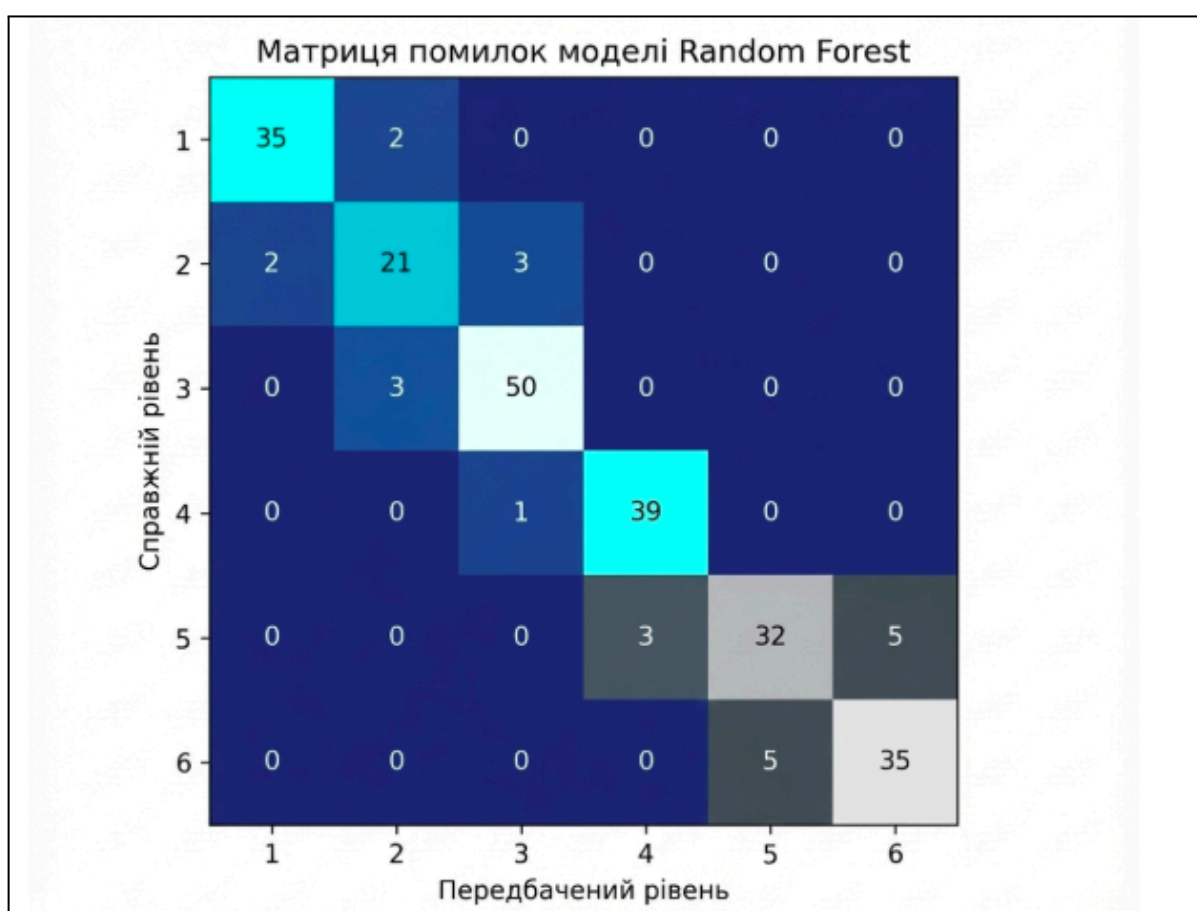


Рис. 2.4 Матриця помилок моделі Random Forest

Отримані результати свідчать про достатньо високу якість побудованої моделі Random Forest для автоматичного визначення рівня складності англomовних текстів. Значення загальної точності близьке до 90 %, що є хорошим показником для задач багатокласової класифікації у сфері обробки природної мови [24]. Аналіз матриці помилок показує, що модель не робить грубі помилки між далекими рівнями, наприклад, A1 і C2, а більшість неточностей пов'язані з близькими рівнями складності, що є очікуваним результатом.

На рис. 2.5 представлено теплову карту кореляції між усіма лінгвістичними метриками, які використовувалися для навчання моделі машинного навчання. Аналіз кореляційних зв'язків дозволяє зробити висновок про наявність чіткої залежності між рівнем складності тексту (level_id) та показниками читабельності, середньою довжиною слів, середньою довжиною речень, а також семантичними характеристиками.

Найсильніша позитивна кореляція з рівнем складності спостерігається для показників Flesch–Kincaid Grade Level, Gunning Fog Index, SMOG та Dale–Chall, що підтверджує їхню здатність адекватно відображати ускладнення текстів при переході від базових рівнів до просунутих. Також суттєвими є кореляції між середньою кількістю складів у слові, часткою рідкісних слів за шкалою Zipf та глибиною синтаксичних залежностей.

Окрему увагу привертає наявність сильних кореляцій між деякими метриками всередині однієї групи, зокрема між показниками читабельності між собою, а також між семантичними характеристиками. Це свідчить про часткову мультиколінеарність ознак, однак обраний алгоритм Random Forest є стійким до таких залежностей, що забезпечує коректну роботу моделі без необхідності жорсткого відбору ознак.

Кореляційний аналіз дозволяє оцінити доцільність одночасного використання різних груп метрик у межах єдиного вектора ознак. Наявність як слабо корельованих, так і сильно пов'язаних показників свідчить про те, що модель отримує інформацію з різних мовних рівнів — лексичного, синтаксичного, морфологічного та семантичного. Це забезпечує комплексне охоплення властивостей тексту та зменшує ризик переорієнтації моделі на окремі вузькі характеристики. Таким чином, сформований набір ознак є збалансованим і придатним для подальшого використання у задачі класифікації рівнів складності текстів.

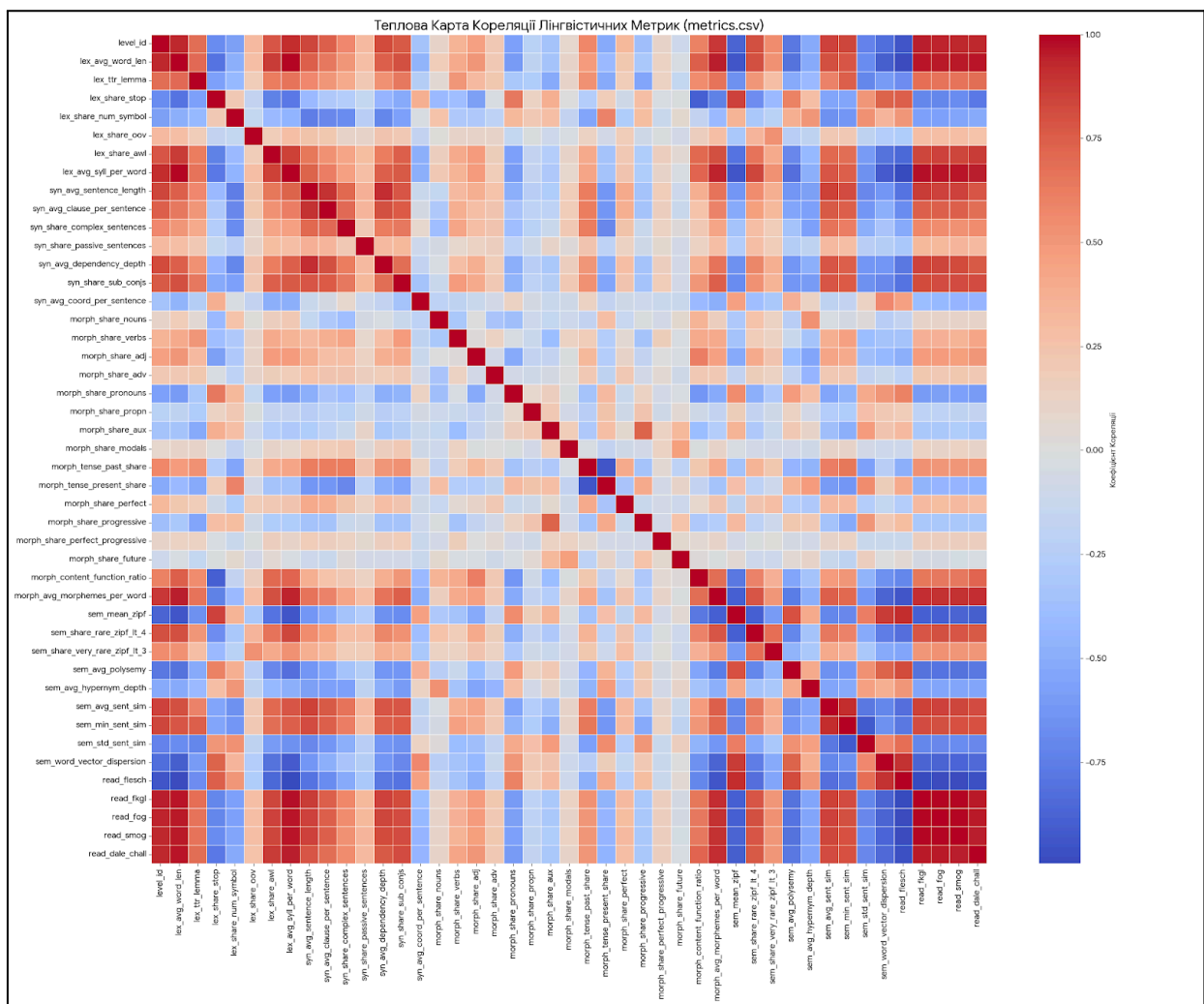


Рис. 2.5 Теплова карта кореляції між лінгвістичними метриками

Для додаткової інтерпретації роботи моделі Random Forest було виконано аналіз важливості лінгвістичних ознак. Даний показник відображає внесок кожної метрики у процес прийняття рішення класифікатором та дозволяє визначити, які характеристики тексту є найбільш інформативними під час прогнозування рівня складності [25].

На рис. 2.6 показано, що найбільший внесок у прийняття рішень має показник Flesch–Kincaid Grade Level (read_fkgi), що свідчить про визначальну роль показників читабельності при оцінюванні складності текстів. Високу важливість також демонструють індекси Flesch Reading Ease (read_flesch) та Gunning Fog Index (read_fog), які комплексно враховують довжину речень і структуру слів.

Серед лексичних характеристик суттєвий вплив мають частка слів з Academic Word List (lex_share_awl) та середня довжина слова (lex_avg_word_len), що підтверджує залежність складності тексту від рівня

формальності та лексичної насиченості. До важливих семантичних показників належать середня частотність слів у шкалі Zipf (`sem_mean_zipf`) та частка рідкісних слів (`sem_share_rare_zipf_lt_4`), що відображає абстрактність та спеціалізованість лексики.

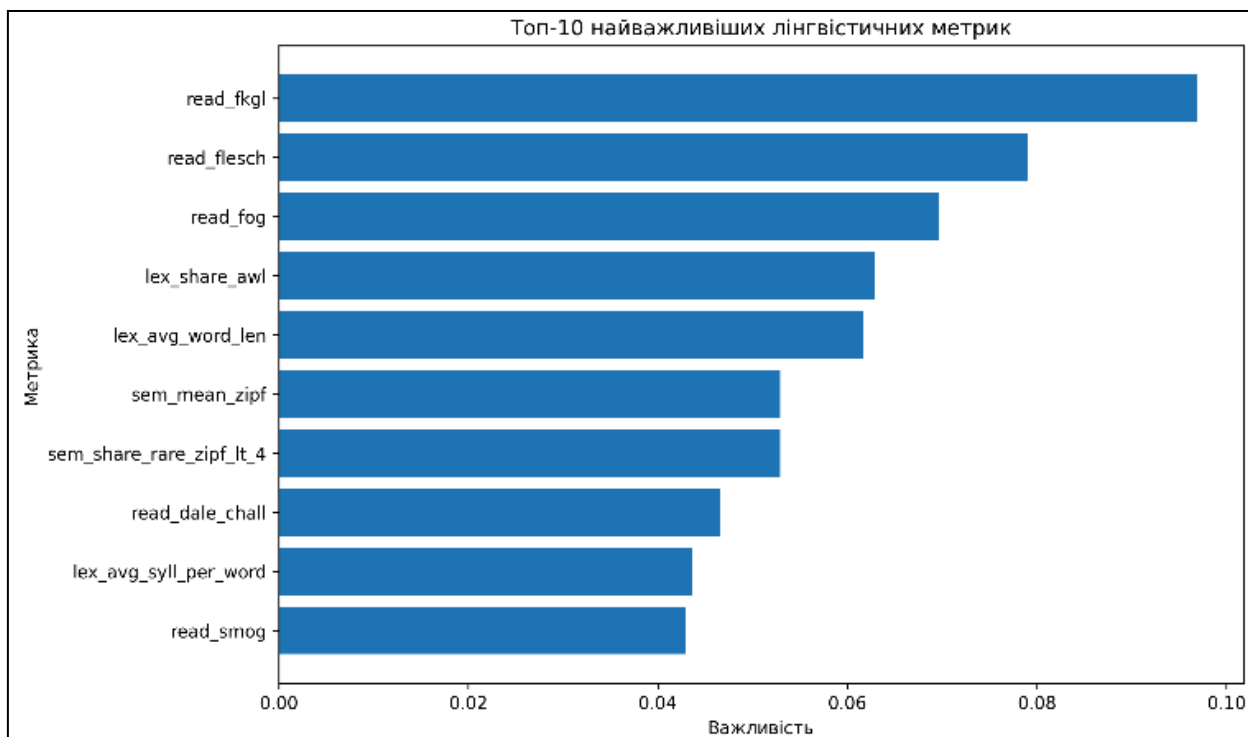


Рис. 2.6 Топ-10 домінантних лінгвістичних метрик за версією моделі

Отримані результати підтверджують, що рівень складності тексту визначається насамперед поєднанням показників читабельності, лексичної насиченості та семантичної рідкісності, що відповідає сучасним уявленням про багатофакторну природу текстової складності.

Висновки до розділу 2

У другому розділі було розроблено методику автоматичного оцінювання складності англomовних текстів відповідно до шкали CEFR. На основі сформованого корпусу текстів різних рівнів складності здійснено повний цикл попередньої обробки даних, що включав нормалізацію, токенізацію, лематизацію та автоматичне визначення мови тексту. Це дозволило забезпечити однорідність і якість вхідних даних для подальшого аналізу.

Було реалізовано систему розрахунку лексичних, синтаксичних, морфологічних, семантичних та показників читабельності, які комплексно

описують структурні та змістові характеристики тексту. Сформований вектор ознак став основою для навчання моделі машинного навчання та дозволив врахувати багатofакторну природу текстової складності.

Для класифікації рівня складності текстів було обрано алгоритм Random Forest, який продемонстрував високу узагальнювальну здатність та стійкість до шуму у даних. Навчання моделі на підготовленій вибірці дозволило досягти високого рівня точності класифікації, що підтверджує доцільність обраного підходу.

У межах інтерпретації результатів було виконано аналіз матриці помилок, звіту класифікації та важливості ознак, що дозволило визначити найбільш інформативні метрики для прогнозування рівня складності. Отримані результати засвідчили, що ключову роль відіграють показники читабельності, лексичної насиченості та семантичної рідкості слів.

Отже, у розділі було створено науково обґрунтовану методику оцінювання складності текстів, яка поєднує сучасні методи обробки природної мови та машинного навчання і може бути ефективно використана як основа для практичної програмної реалізації, описаної в наступному розділі.

РОЗДІЛ 3. РОБОТА ПРОГРАМНОЇ СИСТЕМИ

3.1 Загальна організація та архітектурні рішення програмної системи

Розроблена програмна система призначена для автоматизованого визначення рівня складності англомовних текстів відповідно до шкали CEFR на основі лінгвістичних та статистичних характеристик тексту. Основною метою створення системи є надання користувачеві зручного інструменту для аналізу текстів у режимі реального часу з можливістю отримання не лише прогнозованого рівня, але й набору пояснювальних метрик. Система реалізована як клієнт-серверний вебзастосунок, що забезпечує гнучкість, масштабованість та незалежність користувацького інтерфейсу від серверної частини.

Архітектура програмної системи побудована за модульним принципом і включає такі основні компоненти: користувацький вебінтерфейс, серверну частину з реалізованим API, модуль обробки природної мови та машинного навчання, а також базу даних для збереження результатів аналізу та статистичної інформації. Взаємодія між компонентами здійснюється через HTTP-запити у форматі REST, що забезпечує чіткий поділ відповідальностей між рівнями системи та спрощує її подальше розширення [26].

На рисунку 3.1 представлено загальну архітектуру програмної системи автоматичного визначення рівня складності англомовних текстів за шкалою CEFR.

Користувач взаємодіє із системою через вебінтерфейс, який забезпечує введення тексту або завантаження файлу для подальшої обробки. Вебінтерфейс передає введені дані до серверної частини системи, реалізованої на основі фреймворку FastAPI. Представлена архітектура забезпечує чіткий розподіл функцій між компонентами системи та зручність подальшої інтеграції в освітні інформаційні середовища.

На етапі лінгвістичної обробки текст передається до модуля NLP, який виконує попередню обробку, нормалізацію, токенизацію, лематизацію та обчислення лінгвістичних метрик. Отримані числові ознаки надходять до модуля машинного навчання, реалізованого на основі алгоритму Random Forest [23], який здійснює класифікацію тексту за рівнями складності CEFR.

Результати прогнозування, а також статистичні дані про виконані аналізи зберігаються у реляційній базі даних PostgreSQL. База даних використовується як для накопичення результатів роботи системи, так і для

формування статистичної інформації, що надалі відображається користувачу через вебінтерфейс.

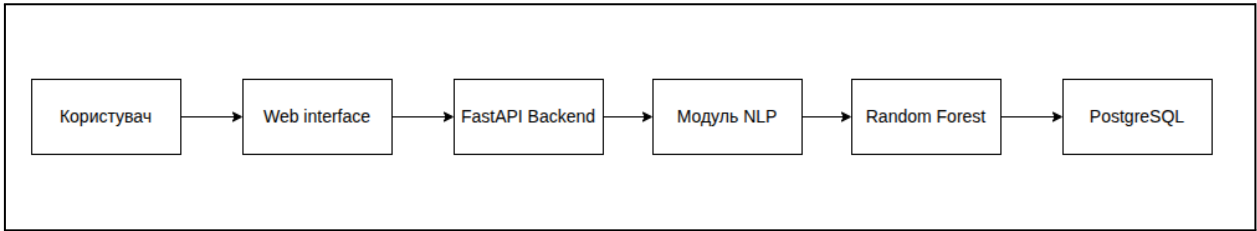


Рис. 3.1 Компонентна архітектура програмної системи

Для більш детального аналізу процесів обробки даних у системі доцільно розглянути діаграму потоків даних, яка відображає послідовність обробки тексту, формування метрик, прогнозування рівня CEFR та збереження результатів.

На рисунку 3.2 представлено діаграму потоків даних програмної системи визначення рівня складності англomовних текстів за шкалою CEFR. Діаграма ілюструє основні процеси обробки даних, а також взаємодію між користувачем, функціональними модулями системи та базами даних [6].

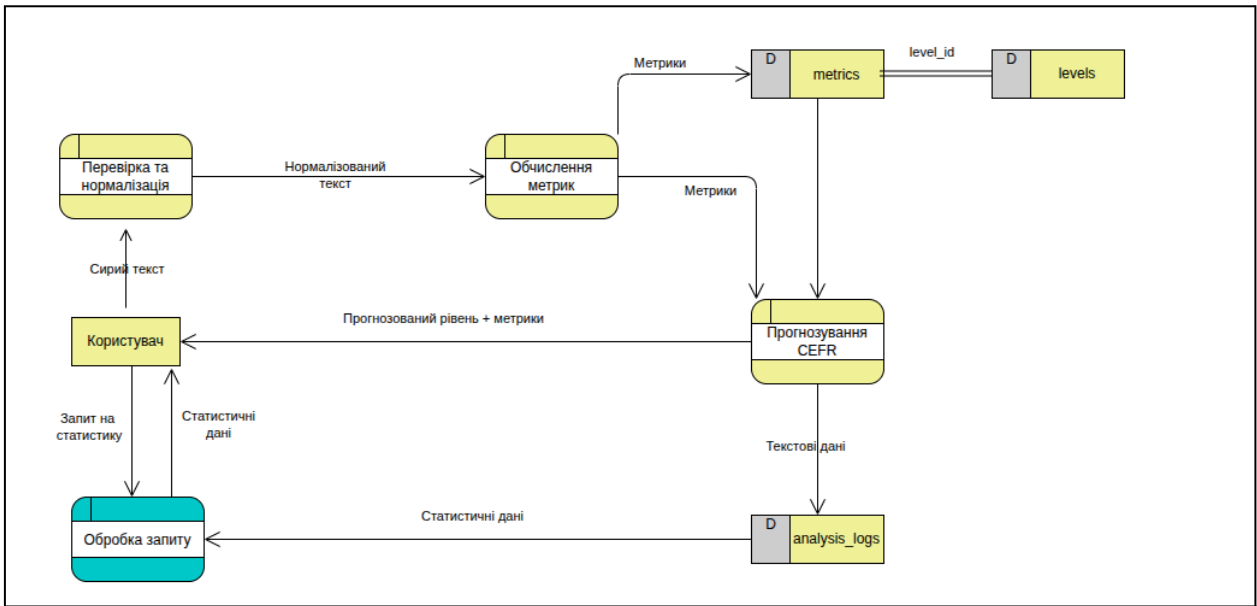


Рис 3.2 Діаграма потоків даних

Початковим джерелом даних є користувач, який вводить або завантажує сирий текст для аналізу. Введений текст надходить до процесу перевірки та нормалізації, де виконується автоматичне визначення мови, очищення тексту, приведення до нижнього регістру, видалення зайвих символів та підготовка

до подальшої лінгвістичної обробки. Результатом цього етапу є нормалізований текст.

Далі нормалізований текст передається до процесу обчислення лінгвістичних метрик, у якому формуються числові ознаки, що характеризують згадані лексичні характеристики. Надалі значення метрик для текстів користувачів безпосередньо у таблиці `metrics` не будуть зберігатися, оскільки ця таблиця використовується виключно для збереження корпусних даних, сформованих на етапі навчання моделі [20].

Після формування вектора ознак метрики надходять до процесу прогнозування рівня CEFR, де застосовується навчена модель машинного навчання на основі алгоритму `Random Forest`. Модель визначає прогнозований рівень складності тексту та формує результат у вигляді прогнозованого рівня разом із розрахованими метриками.

Отримані результати аналізу та текстові дані зберігаються у базі даних `analysis_logs`, яка використовується для накопичення історії аналізів. Користувач має можливість надсилати запит на перегляд статистичних даних. Ці запити обробляються процесом обробки запиту, який звертається до таблиці `analysis_logs` та повертає відповідні статистичні дані користувачу.

Представлена DFD-діаграма наочно демонструє повний життєвий цикл обробки текстових даних у системі [6] – від моменту введення тексту користувачем до отримання прогнозу рівня складності та збереження результатів для подальшого аналізу.

Для зберігання результатів аналізу текстів, обчислених лінгвістичних метрик та довідкової інформації в програмній системі використовується реляційна база даних `PostgreSQL`. Структура бази даних спроектована з урахуванням вимог до цілісності даних, швидкодії доступу та можливості подальшого розширення функціональності системи [27]. Логічну модель бази даних програмної системи подано на рис. 3.3.

Основною таблицею є таблиця `metrics`, у якій зберігаються всі числові ознаки, що використовуються для прогнозування рівня складності тексту. Кожен запис цієї таблиці відповідає одному проаналізованому тексту та включає набір лексичних, синтаксичних, морфологічних, семантичних і показників читабельності. Усі ці показники формують вектор ознак, що подається на вхід моделі машинного навчання.

Таблиця `levels` є довідниковою та використовується для зберігання інформації про рівні володіння мовою відповідно до шкали CEFR. Вона

Для фіксації історії взаємодії користувачів із системою використовується таблиця `analysis_logs`. У ній зберігаються дата та час виконання аналізу, ідентифікатор рівня, довжина тексту, тип джерела введення та інші службові характеристики. Наявність цієї таблиці дозволяє здійснювати статистичний аналіз використання системи, відстежувати динаміку прогнозів та формувати узагальнені звіти.



На рисунку 3.4 представлено початкову сторінку вебінтерфейсу системи прогнозування складності англомовних текстів за шкалою CEFR. Дана сторінка виконує функцію першого рівня взаємодії користувача із системою та слугує для ознайомлення з її призначенням і можливостями. У центральній частині сторінки розміщено інформаційний блок з привітальним повідомленням та коротким описом системи. У нижній частині інформаційного блоку розміщено кнопку навігації «Перейти до аналізу тексту», яка ініціює перехід користувача до основного функціонального модуля введення та аналізу тексту.

Візуальне оформлення початкової сторінки побудоване з урахуванням принципів простоти та інтуїтивної зрозумілості, що знижує поріг входу для користувачів з різним рівнем цифрової підготовки. Фонові графічні елементи у вигляді позначень рівнів CEFR створюють асоціативний зв'язок з основною функцією системи та підкреслюють її освітню спрямованість.

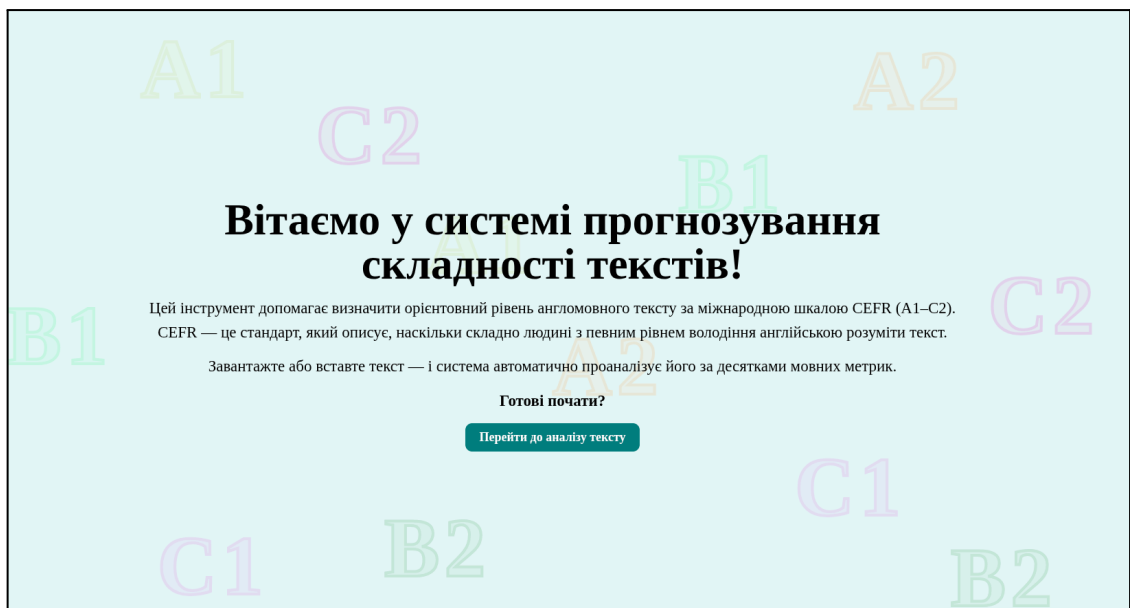


Рис. 3.4 Початкова вітальна сторінка

На рисунку 3.5 представлено головну сторінку вебінтерфейсу. У лівій частині сторінки розміщено поле введення тексту, в яке користувач може безпосередньо ввести фрагмент англомовного тексту або скористатися функцією завантаження текстового файлу. Поруч доступна кнопка «Аналізувати», після натискання якої ініціюється процес обробки тексту та визначення його рівня складності. Права частина інтерфейсу призначена для відображення результатів аналізу та має інформативний характер: до моменту

запуску аналізу користувачеві відображається повідомлення про необхідність введення тексту. Така структура сторінки забезпечує чітке розмежування між зоною введення даних і зоною результатів, що підвищує зручність взаємодії з системою та наочність процесу аналізу.

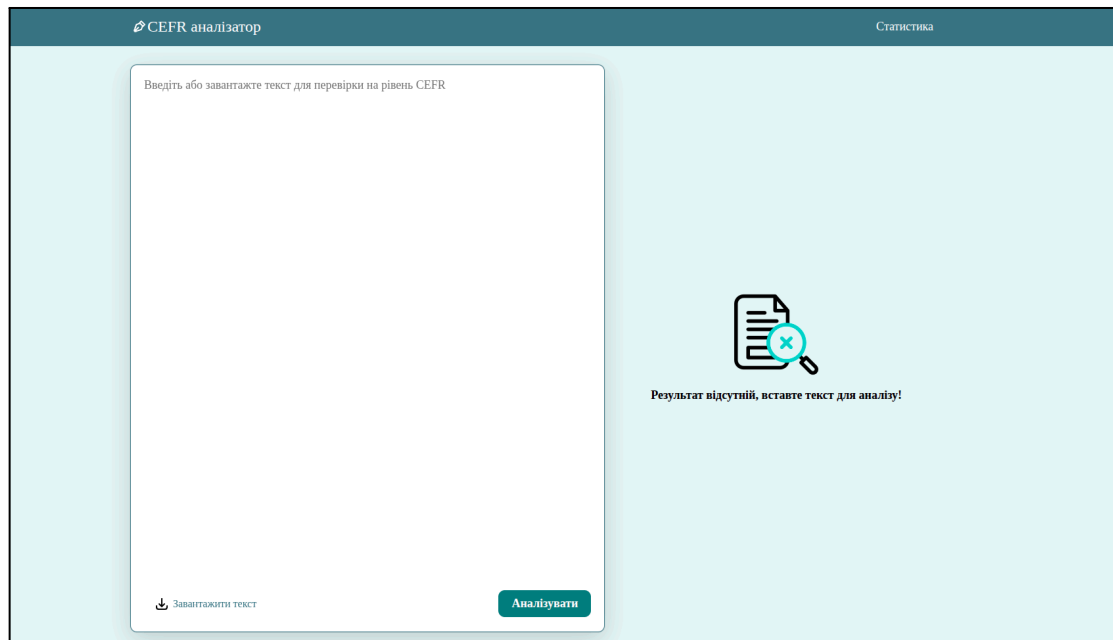


Рис. 3.5 Вигляд основної сторінки

Після обробки введеного тексту система відображає прогнозований рівень складності у вигляді великого маркованого результату, що наочно демонструє користувачу загальну оцінку складності поданого матеріалу. Результат подається у форматі рівня шкали CEFR з його словесною інтерпретацією, що дозволяє швидко співвіднести текст із відповідним рівнем мовної підготовки. Нижче основного результату CEFR відображається набір інформаційних карток, які містять ключові лінгвістичні та статистичні показники тексту, зокрема характеристики лексичної насиченості, синтаксичної структури, морфологічного складу та показники читабельності. Кожна картка супроводжується коротким поясненням відповідного показника, що сприяє кращому розумінню результатів аналізу та підвищує інтерпретованість роботи системи для кінцевого користувача (рис. 3.6).

Відображення окремих лінгвістичних характеристик у наочному вигляді дозволяє користувачеві не лише отримати підсумковий рівень CEFR, а й проаналізувати причини, які вплинули на прийняте рішення моделі. Це розширює можливості системи для використання в освітніх і дослідницьких

цілях, зокрема для добору навчальних матеріалів, самостійного контролю рівня мовної підготовки та аналізу складності текстів різного призначення.

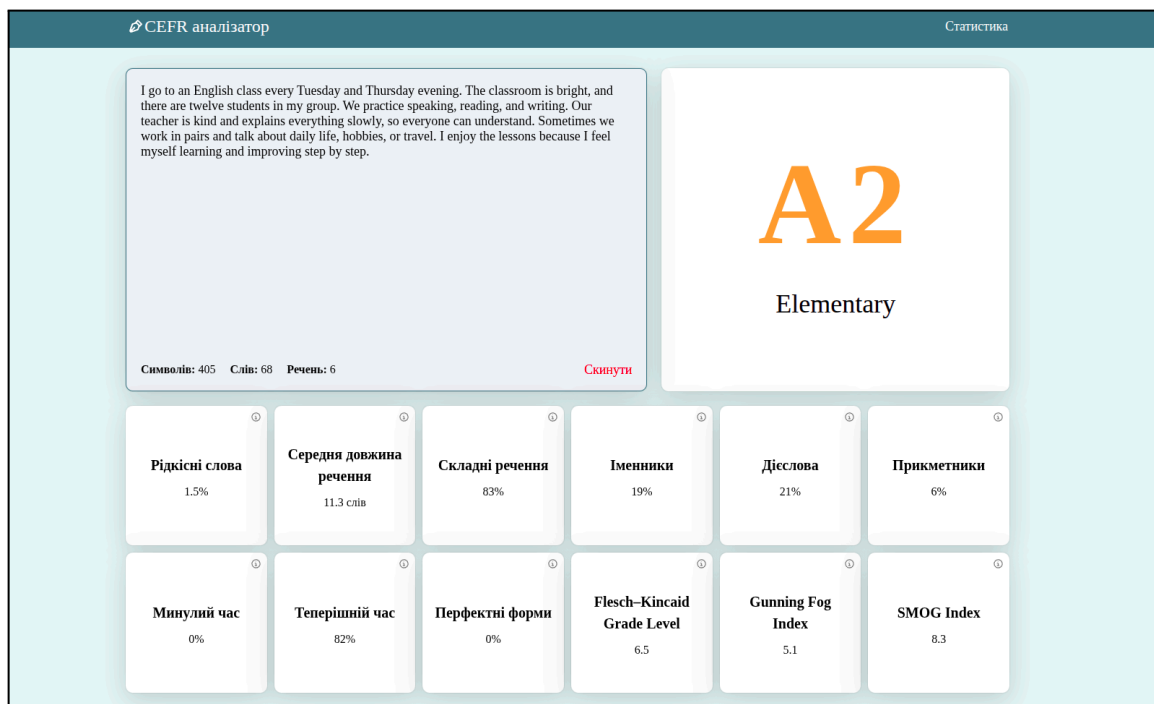


Рис. 3.6 Сторінка з прогнозованим рівнем тексту та характеристиками

На рисунку 3.7 зображено сторінку статистики програмної системи, яка містить узагальнену інформацію про всі виконані аналізи за період використання системи. На сторінці відображається загальна кількість проведених аналізів, середній розмір проаналізованих текстів, а також розподіл текстів за рівнями складності відповідно до шкали CEFR. Подана візуалізація дозволяє оцінити характер використання системи, визначити найбільш поширені рівні складності текстів та проаналізувати загальні тенденції у виборі навчальних матеріалів. Отримані статистичні дані можуть бути використані для подальших досліджень динаміки мовної складності освітнього контенту, а також для вдосконалення методики добору навчальних текстів.

Подання статистичної інформації у вигляді агрегованих показників та візуальних елементів дозволяє швидко оцінити характер використання системи й загальний розподіл складності проаналізованих текстів. Відображення кількості аналізів, середнього розміру тексту та частки кожного рівня CEFR створює цілісне уявлення про типові сценарії застосування системи. Такий функціонал є корисним як для дослідницьких задач, пов'язаних із вивченням динаміки мовної складності текстів у часі.

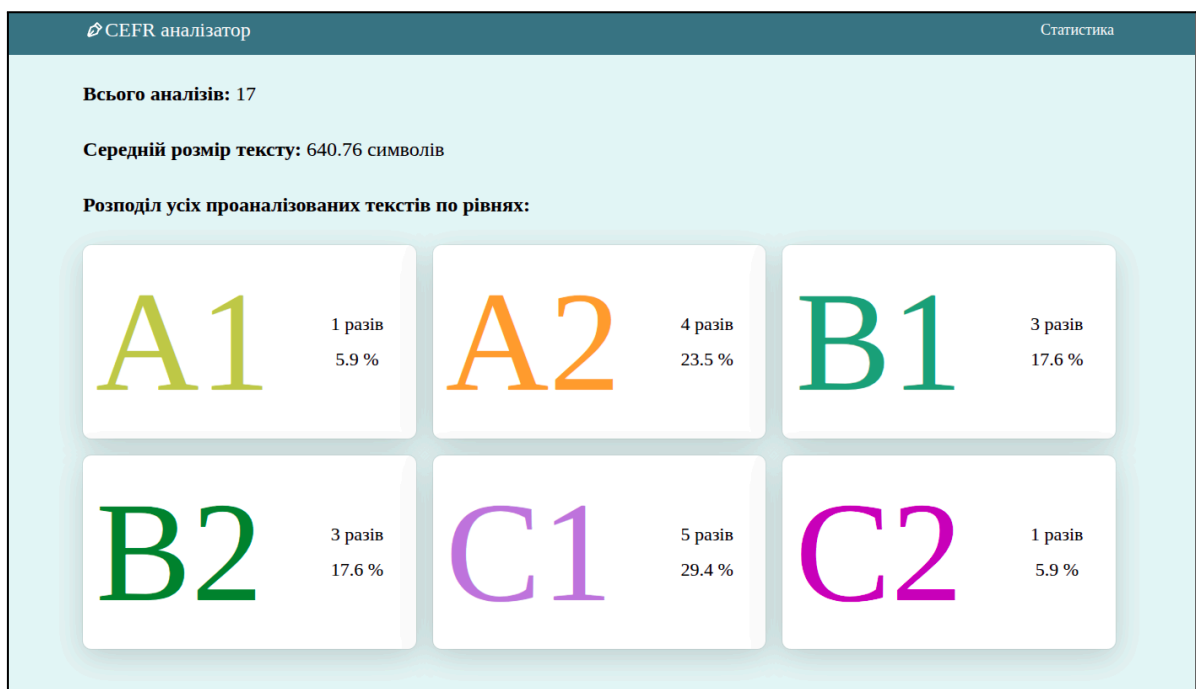


Рис. 3.7 Сторінка статистики

Таким чином, розроблена архітектура програмної системи забезпечує чіткий розподіл функцій між її основними компонентами: вебінтерфейсом, серверною частиною на базі FastAPI, модулем обробки природної мови, моделлю машинного навчання та базою даних. Побудована структура сприяє модульності, масштабованості та зручності супроводу системи, а також дозволяє легко розширювати її функціональні можливості в подальшому [28]. Інтеграція користувацького інтерфейсу з аналітичним ядром через REST API забезпечує ефективну обробку запитів у реальному часі та створює передумови для використання системи в прикладних освітніх та наукових середовищах [26].

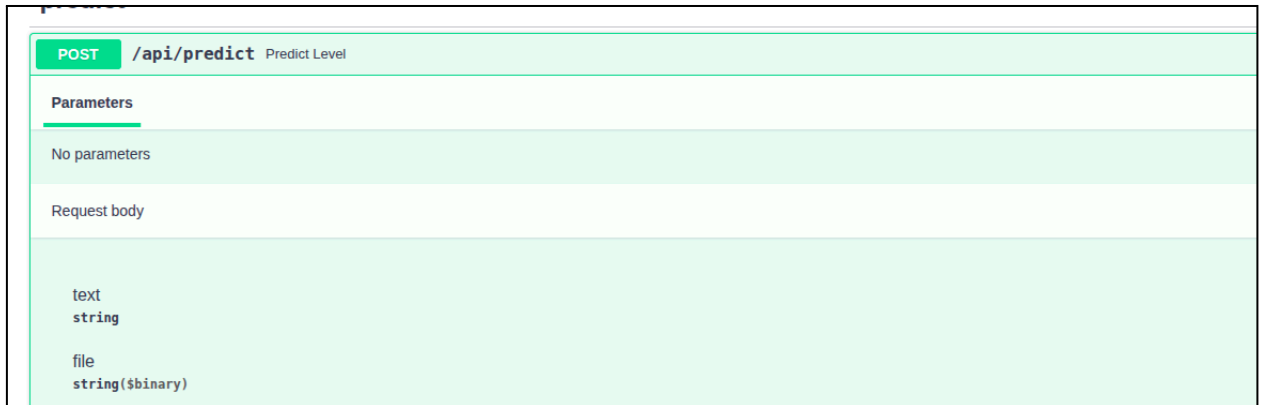
3.2 Визначення рівня CEFR

Розроблена система реалізована у вигляді вебсервісу на базі фреймворку FastAPI, який надає програмний інтерфейс для надсилання запитів на аналіз текстів та отримання результатів класифікації. Для зручності тестування та демонстрації роботи API використовується вбудований інтерфейс документації Swagger, який автоматично генерується під час запуску сервера та дозволяє виконувати запити до системи у графічному режимі без використання сторонніх інструментів.

Інтерфейс Swagger забезпечує наочне представлення доступних методів API, їхніх параметрів, форматів запитів та відповідей [29]. У розробленій

системі реалізовано два основні ендпоінти: для визначення рівня складності тексту та для отримання статистичних даних.

Для виконання прогнозування використовується ендпоінт POST /api/predict, який дозволяє передавати вхідні дані у двох форматах: безпосередньо у вигляді текстового рядка або у вигляді текстового файлу з розширенням .txt (рис. 3.7)



POST /api/predict Predict Level

Parameters

No parameters

Request body

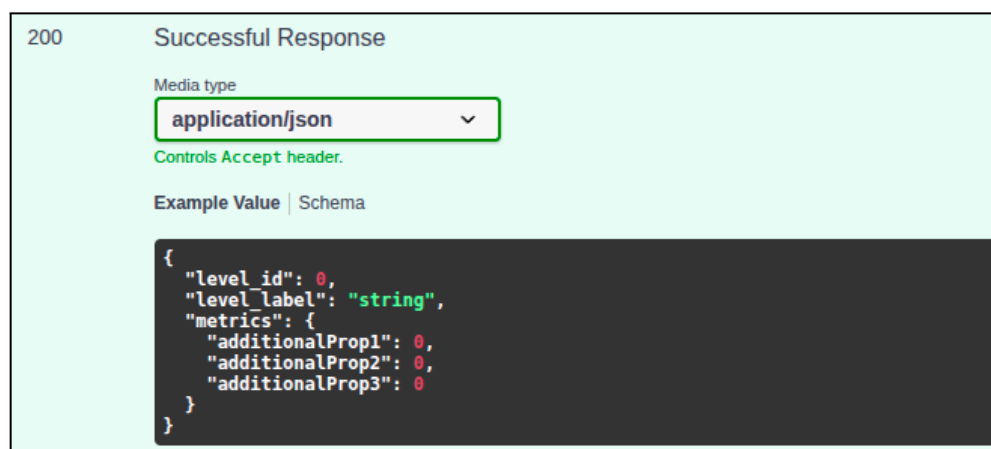
text
string

file
string(\$binary)

Рис. 3.7 Вигляд POST запиту для визначення рівня CEFR

Після відправлення запиту система виконує повний цикл обробки: перевірку мови тексту, нормалізацію, обчислення лінгвістичних метрик, формування вектора ознак та застосування навченої моделі Random Forest для прогнозування рівня CEFR [24]. У відповідь користувач отримує структурований JSON-об'єкт, який містить:

- числовий ідентифікатор (level_id);
- текстове позначення рівня CEFR (level_label);
- набір обчислених лінгвістичних метрик (metrics) (рис. 3.8).



200 Successful Response

Media type

application/json

Controls Accept header.

Example Value | Schema

```
{
  "level_id": 0,
  "level_label": "string",
  "metrics": {
    "additionalProp1": 0,
    "additionalProp2": 0,
    "additionalProp3": 0
  }
}
```

Рис. 3.8 Відповідь POST запиту

Окрім функції прогнозування, у системі реалізовано можливість отримання статистичних даних про використання сервісу за допомогою ендпоінту GET /api/stats. Даний ендпоінт не потребує жодних вхідних параметрів, що робить його універсальним інструментом для перегляду поточного стану бази аналізів. У цьому методі реалізовано автоматичне формування статистичних показників, які відображають загальну активність користувачів, динаміку виконаних запитів та особливості розподілу текстів за рівнями CEFR. Через інтерфейс Swagger користувач може побачити структуру відповіді, допустимі формати даних та приклади виклику, що полегшує ознайомлення з API та прискорює процес інтеграції.

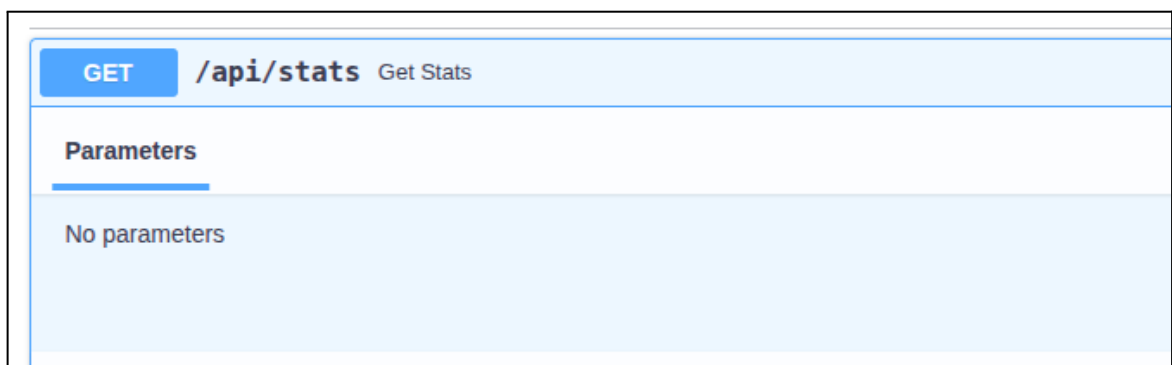


Рис. 3.9 GET запит для статистичних даних

На рисунку 3.10 зображено структуру відповіді сервера на GET-запит /api/stats, яка формується на основі даних, накопичених у таблиці analysis_logs. Результат повертається у форматі JSON і містить кілька груп аналітичних показників: загальну кількість виконаних користувачами аналізів (total_count), середню довжину текстів у символах (avg_text_length), а також детальний розподіл проаналізованих текстів за рівнями CEFR. Кожен елемент масиву levels включає ідентифікатор рівня (level_id), його текстове позначення (level_label), абсолютну кількість текстів цього рівня (count) і відносну частку (share). Такий формат відповіді дає змогу швидко оцінити, які рівні найчастіше визначаються моделлю, виявити можливі дисбаланси у вхідних даних або проаналізувати поведінку користувачів. Статистичні дані можуть бути використані для моніторингу якості роботи системи, налаштування навчального корпусу, а також для прийняття рішень щодо подальшого розширення функціоналу вебзастосунку.

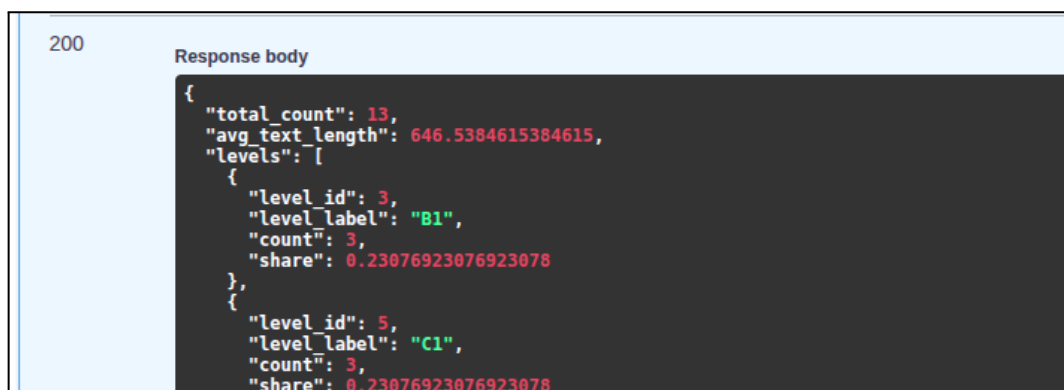


Рис. 3.10 Статистичні дані

Побудована програмна система не лише автоматично визначає рівень складності англomовного тексту за шкалою CEFR, але й дозволяє здійснювати глибший аналіз отриманих результатів з погляду якісних характеристик тексту. Оскільки шкала CEFR є ієрархічною, перехід між рівнями відображає поступове зростання мовної складності [4], що проявляється у зміні лексичних, граматичних та семантичних показників.

Особливо показовим є перехід між рівнями B1 та B2, який у методиках викладання іноземних мов вважається переходом від середнього до впевнено просунутого рівня володіння мовою. На рівні B1 користувач здатний розуміти основний зміст повідомлень, повсякденні тексти та прості аргументації, тоді як рівень B2 характеризується здатністю сприймати складніші тексти, абстрактні теми, аргументовані міркування та стилістично різноманітні мовні конструкції [2].

Для аналізу було обрано два невеликі тексти рівня B1 та B2:

Текст рівня B1: *“Last year, I decided to change my daily routine because I felt tired and unmotivated. I used to wake up late, skip breakfast, and rush to work without any plan. This made my days stressful and unorganised. One day, I started waking up earlier and preparing my tasks for the day in advance. I also began walking to work instead of taking the bus. After a few weeks, I noticed that I had more energy and felt calmer during the day. Even though the change was not easy at first, it helped me feel more positive and in control of my life.”*

Текст рівня B2: *“Many people believe that technology makes our lives easier, yet it also brings challenges that we are only beginning to understand. For example, the constant use of smartphones has changed the way we communicate with one another. While we can now stay in touch with friends and family across*

the world, many people find it difficult to focus for long periods or have meaningful face-to-face conversations. Researchers warn that this constant distraction may affect our ability to think deeply and solve complex problems. However, they also suggest that with conscious effort—such as setting limits on screen time or creating “phone-free” spaces—we can develop healthier digital habits. The key is to use technology as a helpful tool rather than letting it control our daily routines.”

На рис. 3.11 наведено порівняльні результати лінгвістичного аналізу двох текстів, що відповідають рівням B1 та B2 за шкалою CEFR. Для кожного рівня подано значення ключових лексичних, синтаксичних, морфологічних, семантичних показників та індексів читабельності. Таке порівняння дозволяє наочно простежити, які саме мовні характеристики змінюються під час переходу від середнього до просунутого рівня володіння мовою, а також оцінити, наскільки автоматично обчислені метрики узгоджуються з теоретичними уявленнями про складність текстів різних рівнів.

<pre> { "level_id": 3, "level_label": "B1", "metrics": { "lex_avg_word_len": 4.118, "lex_ttr_lemma": 0.676, "lex_share_stop": 0.559, "lex_share_num_symbol": 0.114, "lex_share_oov": 0, "lex_shareawl": 0.029, "lex_avg_syll_per_word": 1.392, "syn_avg_sentence_length": 14.571, "syn_avg_clause_per_sentence": 1.571, "syn_share_complex_sentences": 0.857, "syn_share_passive_sentences": 0, "syn_avg_dependency_depth": 6, "syn_share_sub_conjs": 0.029, "syn_avg_coord_per_sentence": 2, "morph_share_nouns": 0.167, "morph_share_verbs": 0.196, "morph_share_adj": 0.108, "morph_share_adv": 0.069, "morph_share_pronouns": 0.137, "morph_share_propn": 0, "morph_share_aux": 0.01, "morph_share_modals": 0, "morph_tense_past_share": 0.524, "morph_tense_present_share": 0.19, "morph_share_perfect": 0, "morph_share_progressive": 0, "morph_share_perfect_progressive": 0, "morph_share_future": 0, "morph_content_function_ratio": 1.196, "morph_avg_morphemes_per_word": 1.451, "sem_mean_zipf": 6.043, "sem_share_rare_zipf_lt_4": 0.039, "sem_share_very_rare_zipf_lt_3": 0.02, "sem_avg_polysemy": 8.236, "sem_avg_hypernym_depth": 5.091, "sem_avg_sent_sim": 0.809, "sem_min_sent_sim": 0.715, "sem_std_sent_sim": 0.047, "sem_word_vector_dispersion": 0.512, "read_flesch": 75.098, "read_fkgl": 6.405, "read_fog": 8.574, "read_smog": 8.842, "read_dale_chall": 7.455 } } </pre>	<pre> { "level_id": 4, "level_label": "B2", "metrics": { "lex_avg_word_len": 4.984, "lex_ttr_lemma": 0.744, "lex_share_stop": 0.481, "lex_share_num_symbol": 0.123, "lex_share_oov": 0, "lex_shareawl": 0.085, "lex_avg_syll_per_word": 1.636, "syn_avg_sentence_length": 21.5, "syn_avg_clause_per_sentence": 2.167, "syn_share_complex_sentences": 1, "syn_share_passive_sentences": 0, "syn_avg_dependency_depth": 7.333, "syn_share_sub_conjs": 0.039, "syn_avg_coord_per_sentence": 1.667, "morph_share_nouns": 0.248, "morph_share_verbs": 0.171, "morph_share_adj": 0.116, "morph_share_adv": 0.062, "morph_share_pronouns": 0.085, "morph_share_propn": 0, "morph_share_aux": 0.047, "morph_share_modals": 0.107, "morph_tense_past_share": 0.036, "morph_tense_present_share": 0.5, "morph_share_perfect": 0.036, "morph_share_progressive": 0.071, "morph_share_perfect_progressive": 0, "morph_share_future": 0, "morph_content_function_ratio": 1.51, "morph_avg_morphemes_per_word": 1.682, "sem_mean_zipf": 5.852, "sem_share_rare_zipf_lt_4": 0.016, "sem_share_very_rare_zipf_lt_3": 0, "sem_avg_polysemy": 6.792, "sem_avg_hypernym_depth": 5.584, "sem_avg_sent_sim": 0.876, "sem_min_sent_sim": 0.853, "sem_std_sent_sim": 0.022, "sem_word_vector_dispersion": 0.426, "read_flesch": 43.949, "read_fkgl": 12.264, "read_fog": 14.073, "read_smog": 14.314, "read_dale_chall": 9.755 } } </pre>
---	--

Рис. 3.11 Визначення рівня B1 та B2

Наведені значення метрик добре ілюструють якісну різницю між текстами рівнів B1 та B2. Передусім, на лексичному рівні спостерігається зростання середньої довжини слова: для тексту рівня B1 середня довжина становить близько 4,1 символа, тоді як для B2 — майже 5 символів. Аналогічну тенденцію демонструє й середня кількість складів у слові (приблизно 1,39 проти 1,64), що вказує на використання більш багатоскладової та, відповідно, складнішої лексики на рівні B2. Типово для просунутого рівня також збільшується лексичне різноманіття: коефіцієнт *type-token ratio* для лем становить близько 0,68 для B1 та 0,74 для B2. Це означає, що в тексті B2 більша частка унікальних слів і менше повторів, що підвищує як інформаційну насиченість, так і вимоги до словникового запасу читача. Важливим є й зростання частки слів з *Academic Word List*: показник у тексті B2 майже утричі вищий, ніж у B1, тобто текст на рівні B2 значно частіше містить академічну й абстрактну лексику.

Синтаксичні показники також демонструють послідовне ускладнення структури висловлення. У тексті рівня B1 середня довжина речення становить близько 14,5 слів, тоді як у тексті B2 — понад 21 слово. Кількість фінітних предикатів на речення зростає орієнтовно з 1,6 до 2,2, а частка складних речень для B2 досягає максимального значення, усі речення розглянутого фрагмента класифіковано як складні. Вищою є й частка пасивних конструкцій, а також середня глибина дерева залежностей, що свідчить про використання багаторівневих синтаксичних структур із вкладеними підрядними частинами. Натомість кількість координативних зв'язків на речення дещо менша, тобто текст B2 менше спирається на просте поєднання однорідних частин і більше — на підрядність та ієрархічну організацію висловлення.

Морфологічні метрики відображають зміщення балансу між частинами мови та часовими формами. У тексті рівня B2 зростає частка іменників (з $\approx 0,17$ до $\approx 0,25$), тоді як частка займенників помітно зменшується. Це означає, що текст B2 більшою мірою оперує конкретними й абстрактними поняттями, а не займенниковими відсиланнями, що робить його семантично щільнішим. Показник співвідношення повнозначних і службових слів також вищий у B2, тобто на одиницю тексту припадає більше “змістовних” слів, які несуть основне смислове навантаження. Цікавим є й розподіл часових форм: у тексті B1 домінує минулий час (понад половина предикатів), тоді як у B2 переважає теперішній час. Такий перехід від ретроспективного наративу до

узагальнених тверджень і описів типовіший для текстів, орієнтованих на аргументацію, пояснення та опис явищ, що є характерною рисою рівня B2. Крім того, середня кількість морфем на слово зростає приблизно з 1,45 до 1,68, що свідчить про активніше використання афіксації та більш складних словотворчих моделей.

Семантичні показники підкреслюють, що лексика тексту рівня B2 в середньому рідкісніша та семантично “глибша”. Хоча середнє значення частотності за шкалою Zipf (`sem_mean_zipf`) для B2 дещо нижче, ніж для B1, це означає, що в просунутому тексті частіше з’являються менш уживані слова. При цьому середня глибина гіперонімічної ієрархії (`sem_avg_hyponym_depth`) вища саме у B2, тобто використовувані лексеми належать до більш спеціалізованих, “глибоко” розташованих у семантичній ієрархії понять. Показники внутрішньотекстової когерентності (`sem_avg_sent_sim` та `sem_min_sent_sim`) також вищі для тексту B2 за одночасно нижчої дисперсії (`sem_std_sent_sim`), що інтерпретується як більш послідовний розвиток теми без різких семантичних стрибків між сусідніми реченнями. Менше значення `sem_word_vector_dispersion` у B2 підтверджує, що семантичні вектори слів більш згуртовані навколо спільного тематичного центру.

Найбільш наочно різницю між рівнями демонструють індекси читабельності. Для тексту рівня B1 значення індексу Флеша становить близько 75, що відповідає доволі легкому для читання тексту, тоді як для B2 цей показник знижується до приблизно 43, що характеризує текст як значно складніший. Відповідно, індекс Flesch–Kincaid зростає з приблизно 6,5 до понад 12, а індекси Gunning Fog, SMOG та Dale–Chall також демонструють істотне підвищення. Усі ці метрики узгоджено свідчать про те, що текст рівня B2 вимагає від читача вищого рівня мовної компетентності: довші речення, багатший словник, більше багатоскладових і термінологічних слів, складніші граматичні конструкції.

У сукупності наведені відмінності підтверджують, що система коректно відображає концептуальний перехід від B1 до B2. Текст рівня B1 характеризується простішою синтаксичною організацією, вищою часткою займенників та дещо розмовнішою лексикою, тоді як текст B2 є лексично насиченим, структурно складним і семантично щільним. Таким чином, результати автоматичного аналізу добре відповідають описам рівнів CEFR і

демонструють, внаслідок яких саме лінгвістичних параметрів досягається зростання складності навчальних текстів.

3.3 Інтеграція моделі в освітні процеси

Сучасний освітній простір характеризується активним впровадженням цифрових технологій, спрямованих на підвищення ефективності навчального процесу, індивідуалізацію навчання та об'єктивізацію оцінювання результатів [30]. У цьому контексті автоматичне визначення рівня складності текстів відповідно до шкали CEFR набуває особливої актуальності, оскільки дозволяє адаптувати навчальні матеріали до реального рівня мовної підготовки здобувачів освіти. Розроблена програмна система може бути інтегрована в освітні процеси як допоміжний аналітичний інструмент для викладачів, студентів, розробників навчального контенту та освітніх онлайн-платформ.

Одним із ключових напрямів застосування моделі є підтримка процесу формування навчальних матеріалів. Під час створення підручників, методичних посібників, тестових завдань і додаткових навчальних ресурсів викладач може скористатися системою для автоматичної перевірки відповідності текстів заявленому рівню CEFR [31]. Це дозволяє уникнути ситуацій, коли навчальний матеріал є надмірно складним або, навпаки, занадто спрощеним для певної категорії студентів. У результаті забезпечується точніше узгодження змісту навчальних матеріалів із навчальними програмами, стандартами освіти та очікуваними результатами навчання [17].

Важливим практичним напрямом є використання розробленої моделі в системах дистанційного та змішаного навчання. Освітні платформи можуть інтегрувати API для автоматичного аналізу текстових матеріалів, що завантажуються користувачами. Зокрема, у системах управління навчанням модель може застосовуватися для перевірки рівня складності текстів, які студенти подають у межах індивідуальних проєктів, есе або наукових рефератів. Це надає викладачеві додаткову аналітичну інформацію щодо відповідності складності виконаних робіт заявленому рівню навчання.

Окрему цінність розроблена система має для організації самостійної роботи студентів. Інтерфейс вебзастосунку дозволяє користувачеві завантажити власний текст або вставити фрагмент для аналізу та отримати не лише прогнозований рівень CEFR, але й детальні значення лінгвістичних

метрик. Завдяки цьому студент може контролювати складність текстів, які обирає для читання, а також аналізувати власні письмові роботи з погляду відповідності бажаному рівню володіння мовою. Такий підхід сприяє розвитку рефлексії, самоконтролю та усвідомленого формування мовної компетентності.

Не менш важливим є використання системи у процесах моніторингу та аналітики освітнього контенту. Накопичення статистичних даних у базі дозволяє здійснювати аналіз тенденцій використання текстів різного рівня складності, частоти звернень користувачів, середнього обсягу проаналізованих матеріалів тощо. Отримані статистичні показники можуть застосовуватися адміністрацією навчального закладу або розробниками освітніх платформ для оцінки ефективності навчальних матеріалів, коригування навчальних програм та оптимізації освітнього процесу.

Практичне значення має також можливість використання моделі у процесі підготовки до міжнародних мовних іспитів, таких як IELTS, TOEFL, Cambridge Exams та інші. Автоматичне визначення рівня складності текстів дозволяє відбирати навчальні матеріали відповідно до вимог конкретного іспиту та поточного рівня підготовки студента. Це підвищує ефективність підготовки та сприяє більш раціональному розподілу навчального навантаження [31].

Таким чином, інтеграція запропонованої моделі в освітні процеси забезпечує комплексну підтримку навчання англійської мови на різних рівнях. Система може використовуватися як інструмент контролю якості навчальних матеріалів, засіб індивідуалізації навчання, модуль аналітики в освітніх платформах та допоміжний сервіс для самостійної підготовки студентів. Завдяки відкритій архітектурі, використанню сучасних вебтехнологій та можливості масштабування розроблена програмна система має значний потенціал для подальшого впровадження в освітнє середовище.

Висновки до розділу 3

У третьому розділі роботи було детально розглянуто практичну реалізацію програмної системи автоматичного визначення рівня складності англійських текстів за шкалою CEFR та особливості її функціонування. Описано загальну архітектуру системи, принципи взаємодії між основними компонентами, а також засоби реалізації серверної та клієнтської частин.

Проаналізовано архітектурні рішення програмної системи, обґрунтовано вибір клієнт-серверної моделі та використання REST API для організації обміну даними між вебінтерфейс і серверною частиною. Побудована компонентна архітектура, DFD-діаграма та структура бази даних продемонстрували цілісність і логічну організацію системи, а також забезпечили чіткий розподіл відповідальностей між модулями обробки тексту, машинного навчання та збереження результатів. Додатково показано реалізацію вебінтерфейс, який забезпечує зручну взаємодію користувача із системою та візуалізацію результатів аналізу.

Продемонстровано роботу Swagger-інтерфейсу, описано структуру запитів і відповідей, а також наведено приклад практичного використання системи для аналізу текстів різних рівнів складності. Проведене порівняння текстів рівнів B1 та B2 підтвердило коректність автоматично обчислених лінгвістичних метрик та їхню відповідність теоретичним уявленням про зростання мовної складності.

Обґрунтовано можливості інтеграції розробленої моделі в освітні процеси. Показано, що система може бути використана для підбору навчальних матеріалів, аналізу письмових робіт студентів, підтримки дистанційного навчання, формування статистики використання навчального контенту та підготовки до міжнародних мовних іспитів. Практична спрямованість системи дозволяє застосовувати її як допоміжний інструмент у роботі викладачів і студентів, а також у складі освітніх інформаційних платформ.

Отже, у третьому розділі продемонстровано повноцінну реалізацію програмної системи, отримані результати підтверджують ефективність обраних архітектурних та програмних рішень і створюють основу для подальшого розвитку та масштабування системи.

ВИСНОВКИ

У межах даної магістерської кваліфікаційної роботи було досліджено проблему автоматичного визначення рівня складності англомовних текстів відповідно до міжнародної шкали CEFR та розроблено програмну систему, яка реалізує дану задачу із застосуванням сучасних методів NLP та ML.

Здійснено ґрунтовний аналіз теоретичних основ оцінювання складності текстів, розкрито сутність шкали CEFR як міжнародного стандарту опису мовної компетентності, а також охарактеризовано основні групи лінгвістичних показників, що впливають на сприйняття текстів різних рівнів складності. Розглянуто сучасні підходи NLP та методи машинного навчання, які використовуються для класифікації текстів, а також здійснено огляд існуючих систем автоматичного визначення рівня складності.

Розроблено методику автоматичного оцінювання складності англомовних текстів. Сформовано корпус текстів рівнів A1–C2, виконано їх попередню обробку, реалізовано систему розрахунку метрик. На основі отриманих метрик побудовано модель машинного навчання на базі алгоритму Random Forest. Проведене навчання та тестування моделі продемонструвало високий рівень точності класифікації близько 90%. Додатково виконано інтерпретацію результатів.

Реалізовано повнофункціональну клієнт-серверну програмну систему прогнозування складності англомовних текстів. Побудовано модульну архітектуру вебзастосунку з використанням фреймворку FastAPI, модуля NLP, моделі машинного навчання та реляційної бази даних PostgreSQL. Реалізовано інтерфейс користувача для взаємодії з системою, систему обміну даними через REST API, а також механізми збереження результатів аналізу та формування статистичних даних.

Подальший розвиток розробленої програмної системи може бути спрямований на підвищення точності визначення рівня складності текстів шляхом використання сучасних нейромережевих моделей глибокого навчання. Окремим напрямом розвитку є інтеграція системи з освітніми платформами та системами дистанційного навчання, аби використовувати систему як компонент адаптивного освітнього середовища.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. М'яновська А. Ю., Савіцький Р. С. Інтеграція бібліотек NLP та попередньо навчених моделей для автоматичної оцінки складності тексту. Вісник Херсонського національного технічного університету. 2025. № 2. С. 319–326. DOI: 10.35546/kntu2078-4481.2025.2.2.39.
2. Рада Європи. Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume. Strasbourg : Council of Europe Publishing, 2020. 278 p. URL: <https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16802fc1bf>.
3. Fulcher G. Assessing second language speaking. Language Teaching. 2015. Vol. 48. № 2. P. 198–216. DOI: 10.1017/S0261444814000391.
4. North B., Piccardo E. Developing illustrative descriptors of aspects of mediation for the CEFR. Strasbourg : Council of Europe, 2017. 65 p.
5. Deutsch T., Jasbi M., Shieber S. Linguistic features for readability assessment. Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. Stroudsburg: Association for Computational Linguistics, 2020. P. 1–17. DOI: 10.48550/arXiv.2006.00377.
6. Jurafsky D., Martin J. H. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, NJ: Prentice Hall, 2008. 1044 p. URL: [https://pages.ucsd.edu/~bakovic/compphon/Jurafsky,%20Martin.-Speech%20and%20Language%20Processing_%20An%20Introduction%20to%20Natural%20Language%20Processing%20\(2007\).pdf](https://pages.ucsd.edu/~bakovic/compphon/Jurafsky,%20Martin.-Speech%20and%20Language%20Processing_%20An%20Introduction%20to%20Natural%20Language%20Processing%20(2007).pdf).
7. Parks L., Peters W. Natural language processing in mixed-methods text analysis: A workflow approach. International Journal of Social Research Methodology. 2022. Vol. 26. № 1. P. 1–13. DOI: 10.1080/13645579.2021.2018905.
8. Roldan-Baluis W. L., Zapata N. A., Mañaccasa Vásquez M. S. The effect of natural language processing on the analysis of unstructured text: A systematic review. International Journal of Advanced Computer Science and Applications. 2022. Vol. 13. № 5. P. 43–51. DOI: 10.14569/IJACSA.2022.0130507.

9. Indurkha N., Damerau F. J. Handbook of natural language processing. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC, 2010. 679 p. URL: <https://karczmarczyk.users.greyc.fr/TEACH/TAL/Doc/Handbook%20Of%20Natural%20Language%20Processing.%20Second%20Edition%20Chapman%20&%20Hall%20Crc%20Machine%20Learning%20&%20Pattern%20Recognition%202010.pdf>.
10. Partalidou E., Spyromitros-Xioufis E., Doropoulos S., Vologiannidis S., Diamantaras K. Design and implementation of an open source Greek POS Tagger and Entity Recognizer using spaCy. 2019. DOI: 10.48550/arXiv.1912.10162.
11. Aluisio S., Specia L., Gasperin C., Scarton C. Readability assessment for text simplification. Proceedings of the Fifth Workshop on Innovative Use of NLP for Building Educational Applications (NAACL HLT 2010). Stroudsburg: Association for Computational Linguistics, 2010. P. 1–9.
12. Paaß G., Giesselbach S. Foundation models for natural language processing: Pre-trained language models integrating media. Cham, Switzerland: Springer, 2023. DOI: 10.48550/arXiv.2302.08575.
13. Shah K., Patel H., Sanghvi D. A comparative analysis of logistic regression, random forest and KNN models for text classification. Augmented Human Research. 2020. Vol. 5. Art. 12. DOI: 10.1007/s41133-020-00032-0.
14. Garrido-Merchán E. C., Gozalo-Brizuela R., Gonzalez-Carvajal S. Comparing BERT against traditional machine learning models in text classification. Journal of Computational and Cognitive Engineering. 2023. Vol. 2. № 4. P. 352–356. DOI: 10.47852/bonviewJCCE3202838.
15. Huawei S., Aryadoust V. A systematic review of automated writing evaluation systems. Education and Information Technologies. 2023. Vol. 28. P. 771–795. DOI: 10.1007/s10639-022-11200-7.
16. Кочубейник О. Критичний дискурс-аналіз: як забезпечувати репрезентативність корпусу текстів? Psychology of Personality. 2020. № 10. С. 140–148. DOI: 10.15330/ps.10.1.140-148.
17. Мартиненко В. Лінгвістична складність навчальних текстів у підручниках читання для молодших школярів. Acta Paedagogica Volynienses. 2024. № 34. DOI: 10.32782/apv/2024.2.5.
18. Бистріцький А., Толкунов І., Gavrilenko О., Богданова Н. Аналіз сучасних методів реферування текстів. Адаптивні системи

- автоматичного управління. 2025. № 1. С. 46–54. DOI: 10.20535/1560-8956.46.2025.323679.
19. Pawłowski A., Walkowiak T. NLP for digital humanities: Processing chronological text corpora. *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*. Stroudsburg: Association for Computational Linguistics, 2024. P. 105–112. DOI: 10.18653/v1/2024.nlp4dh-1.10.
20. Петрина В., Дорошенко А. Ефективність застосування методів класифікації для задач інтелектуального аналізу великих даних. *Scientific Bulletin of UNFU*. 2024. Vol. 34. С. 119–128. DOI: 10.36930/40340516.
21. Palma D., Soto C. Combining large language models with linguistic features for the readability complexity assessment of texts. 2025. DOI: 10.54941/ahfe1007028.
22. Salman A., Al-Jawher W. Performance comparison of support vector machines, AdaBoost, and random forest for sentiment text analysis and classification. *Journal Port Science Research*. 2024. Vol. 7. P. 300–311. DOI: 10.36371/port.2024.3.8.
23. Islam Z., Liu J., Li J., Liu L., Kang W. A semantics aware random forest for text classification. *Proceedings of the conference on knowledge discovery and data mining*. 2019. P. 1061–1070. DOI: 10.1145/3357384.3357891.
24. Salles T., Gonçalves M., Rodrigues V., Rocha L. Improving random forests by neighborhood projection for effective text classification. *Information Systems*. 2018. Vol. 77. DOI: 10.1016/j.is.2018.05.006.
25. Song S., He R., Shi Z., Zhang W. Variable importance measure system based on advanced random forest. *Computer Modelling in Engineering & Sciences*. 2021. Vol. 128. P. 65–85. DOI: 10.32604/cmescs.2021.015378.
26. Чаус О., Крук О. Мобільний застосунок для групового вивчення іноземних мов з автоматичним оцінюванням вимови. *Наука і техніка сьогодні*. 2025. № 6 (47). С. 1664–1681. DOI: 10.52058/2786-6025-2025-6(47)-1664-1681.
27. Волкова А., Харахаш О., Сакалюк О., Козуб О. Розробка електронного дайджесту наукових доробків. *Automation of Technological and Business Processes*. 2020. Vol. 12. С. 12–17. DOI: 10.15673/atbp.v12i1.1698.
28. Mohebian M., Momeni K., Pirhayati M. A comparative analysis of sustainability education in undergraduate architecture programs: Case

- studies of domestic and international architecture schools. 2025. Vol. 13. P. 69–86. DOI: 10.30473/ee.2025.56173.2285.
29. Kim J.-L., Shin M. Assessment of swagger perception by linear shape variation on the static phenomenon. Journal of Engineering and Applied Sciences. 2017. Vol. 12. P. 2702–2705. DOI: 10.3923/jeasci.2017.2702.2705.
30. Аршава О., Дейнега О., Панченко А., Жовтоніжко І. Цифрові технології в освіті: виклики, можливості, досвід впровадження. Наукові інновації та передові технології. 2025. № 7 (47). С. 1505–1517. DOI: 10.52058/2786-5274-2025-7(47)-1505-1517.
31. Ковальчук А. Розвиток цифрової компетентності майбутніх педагогів професійного навчання в умовах цифровізації. Молодь і ринок. 2024. С. 148–153. DOI: 10.24919/2308-4634.2023.296992.