Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО–МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра мультимедійних систем факультету інформатики

СТВОРЕННЯ ГІБРИДНОЇ NLP-СИСТЕМИ МОРФЕМНОГО АНАЛІЗУ УКРАЇНСЬКОМОВНИХ СЛІВ

Текстова частина до кваліфікаційної роботи
за спеціальністю 122 «Комп’ютерні науки»

Керівник кваліфікаційної роботи
ст.викл. Смиш О.Р.

_____

(підпис)

«____» _____ 2026 р.

Виконав студент МП КН–2
Кирилін Є.С.
«____» _____ 2026 р.

Київ – 2026

Міністерство освіти і науки України

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»

Кафедра мультимедійних систем факультету інформатики

ЗАТВЕРДЖУЮ

Зав.кафедри мультимедійних систем,

Доцент., к. ф.-м. н. О.П. Жежерун

_____(підпис)

«____»_____2025 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

на кваліфікаційну роботу

студенту 2-го року МП «Комп'ютерні науки» факультету інформатики

Кириліну Єгору Сергійовичу

ТЕМА: Створення гібридної NLP-системи морфемного аналізу українськомовних слів

Зміст ТЧ до кваліфікаційної роботи:

Анотація

Вступ

Огляд інструментів та наявних рішень

Розроблення гібридної системи морфемного аналізу українськомовних слів

Створення вебзастосунку та апробація отриманих результатів

Висновки

Список використаних джерел

Дата видачі «___» _____ 2025 р. Керівник _____

(підпис)

Завдання отримав _____

(підпис)

Тема:

_____

Календарний план виконання роботи:

№ п/п	Назва етапу кваліфікаційної роботи	Термін виконання етапу	Примітка
1	Отримання завдання на кваліфікаційну роботу	01.09.2025	
2	Огляд технічної літератури за темою роботи	02.09.2025 – 01.03.2026	
3	Формування ідеї й архітектури проєкту	02.09.2025 – 01.10.2025	
4	Збір корпусу та підготовка набору даних	02.10.2025 – 21.12.2025	
5	Розроблення та навчання моделей	22.12.2025 – 28.02.2026	
6	Створення п’ятиетапної гібридної системи	01.03.2026 – 15.03.2026	
7	Розроблення вебзастосунку для морфемного аналізу	16.03.2026 – 31.03.2026	
8	Оцінювання якості системи та порівняння з мовними моделями	01.04.2026 – 10.04.2026	
9	Написання пояснювальної роботи	01.02.2026 – 27.04.2026	
10	Створення слайдів для доповіді та написання доповіді	14.04.2026 – 27.04.2026	
11	Аналіз отриманих результатів з керівником	11.04.2026 – 18.04.2026	
12	Корегування роботи згідно з результатами перевірки наукового керівника	19.04.2026 – 27.04.2026	
13	Попередній захист кваліфікаційної роботи	28.04.2026	
14	Остаточне оформлення пояснювальної роботи та слайдів	29.04.2026 – 24.05.2026	
15	Захист кваліфікаційної роботи	08.06.2026	

Графік узгоджено «___» _____

Студент Кирилін Єгор Сергійович

Керівник Смиш Олег Русланович

АНОТАЦІЯ

Кваліфікаційну роботу присвячено розробленню гібридної NLP-системи морфемного аналізу українськомовних слів. Досліджено специфіку морфемної будови синтетичних флективних мов та проаналізовано обмеження наявних підходів: морфосинтаксичних аналізаторів, лексикографічних баз даних і великих мовних моделей у контексті завдання морфемної сегментації.

Сформовано навчальний корпус з пар «словоформа–BIO-розмітка» на основі «Словника афіксальних морфем» та ВЕСУМ. Навчено та порівняно три архітектурні підходи: базовий класифікатор на основі випадкового лісу (44,70 % точності повного розбору слова), ансамбль спеціалізованих трансформерних моделей за частинами мови з маршрутизацією через UDPipe (69,96 %) та єдину універсальну трансформерну модель (81,04 %).

На основі порівняльного аналізу спроєктовано п’ятирівневий гібридний каскадний конвеєр, що поєднує пошук за словником, нейромережеву класифікацію, перевірку структурної цілісності, верифікацію афіксів і комбінаторний генератор. Загальна точність гібридної системи на тестовій вибірці становить 94 %, що перевищує результати великих мовних моделей. Розроблено вебзастосунок із діагностичними візуалізаціями процесу аналізу.

Ключові слова: обробка природної мови, українська мова, морфеміка, морфемна сегментація, трансформерна нейромережа, BIO-маркування, гібридна система, RoBERTa, класифікація токенів, комп’ютерні науки, комп’ютерні системи, навчання моделі, модель, нейронні мережі, великі мовні моделі.

ЗМІСТ

ПЕРЕЛІК ПРИЙНЯТИХ СКОРОЧЕНЬ

ВЕСУМ — Великий електронний словник української мови

BERT — Bidirectional Encoder Representations from Transformers

BIO — Begin-Inside-Outside

CoNLL-U — Computational Natural Language Learning

CSS — Cascading Style Sheets — Каскадні таблиці стилів

GPT — Generative Pre-training Transformer

HTML — HyperText Markup Language — Мова гіпертекстової розмітки

LLM — Large Language Models — Великі мовні моделі

ML — Machine Learning — Машинне навчання

NLP — Natural Language Processing — Обробка природної мови

POS — Part of Speech — Частина мови

RoBERTa — Robustly Optimized BERT Approach

SIGMORPHON — Special Interest Group on Computational Morphology and Phonology

UD — Universal Dependencies — Універсальні залежності

ВСТУП

Обґрунтування вибору теми кваліфікаційної роботи

Морфема — мінімальна значуща одиниця мови, що несе лексичне або граматичне значення. Знання морфемного складу слова є базовою умовою для низки завдань комп'ютерної лінгвістики: стемінгу, лематизації, машинного перекладу, автоматичного виправлення помилок і визначення семантичної спорідненості слів. Попри це, автоматичний морфемний розбір для української мови залишається нетривіальним завданням.

Наявні морфосинтаксичні аналізатори оптимально визначають частини мови та граматичні ознаки, але розглядають слово як атомарну одиницю і не здійснюють його поділу на морфеми [1]. Лексикографічні бази містять лінгвістично достовірні розбори, однак працюють за принципом прямого пошуку: якщо слова немає в реєстрі, система не повертає результату. Великі мовні моделі завдяки алгоритму підслівного токенування не мають прямого доступу до посимвольного складу слова, що робить їх архітектурно непридатними для точної морфемної сегментації [2].

У попередньому дослідженні [3] продемонстровано потенціал детермінованого комбінаторного підходу та сформовано базовий корпус афіксальних морфем. Водночас виявлено значуще обмеження: якість розбору невідомої лексики є нестабільною через брак механізму узагальнення за межами словника. Це зумовлює необхідність гібридного підходу, що поєднає переваги детермінованих правил і нейромережевої гнучкості.

Мета і завдання кваліфікаційної роботи

Метою роботи є автоматизація процесу морфемної сегментації українськомовних слів через розроблення гібридної системи, що визначає структуру слова й класифікує його морфеми.

Об'єктом дослідження є морфемна структура слів української мови в контексті автоматизованої обробки природної мови.

Предметом дослідження є методи та моделі машинного навчання для автоматичної морфемної сегментації українськомовних слів.

Для досягнення поставленої мети виконано такі завдання:

- досліджено специфіку української мови та проаналізовано обмеження наявних NLP-рішень у контексті задачі морфемного аналізу;
- сформовано навчальний корпус із застосуванням BIO-маркування та реалізовано стратегію поділу за лемами;
- навчено та порівняно три архітектурні підходи — базовий класифікатор, ансамбль спеціалізованих трансформерних моделей та єдину універсальну трансформерну нейромережу;
- спроєктовано п’ятирівневий гібридний каскадний конвеєр та оцінено його ефективність у порівнянні з великими мовними моделями;
- розроблено вебзастосунок із комплексними візуалізаціями нейромережевих передбачень.

Практичне значення отриманих результатів

Уперше розроблено метод гібридної морфемної сегментації українськомовних слів, що поєднує детерміновані лінгвістичні правила з трансформерною нейромережевою моделлю посимвольної класифікації, навченою на сформованому наборі даних з морфемною розміткою у форматі BIO на основі лексикографічних ресурсів та автоматичної генерації словоформ. На відміну від наявних підходів, що обмежені статичним словниковим пошуком або схильні до генерації хибних морфем, запропонований метод узагальнює невідому лексику та формує лінгвістичну коректність результату.

Розроблена гібридна система досягає точності повного морфемного розбору 94,0 % на тестовій вибірці зі 100 слів, що перевищує результати провідних великих мовних моделей. Вебзастосунок забезпечує інтерактивний морфемний аналіз із відображенням результату в стандартній нотації та набором комплексних візуалізацій для дослідження внутрішніх представлень нейромережі. Сформований корпус афіксальних морфем та навчена збірка

даних є самостійними лінгвістичними ресурсами, придатними для подальших досліджень у галузі комп’ютерної лінгвістики.

Структура кваліфікаційної роботи

Кваліфікаційна робота містить вступ, 3 основні розділи, висновки та використані джерела. Загальний обсяг роботи становить 60 сторінок. У роботі наведено 14 рисунків та 4 таблиці. Список використаних джерел має 24 найменування.

РОЗДІЛ 1
ОГЛЯД ІНСТРУМЕНТІВ І НАЯВНИХ РІШЕНЬ

Розділ присвячено аналізу специфіки української морфології, огляду наявних інструментів та дослідженню методів машинного навчання, застосовних для завдання автоматизованої морфемної сегментації. Окреслено базові поняття обробки природної мови та проаналізовано структурні особливості української мови, що ускладнюють алгоритмічний розбір. Оглянуто наявні лексикографічні ресурси і морфосинтаксичні аналізатори та проаналізовано детермінований словниковий підхід, розроблений у попередньому дослідженні, його лінгвістичні переваги та вразливість до проблеми невідомої лексики. Досліджено еволюцію застосування методів машинного навчання: від статистичних алгоритмів навчання без вчителя до трансформерних архітектур у завданні посимвольної класифікації з використанням BIO-маркування. Висвітлено проблему виникнення статистичних «галюцинацій» під час роботи нейромереж. Проаналізовано обмеження сучасних великих мовних моделей, що зумовлені застосуванням алгоритмів підслівного токенування. На основі проведеного аналізу обґрунтовано доцільність розроблення гібридної каскадної системи, що поєднує детерміновані правила та нейромережеві передбачення.

1.1. Обробка природної української мови та специфіка української морфології

Обробка природної мови є однією з вагомих галузей інформатики, комп’ютерної лінгвістики та штучного інтелекту, що вивчає методи аналізу та синтезу людської мови комп’ютерними системами. Ця галузь комп’ютерних наук уможливлює написання програм, що здатні взаємодіяти з людьми, розуміти глибинний контекст, генерувати та обробляти текстові масиви великих обсягів. Комп’ютерний аналіз тексту охоплює кілька рівнів мовного аналізу:

графематичний (токенізація), морфемний, морфологічний, синтаксичний, семантичний та прагматичний [4].

У цій ієрархії морфологічний та морфемний рівні є фундаментальними, оскільки без точного визначення структури слова та його граматичних ознак неможливо здійснити коректний синтаксичний та семантичний аналіз. Комп'ютерна система має виокремлювати слово не як набір символів, а як структуровану одиницю з власними характеристиками.

З лінгвістичного погляду, ці завдання вирішуються на перетині двох дисциплін: морфології та морфеміки. Морфологія — це галузь мовознавства, яка досліджує слово як граматичну одиницю мови, а саме: будову, граматичні категорії, класи слів і словозміну. Морфеміка ж досліджує безпосередньо морфемну будову слова, порушуючи питання поділу слів на найменші значущі одиниці — морфеми, їхню класифікацію та принципи виокремлення [5].

Відповідно до структурної класифікації, морфеми поділяються на кореневі та афіксальні. Корінь є обов'язковою частиною слова, оскільки саме він несе основне лексичне значення. Афіксальні морфеми уточнюють це значення або вказують на граматичні зв'язки та діляться на префікси та постфікси. Префікси розташовані перед коренем і переважно виконують словотвірну функцію. Морфеми, що стоять після кореня, постфікси, в українській мові мають дві основні категорії: суфікси, що мають словотвірне значення, та закінчення, які мають граматичне значення. Окремо виділяється службова морфема інтерфікс, яка поєднує корені у складних словах, як-от літера «о» у слові «лісостеп».

Незважаючи на стрімкий розвиток технологій обробки природної мови, автоматичний аналіз української мови на морфемному рівні залишається нетривіальним завданням через її особливості. Українська мова належить до групи синтетичних флективних мов [6]. На відміну від аналітичних мов, наприклад, англійської, де граматичні відношення передаються переважно порядком слів та прийменниками, в українській мові ці відношення виражаються всередині самого слова за допомогою системи афіксації. Ця специфіка породжує низку складнощів для комп'ютерного оброблення.

По-перше, високофлективність та парасинтетичний словотвір. В українській мові слова переважно утворюються через одночасне додавання префікса та суфікса, як-от у слові «непереможний». Крім того, до одного кореня може додаватися ланцюг із кількох суфіксів та префіксів поспіль. Це створює низку унікальних словоформ, що унеможливлює використання суто словникових методів, тобто пошуку слова у фіксованій базі даних, оскільки алгоритму постійно траплятиметься проблема невідомих слів.

По-друге, під час словозміни чи словотворення корінь слова може змінювати свій літерний склад. Наприклад, чергування голосних [е]-[і] у словах «пекти» та «піч», або чергування приголосних [к]-[ч]-[ц] у словах «рука», «ручка» та «руці». Для простих алгоритмічних стеммерів, які відсікають кінець слова, такі зміни є критичними, оскільки система ідентифікує різні форми одного слова як різні лексеми з різними коренями.

По-третє, граматичне значення в українській мові може виражатися і наявною, і нульовою морфемою. Наприклад, в іменниках чоловічого роду називного відмінка, як-от «ліс» чи «брат», закінчення є нульовим, що комп’ютерному алгоритму нетривіально ідентифікувати без додаткового морфологічного аналізу. Крім того, наявність постфіксів, як-от зворотних часток «-ся» або «-сь» в окремих дієсловах інфінітивної форми та займенниках, руйнує класичну парадигму алгоритмів машинного навчання, згідно з якою закінчення завжди є останньою морфемою у слові. В українській мові закінчення може міститися всередині слова, що вимагає створення багаторівневих каскадних правил відсікання.

Окремою проблемою є омонімія морфем, коли однакові кластери літер виконують різні функції залежно від контексту. Наприклад, буквосполука «ов» є суфіксом у прикметнику «слив-ов-ий», але належить до кореня в дієслові «повз-а-ти». Без розуміння повної структури слова комп’ютерна система може хибно сегментувати такі лексеми.

Отже, високий рівень морфологічної варіативності, багатство словозміни та складні структурні парадигми роблять українську мову нетривіальною для

машинного розбору. Традиційні детерміновані алгоритми не здатні охопити всі винятки, тоді як статистичні моделі машинного навчання можуть генерувати хибні розбори через омонімію афіксів. Цим обґрунтовано необхідність пошуку нових, гібридних підходів до обробки української природної мови, які поєднуватимуть строгі правила лінгвістики з гнучкістю нейромережевих архітектур.

1.2. Огляд наявних рішень для морфологічного та морфемного аналізу

Досліджено наявні програмні та лексикографічні рішення для вирішення лінгвістичних завдань українською мовою. Для об’єктивного оцінювання їхньої ефективності розмежовано засоби морфологічного аналізу — визначення граматичних характеристик слова як цілісної одиниці, та морфемного аналізу, тобто внутрішньослівної сегментації. Аналіз наявних рішень демонструє суттєвий дисбаланс у розвитку цих двох напрямів.

Серед найпоширеніших систем морфологічного та синтаксичного аналізу є модель UDPipe — комплексний конвеєр для обробки природних мов, що виконує токенізацію, лематизацію, розмічування частин мови та аналіз залежностей. Цю архітектуру побудовано на основі фреймворку універсальних залежностей, який є міжнародним стандартом для анотування граматичних структур, синтаксичних залежностей та морфологічних особливостей. Результати аналізування тексту моделлю повертаються у форматі CoNLL-U (Computational Natural Language Learning) [1], як показано на рисунку 1.1.

Модель UDPipe виконує ідентифікацію частин мови, визначення початкової форми слова, а також виокремлення його граматичних ознак, як-от відмінка, роду, числа, часу або особи [1]. Однак значущим обмеженням систем, як UDPipe, у контексті дослідження є розгляд слова як неподільної, атомарної одиниці тексту. Алгоритми, засновані на універсальних залежностях, не

здійснюють поділу слова на значущі структурні частини. Наприклад, класифікуючи лексему «перероблений» як дієприкметник із відповідними морфологічними ознаками, система не здатна відділити префікс «пере-», корінь «роб», суфікси «-л-» та «-ен-» та закінчення «-ий». Отже, наявні морфологічні аналізатори вирішують завдання визначення граматичних класів, але є непридатними для безпосереднього морфемного розбору слів української мови.

Id	Form	Lemma	UPosTag	XPosTag	Feats	Head	DepRel	Deps	Misc
# generator = UDPipe 2, https://lindat.mff.cuni.cz/services/udpipe									
# udpipe_model = ukrainian-iu-ud-2.17-251125									
# udpipe_model_licence = CC BY-NC-SA									
# newdoc									
# newpar									
# sent_id = 1									
# text = Морфемна сегментація — це поділ слова на його мінімальні значущі частини.									
1	Морфемна	морфемний	ADJ	Ao-fsns	Case=Nom|Gender=Fem|Number=Sing	2	amod	_	TokenRange=0:8
2	сегментація	сегментація	NOUN	Ncfsnn	Animacy=Inan|Case=Nom|Gender=Fem|Number=Sing	5	nsubj	_	TokenRange=9:20
3	—	—	PUNCT	U	PunctType=Dash	5	punct	_	TokenRange=21:22
4	це	це	PRON	Pd--nnsnn	Animacy=Inan|Case=Nom|Gender=Neut|Number=Sing|PronType=Dem	5	expl	_	TokenRange=23:25
5	поділ	поділ	NOUN	Ncmsnn	Animacy=Inan|Case=Nom|Gender=Masc|Number=Sing	0	root	_	TokenRange=26:31
6	слова	слово	NOUN	Ncnsgn	Animacy=Inan|Case=Gen|Gender=Neut|Number=Sing	5	nmod	_	TokenRange=32:37
7	на	на	ADP	Spsa	Case=Acc	11	case	_	TokenRange=38:40
8	його	його	DET	Pps3--paa	Case=Acc|Number=Plur|Person=3|Poss=Yes|PronType=Prs|Uninflect=Yes	11	det	_	TokenRange=41:45
9	мінімальні	мінімальний	ADJ	Ao--pasn	Animacy=Inan|Case=Acc|Number=Plur	11	amod	_	TokenRange=46:56
10	значущі	значущий	ADJ	Ao--pasn	Animacy=Inan|Case=Acc|Number=Plur	11	amod	_	TokenRange=57:64
11	частини	частина	NOUN	Ncfpan	Animacy=Inan|Case=Acc|Gender=Fem|Number=Plur	5	nmod	_	SpaceAfter=No|TokenRange=65:72
12	.	.	PUNCT	U	_	5	punct	_	SpaceAfter=No|TokenRange=72:73

Рисунок 1.1. Приклад результату морфологічного аналізу сервісом UDPipe

Водночас, інструментарій для дослідження внутрішньої будови слова представлений переважно лексикографічними ресурсами — оцифрованими академічними словниками. Вагомим здобутком у цій галузі є Великий електронний словник української мови (ВЕСУМ), який містить понад 440 тисяч лематизованих форм [7] та генерує мільйони словоформ для різних граматичних категорій [8]. Для глибшого дослідження морфеміки існують спеціалізовані джерела, зокрема «Словник афіксальних морфем української

мови» [9], що містить вичерпний перелік префіксів та суфіксів із прикладами їхнього використання, та інші морфемні словники-довідники, у яких слова поділені на значущі частини за допомогою умовних позначень.

Ці ресурси містять обширну лінгвістично достовірну базу даних, яка є безцінною для наукових досліджень, статистичного аналізу та формування навчальних вибірок для нейромереж. Деякі вебпортали інтегрували ці бази даних у свої пошукові системи, даючи змогу користувачам переглядати морфемну будову конкретних лексем.

Головним недоліком усіх наявних лексикографічних баз та побудованих на них вебсервісів є статична репрезентація інформації. У їхній основі лежить принцип прямого пошуку за словником і такі системи не містять жодних алгоритмів автоматичного розбору, евристик чи моделей штучного інтелекту. Якщо користувач вводить слово, яке наявне в базі даних, система виводить заздалегідь підготовлений результат. Але мова є динамічною системою: щодня в ній з’являються нові слова, неологізми, терміни іншомовного походження, професійний сленг та оказіоналізми, наприклад, утворені парасинтетичним способом слова «айтівець» чи «кіберзахист». Якщо введеного слова немає в статичному реєстрі словника, система не може повернути результат, оскільки не має алгоритмічного апарату для того, щоб самостійно проаналізувати невідому лексему та виокремити з неї морфеми.

Наразі у відкритому доступі не існує готових та архітектурно гнучких автоматизованих сервісів або програмних бібліотек, які б здійснювали динамічний морфемний розбір українських слів, незалежно від наявності в статичних словниках. Інструменти морфологічної розмітки не ділять слово на морфеми, а морфемні словники не здатні обробляти нові, невідомі системі слова.

Це обґрунтовує наукову та практичну необхідність у розробленні спеціалізованої системи автоматичної морфемної сегментації. Вона має ґрунтуватися на інтеграції алгоритмічних методів та моделей машинного

навчання, що уможливить морфемний аналіз не лише загальновживаної лексики, а й неологізмів, розбір яких є неможливим у межах словникового підходу.

1.3. Детермінований словниковий підхід

У межах попереднього дослідження цієї теми [3] розроблено детермінований словниковий підхід до морфемного аналізу. Автоматизовано процес збору морфем із наявних лексикографічних баз української мови. У результаті оброблення та фільтрування даних сформовано морфемний корпус, що налічує 13 240 одиниць.

На основі цього корпусу створено алгоритм морфемного розбору, який працює за принципом перевірки потенційних комбінацій знайдених морфем у введеному слові. Метод базується на 64-х розроблених структурних варіантах поділу слова. Ці шаблони описують різні словотвірні моделі: від найпростіших одноморфемних слів, як-от «ліс», до складних конструкцій, які містять до семи морфем, як-от «багатофункціональний». Алгоритм генерує комбінації з наявних у словнику префіксів, коренів, суфіксів і закінчень; якщо згенерована послідовність морфем під час конкатенації збігається з вхідним словом, виведено успішний розбір.

Головною перевагою такого детермінованого підходу є лінгвістична коректність отриманих результатів. Оскільки система використовує для генерації гіпотез суто ті морфеми, які існують у словниках української мови, це унеможливлює проблему хибних передбачень. Наприклад, алгоритм не здатен помилково виділити буквосполуку «ро-» як префікс у слові «робити», оскільки такого префікса не існує в зібраному корпусі, а залишок слова не утворить валідної комбінації.

Водночас цей підхід має два недоліки. По-перше, алгоритм перевірки 64-х структурних комбінацій за допомогою декартового добутку є ресурсомістким. По-друге, система не є здатною обробити невідомі слова. Якщо користувач

вводить неологізм, запозичення чи професійний сленг, кореня якого немає в зібраному корпусі, алгоритм не побудує жодної комбінації і поверне помилку.

Проблема невідомих слів та значна ресурсомісткість доводять, що алгоритмічний підхід не може бути самостійним універсальним рішенням для української мови. Це зумовлює застосування методів машинного навчання. На відміну від правилозалежних алгоритмів, нейромережеві моделі здатні виявляти приховані лінгвістичні шаблони та узагальнювати контекст, що уможливлює сегментацію на морфеми слів, яких не було в навчальній вибірці.

1.4. Машинне навчання та нейромережі в завданні морфемної сегментації

Обмеження детермінованих словникових підходів, зокрема їхня вразливість до проблеми невідомої лексики та нездатність до екстраполяції правил на нові словоформи, зумовили використання методів машинного навчання. Використання імовірнісних та статистичних алгоритмів змістило фокус із словникового пошуку на виявлення прихованих лінгвістичних шаблонів і закономірностей усередині структури слова.

Першими спробами автоматизації морфемного аналізу за допомогою машинного навчання є алгоритми навчання без вчителя [10]. Прикладом такого підходу є родина алгоритмів Morfessor [11], розроблена для морфологічно багатих мов. Цей метод базується на принципі мінімальної довжини опису. Алгоритм аналізує великі масиви нерозміченого тексту й намагається знайти такий набір морфем, який би закодував весь корпус тексту найменшою кількістю символів. Алгоритм Morfessor здійснює базову сегментацію слова, знаходячи межі між морфемами на основі частотності їхньої сумісної появи. Обмеженість зазначених статистичних моделей зумовлена їхньою вузькою спрямованістю на ідентифікацію морфемних меж без можливості подальшої класифікації виділених одиниць — система ділить слово на частини, але не може визначити, яка з цих частин є префіксом, яка коренем, а яка — суфіксом.

Розвитку методів машинного навчання у сфері морфології сприяло проведення щорічних міжнародних кампаній під егідою SIGMORPHON. У рамках цих кампаній дослідницькі команди розробляють нейромережеві моделі для виконання завдань морфологічної словозміни та морфемної сегментації для типологічно різних та високофлективних мов. Результати цих досліджень підтвердили вищу продуктивність глибоких нейронних мереж у завданнях морфологічної словозміни порівняно з класичними алгоритмами та системами на основі правил, особливо для мов аглютинативного та синтетичного типів [12].

Для застосування нейромережевих архітектур до завдання морфемного розбору, його переформульовують із поділу тексту в посимвольну класифікацію послідовностей. У цьому підході слово розглянуто не як єдине ціле, а масив окремих літер. Кожній літері в слові нейромережа присвоює певний клас.

Для маркування літер використано загальноприйняту в комп'ютерній лінгвістиці схему BIO-маркування — «Begin, Inside, Outside». Згідно з цією схемою, кожну морфему поділено на літеру-початок (тег із префіксом «B-») та літери-продовження (тег із префіксом «I-»). Наприклад, коректне BIO-маркування символів слова «надточний» можна побачити в таблиці 1.1.

Такий формат перетворює лінгвістичне завдання на математичну задачу багатокласової класифікації, яку безперешкодно вирішують нейромережі.

Найвищу ефективність у задачах посимвольної класифікації демонструють архітектури на базі енкодерних трансформерів, зокрема моделі сімейства BERT та його оптимізованої версії RoBERTa [13][14]. На відміну від рекурентних нейронних мереж, які обробляли літери послідовно, трансформери використовують механізм самоуваги [15]. Цей механізм дає змогу моделі під час оброблення кожної окремої літери аналізувати весь контекст слова одночасно. Завдяки двосторонньому контексту модель розуміє взаємозв'язки між усіма символами лексеми. Наприклад, класифікуючи літери «ов», архітектура RoBERTa зважає на кінець слова: якщо далі закінчення «-ий», модель з високою

ймовірністю визначить буквосполуку як суфікс прикметника, а якщо слово закінчується на специфічний корінь — класифікує ці літери як частину основи.

Таблиця 1.1. Приклад BIO-маркування символів слова «надточний»

Літера	**Тип морфеми**	**BIO-мітка**
Н	початок префікса	B
А	продовження префікса	I
Д	продовження префікса	I
Т	початок кореня	B
О	продовження кореня	I
Ч	продовження кореня	I
Н	початок суфікса	B
И	початок закінчення	B
Й	продовження закінчення	I

Головною перевагою використання трансформерних нейромереж є їхня здатність до узагальнення [13]. Навчившись на великому корпусі розмічених слів, модель запам’ятовує глибинні структурні шаблони українського словотвору. Завдяки цьому вона може обробляти слова, яких не було в навчальній вибірці. Зіткнувшись із сучасним новотвором, модель виявить знайомі суфікси та префікси на краях слова й логічно визначить залишок літер у центрі як новий корінь, як-от у слові «блогер-ств-о».

Попри високу здатність до контекстуалізації, нейронні мережі характеризуються браком механізмів застосування граматичних правил. Процес формування результатів базується на розрахунку імовірнісних розподілів, що детерміновано під час опрацювання навчальної вибірки. Це призводить до явища хибних передбачень, яке в машинному навчанні називається «галюцинаціями».

Галюцинації у морфемній сегментації проявляються тоді, коли модель знаходить у слові популярний літерний кластер і класифікує його як афікс, ігноруючи лінгвістичну структуру слова. Наприклад, українська мова має префікс «при-», як-от в словах «при-йти», «при-нести». Якщо у навчальній вибірці містилася значна кількість слів із цим префіксом, у моделі формується статистичне упередження. Отримавши на вхід слово «приз», нейромережа може з високою впевненістю виділити буквосполуку «при» як префікс, залишивши єдину літеру «з» як корінь. З погляду математичної статистики для моделі такий поділ має логічний вигляд, проте лексикологічно це є помилкою.

Подібні статистичні похибки виникають на омонімічних морфемах, коли частина кореня збігається за написанням із афіксом, наприклад, хибне виділення суфікса «-ець-» у слові «перець». Оскільки модель не містить вбудований словник для перевірки згенерованого кореня, вона не усвідомлює, що залишок слова не має лексичного сенсу.

Отже, використання нейромереж та формату BIO-маркування забезпечує високу узагальнювальну здатність алгоритму і здатність до аналізу невідомої лексики, однак схильність таких моделей до генерації хибних морфемних структур через статистичні упередження свідчить про неможливість повної відмови від детермінованих методів контролю. Це зумовлює створення верифікаційних механізмів для обмеження вихідних даних нейромережі рамками словникових правил.

1.5. Можливості та обмеження великих мовних моделей

Сучасний етап розвитку галузі обробки природної мови характеризується широким впровадженням великих мовних моделей, побудованих на архітектурі породжувального штучного інтелекту. Ці системи вирішують широкий спектр лінгвістичних завдань і є здатними до глибокого розуміння контексту та продукування змістовних текстів різними мовами. Результати тестування мовних моделей у завданнях морфологічного аналізу підтверджують їхню

здатність успішно виконувати контекстно-залежні макрозавдання, зокрема безпомилково ідентифікувати частини мови у заданих масивах тексту [16].

Незважаючи на високу результативність у завданнях синтаксичного та семантичного рівнів, застосування великих мовних моделей для морфемної сегментації є обмеженим. Зниження правильності виокремлення кореневих та афіксальних морфем не є наслідком недоліків навчання, а зумовлене архітектурними особливостями етапу попередньої обробки вхідних даних. Функціонування переважної більшості сучасних мовних моделей базується на алгоритмі підслівного токенування. Його застосування має на меті оптимізацію обчислювальних ресурсів через встановлення балансу між посимвольним поданням тексту та повнослівним словником. Відповідно, такий алгоритм стискає вхідні дані завдяки ітеративному об’єднанню найчастотніших пар символів у єдині неподільні інформаційні блоки — токени.

З точки зору генерації тексту цей механізм є ефективним, однак для завдань морфеміки він створює архітектурну перешкоду. Алгоритм формує токени суто на основі статистичної частотності сумісної появи літер, ігноруючи лінгвістичні, етимологічні та орфографічні правила словотвору. Через це структурні межі згенерованих токенів не збігаються з реальними межами морфем.

Після завершення етапу токенізації нейромережа отримує на вхід масив токенів і сприймає кожен із них як базову, атомарну одиницю інформації. Велика мовна модель не має доступу до посимвольного розбору всередині сформованого токена. Відповідно, коли від генеративної моделі вимагається здійснити точний поділ слова на морфеми, вона змушена вгадувати межі афіксів та коренів.

Наприклад, при обробці складного українського слова статистичний токенізатор може фрагментувати його на випадкові підслівні блоки: розірвати єдиний корінь навпіл або, навпаки, склеїти кінець кореня із формотворчим суфіксом в один неподільний токен. Оскільки великі мовні моделі оперують

цілісними блоками, вони втрачають здатність реконструювати правильну лінгвістичну структуру, закладену на рівні окремих літер.

Отже, використання великих мовних моделей є результативним у вирішенні семантичних задач та розпізнаванні граматичних класів слів, проте їхнє фундаментальне архітектурне обмеження, зумовлене використанням алгоритмів підслівного кодування, унеможливлює доступ до літерного складу слова. Це робить такі системи непридатними для вирішення завдань, що вимагають точної структурної сегментації, зокрема для автоматичного морфемного розбору.

1.6. Висновок до першого розділу

Проведений аналіз специфіки української морфології та огляд наявних методів обробки природної мови дають змогу виокремити основні висновки щодо вирішення завдання автоматизованого морфемного аналізу.

По-перше, наявний інструментарій не пропонує комплексних рішень для внутрішньослівної сегментації. Сучасні морфологічні аналізатори успішно визначають граматичні категорії, проте сприймають слово як атомарну одиницю і не здійснюють його поділу на морфеми, а лексикографічні бази даних надають суто статичну інформацію без алгоритмів розбору невідомої лексики.

По-друге, використання великих мовних моделей для цієї задачі є неефективним. Попри високі результати в розпізнаванні частин мови, їхнє архітектурне обмеження унеможливлює доступ до літерного складу слова, що призводить до помилок при визначенні меж морфем.

По-третє, аналіз двох підходів до морфемної сегментації, детермінованого та нейромережевого, виявив їхні переваги та недоліки. Суто словниковий алгоритмічний підхід забезпечує лінгвістичну та граматичну достовірність, проте є вразливим до проблеми невідомих слів і не має гнучкості. Натомість моделі посимвольної класифікації на базі трансформерів демонструють високу

швидкість і здатність до узагальнення нових словоформ, але схильні до статистичних галюцинацій — хибного виділення афіксів, яких не існує.

Отже, використання одного методу унеможливлює створення універсального та точного інструменту. Це обґрунтовує необхідність розроблення гібридної каскадної системи морфемного та орфографічного аналізів. Така система об’єднає жорсткі детерміновані правила із контекстною передбачувальною здатністю нейромережевої архітектури.

Інтеграція цих підходів уможливить подолання слабких сторін кожного з них, забезпечивши як гнучкість в обробці неологізмів, так і лінгвістичну коректність фінального результату.

РОЗДІЛ 2
РОЗРОБЛЕННЯ ГІБРИДНОЇ СИСТЕМИ МОРФЕМНОГО АНАЛІЗУ УКРАЇНСЬКОМОВНИХ СЛІВ

Описано процес розроблення гібридної системи морфемного аналізу українськомовних слів — від формування навчального корпусу до створення фінального гібридного конвеєру та вебзастосунку для його демонстрації. Розроблення здійснювалося ітераційно: кожен наступний архітектурний підхід ґрунтується на аналізі слабких місць попереднього.

Окреслено процес збору та підготовки навчальних даних, вибір формату розмітки та стратегію поділу вибірки. Розглянуто три архітектурні підходи: базовий класифікатор, ансамбль спеціалізованих трансформерних моделей та єдина універсальна трансформерна нейромережа. По кожному підходу наведено архітектурне обґрунтування, деталі навчання та результати оцінювання на тестовій вибірці.

2.1. Формування та підготовка навчальної вибірки

Якість нейромережевої моделі залежить насамперед від якості та обсягу навчальних даних. Оскільки публічно доступного розміченого корпусу для морфемної сегментації українських слів у форматі, придатному для машинного навчання, не існувало, навчальну вибірку сформовано на основі двох лексикографічних джерел.

Першим джерелом є «Словник афіксальних морфем української мови» [9]. Цей ресурс містить зібрані українські слова з явно позначеними межами між морфемами завдяки умовним розділовим знакам, де кожному типу морфеми відповідає окреме позначення. Словник охоплює іменники, прикметники, дієслова, прислівники та інші частини мови, забезпечуючи типологічно різноманітний матеріал. Для дослідження зі Словника витягнуто слова та послідовність морфем із їхніми типами.

Другим джерелом є Великий електронний словник української мови (ВЕСУМ) [7], що містить понад 420 тисяч лем із повними парадигмами відмінювання та дієвідмінювання. ВЕСУМ не є морфемним словником — він не містить явних позначок морфемних меж, проте наявність у ВЕСУМ словозмінних парадигм, тобто сукупностей усіх форм слова, застосовано для автоматичного розширення вибірки: до навчальних даних додано словоформи вже розібраних на морфеми лем. Корінь і словотвірні суфікси залишаються незмінними, тоді як закінчення та префікси змінюються відповідно до конкретної словоформи.

Після злиття інформації із двох джерел і їхньої початкової нормалізації кожен запис у збірці даних зберігається у вигляді пари: словоформа та відповідна послідовність морфемних одиниць із типами. Типи морфем відповідають стандартним лінгвістичним категоріям: ROOT (корінь), PREF (префікс), SUFF (суфікс), ENDING (закінчення), INTER (інтерфікс).

Для кожного слова послідовність морфем є впорядкованим списком пар «буквосполука–тип». До прикладу, для слова «перехід» морфемна структура записується як: [(`пере`, PREF), (`хід`, ROOT)], а для «виробника» — [(`ви`, PREF), (`роб`, ROOT), (`ник`, SUFF), (`а`, ENDING)]. Такий формат є проміжним між лінгвістичним словником та навчальними даними для моделі та безпосередньо застосовується в перетворенні в BIO-послідовність.

BIO-маркування (Begin-Inside-Outside) є стандартною схемою для завдань розпізнавання іменованих сутностей та сегментації послідовностей у комп’ютерній лінгвістиці [4]. У контексті морфемного аналізу кожній літері слова присвоюється мітка, що вказує на початок морфеми («B-») або її продовження («I-»). Схему «Outside» у цій постановці завдання не задіяно, оскільки кожна літера обов’язково належить до однієї з морфем. Використання уніфікованої мітки ROOT без її диференціації на компоненти «B-» та «I-» зумовлене тим, що межі кореня визначаються на перетині суміжних афіксів, які вже мають суворе BIO-маркування. Аналогічний підхід без застосування префіксів «B-» та «I-» використано для класів закінчення (ENDING) та

інтерфікса (INTER). Це обґрунтовано граматичною специфікою української мови, згідно з якою такі морфеми не можуть дублюватися поспіль у межах однієї словоформи, що робить алгоритмічне розмежування їхніх стиків непотрібним. Це дає змогу оптимізувати простір вихідних класів моделі та підвищити стабільність її навчання за рахунок усунення надлишкового кодування.

Для морфемного розбору визначено 7 класів: ROOT, B-PREF, I-PREF, B-SUFF, I-SUFF, ENDING, INTER. Перетворення здійснено посимвольно: перша літера кожної морфеми отримує мітку із префіксом «B-», усі наступні літери тієї самої морфеми — мітку «I-» з тим самим типом. Завдяки такому перетворенню завдання морфемного розбору переформульовано як задачу посимвольної класифікації послідовностей — стандартну математичну задачу, для розв’язання якої розроблено широкий клас нейромережевих архітектур. Приклад обраного формату BIO-розмітки для символів слова «лісостеповий» можна побачити на таблиці 2.1.

Сирі дані з обох джерел містили низку артефактів, що потребували очищення. Виконано певні кроки фільтрації. По-перше, видалено рядки з неалфавітними символами — цифрами, дефісами та пунктуацією в основі слова, що можуть виникати через помилки розмітки. По-друге, видалено дублікати за парою «словоформа–морфемна структура»; одна й та сама форма слова з однаковим розбором траплялися в різних частинах словника. По-третє, застосовано балансування за лемами: для кожної леми залишено не більше п’яти словоформ, щоб запобігти домінуванню частотних парадигм над рідкісними. Це зменшує ефект перенавчання на типових закінченнях частотних слів та покращує здатність моделі до узагальнення.

Для збільшення охоплення вибірки щодо рідкісних словотвірних шаблонів застосовано два методи розширення. Перший — автоматична генерація словоформ префіксальних дієслів. Для кожного дієслова без префікса в корпусі генерувалися форми з найпоширенішими дієслівними префіксами (пере-, від-, до-, за-, при-, ви-, на-, по- тощо). Кожна згенерована пара перевірялася на

присутність у базі ВЕСУМ, щоб відсіяти лексично хибні комбінації. Такий підхід збільшив кількість прикладів із дієслівними префіксами, які є статистично рідкіснішими за суфіксальні конструкції.

Таблиця 2.1. Приклад BIO-маркування символів морфем слова «лісостеповий»

Літера	**Тип морфеми**	**BIO-мітка**
Л	корінь	ROOT
І	корінь	ROOT
С	корінь	ROOT
О	інтерфікс	INTER
С	корінь	ROOT
Т	корінь	ROOT
Е	корінь	ROOT
П	корінь	ROOT
О	початок суфікса	B-SUFF
В	продовження суфікса	I-SUFF
И	закінчення	ENDING
Й	закінчення	ENDING

Другий метод — відбір слів із ВЕСУМ за порогом впевненості нейромережевої моделі. На проміжних етапах навчання модель запускалася на нерозмічених словах із ВЕСУМ; слова, для яких модель демонструвала високу впевненість, відбиралися для навчальної вибірки.

Стандартне випадкове перемішування та поділ даних на навчальну, валідаційну та тестову вибірки призводить до витоку даних у завданнях, де різні форми одного слова присутні одночасно в навчальній і тестовій частинах. Модель, що навчалася на слові «виходити», а на тесті їй трапляється

«виходимо», фактично вирішує завдання впізнавання, а не узагальнення. Для коректного вимірювання здатності моделі до узагальнення реалізовано поділ без витоку даних за лемами.

Усі словоформи однієї леми потрапляють до одного з трьох поділів. Леми перемішуються випадково та розподіляються в пропорції 80 % / 10 % / 10 % для навчальної, валідаційної та тестової вибірок відповідно. Такий підхід гарантує, що тестова вибірка містить ті леми, на яких модель не навчалася, що є коректним оцінюванням узагальнюваної здатності.

2.2. Базовий класифікатор на основі випадкового лісу

Перед навчанням нейромережевих архітектур доцільно встановити базову лінію — простішу модель, що розв’язує те саме завдання. Така модель виконує дві функції: слугує нижньою межею очікуваної якості та дає змогу оцінити, наскільки складніша архітектура справді покращує результат. Для цієї ролі обрано класифікатор на основі випадкового лісу.

Випадковий ліс — ансамблевий метод машинного навчання, що будує значну кількість дерев рішень на випадкових підвибірках навчальних даних і приймає рішення голосуванням [17]. Істотна перевага цього алгоритму полягає в інтерпретованості: вдається отримати рейтинг важливості ознак та зрозуміти, на які символьні характеристики модель спирається найбільше. Окрім того, цей метод оптимально справляється з категоріальними ознаками та не потребує масштабування. Важливо, що цей алгоритм приймає рішення локально — для кожного символу незалежно, без урахування глобального контексту слова, що є вагомим обмеженням і головним предметом порівняння з трансформерними архітектурами.

Для того, щоб класифікатор міг обробити послідовність символів, кожну літеру слова перетворено в числовий вектор ознак. Використано два типи ознак: символьні n-грами у вікні ±3 та позиційні характеристики.

Символьні n-грами — це підрядки символів фіксованої довжини. Для кожної позиції і в слові вичленено всі символи у вікні від і-3 до і+3. З цих символів сформовано уніграми, біграми та триграми, закодовані у відповідному словнику. Такий підхід дає класифікатору інформацію про найближчий контекст позиції, однак обмежений вікном спостереження: модель не бачить початок чи кінець слова під час аналізу символу в середині.

Попри це, до вектора ознак додано позиційну інформацію: абсолютну позицію символу, відносну позицію (частку від довжини слова), відстань від початку та від кінця слова, загальну довжину слова. Ці ознаки дають алгоритму зрозуміти, що певні морфеми, зокрема закінчення, концентруються в кінці слова, а префікси — на початку.

Модель навчалася на 80-відсотковій частині збірки даних із безвитоковим поділом. Оптимізацію гіперпараметрів здійснено методом пошуку за сіткою на валідаційній вибірці. Варіювалися зокрема кількість дерев, максимальна глибина дерева, мінімальна кількість прикладів у листі, та критерій поділу. Найкращу конфігурацію встановлено за метрикою F1-macro на валідаційній вибірці.

Остаточною конфігурацією є 500 дерев, максимальна глибина 30, мінімальна кількість прикладів у листі — 5, критерій Gini. Кожне дерево навчалося на випадковій підвибірці навчальних прикладів та випадковій підмножині ознак на кожному вузлі поділу, що забезпечує різноманіття ансамблю та запобігає перенавчанню.

Якість моделі оцінено на тестовій вибірці за трьома основними метриками. Символьна точність — частка символів, клас яких визначено правильно, незалежно від правильності розбору слова в цілому. Точність повнослівного розбору — частка слів, у яких усі символи класифіковано правильно; ця метрика є вимогливішою, оскільки одна помилка в слові дає нульовий бал за все слово. F1-міра за типами морфем — середнє значення F1-міри для кожного з 7 BIO-класів.

Базовий класифікатор досяг символьної точності 95,2 %, точності повнослівного розбору 44,70 % та середньозваженої F1-міри 90,60 %. Різниця між символьною точністю та точністю повнослівного розбору пояснюється тим, що невелика частка неправильно класифікованих символів впливає на відсоток повністю правильних розборів. Детальні результати оцінювання базового класифікатора на основі алгоритму випадкового лісу на тестовій вибірці, охопно з покласовими метриками та загальною точністю системи, наведено в таблиці 2.2.

Таблиця 2.2. Результати оцінювання базового класифікатора на основі випадкового лісу

Клас	**Точність**	**Повнота**	**F1-міра**
Префікс	0,93	0,86	0,8934
Корінь	0,87	0,90	0,8876
Суфікс	0,93	0,92	0,9255
Закінчення	0,89	0,94	0,9136
Інтерфікс	0,92	0,61	0,7347
Метрика			**Результат, %**
Зважена середня F1-міра	—	—	**90,60**
Посимвольна точність	—	—	**87,10**
Точність повного розбору	—	—	**44,70**

Проаналізовано матрицю помилок моделі, яку зображено на рисунку 2.1, та виявлено, що найбільше модель плутає корінь із сусідніми йому суфіксами та префіксами: отримавши мітку ROOT для першої літери кореня, вона подекуди переривала корінь раніше, аніж слід, помилково класифікуючи наступні літери як суфіксальні, та навпаки, відносила до префіксу символи, що належали кореню. Також спостерігаються хибні передбачення символів суфіксів та закінчень.

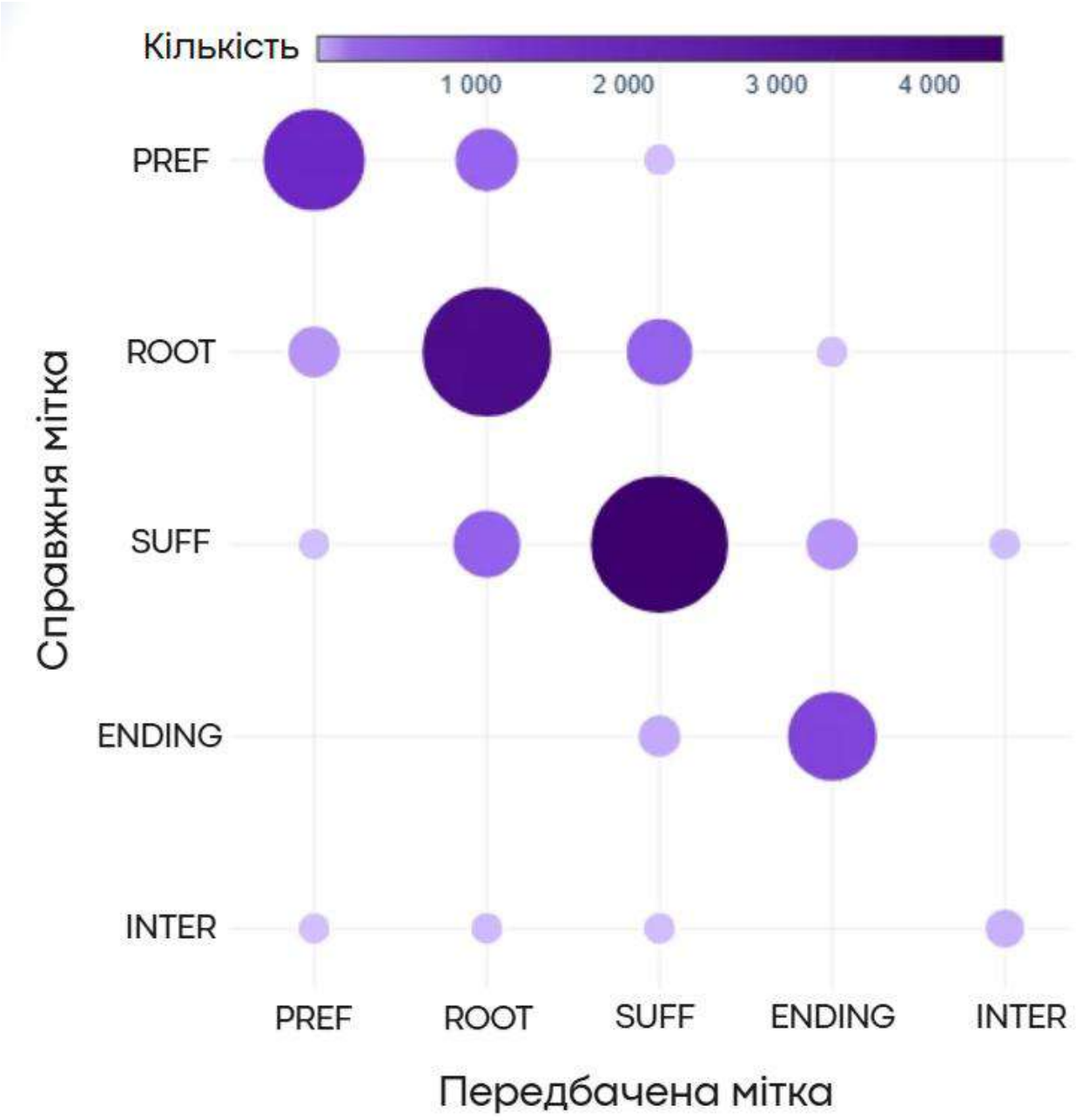

Рисунок 2.1. Матриця помилок базового класифікатора на основі випадкового лісу

Отже, підтверджено обмеження підходу: випадковий ліс приймає рішення локально на основі вікна ±3 символи без доступу до загальної структури слова. Клас символу на позиції і не залежить від того, до яких класів віднесено попередні символи, що спричиняє некоректні розбори.

2.3. Ансамбль спеціалізованих частиномовних трансформерних моделей

Аналіз помилок базового класифікатора засвідчив, що крім неправильних передбачень кореня з сусідніми афіксами, значна частина хибних передбачень також зосереджена на межах між суфіксами та закінченнями — ділянці, де морфемна структура відрізняється між частинами мови. Іменники, прикметники та дієслова мають різні парадигми суфіксації та флексії: дієслівні закінчення (-у, -еш, -емо) відрізняються від іменникових (-а, -у, -ові) та прикметникових (-ий, -ого, -ому). Висунуто гіпотезу: якщо навчати окремі моделі для кожної частини мови, кожна з них зможе вивчити специфічні шаблони своєї групи краще, аніж єдина модель, що обробляє всі класи одночасно.

Окрім того, поділ на частини мови балансує вибірку: дієслова, прикметники та іменники характеризуються відносною рівномірністю представлення в мовному корпусі, але мають різну морфологічну складність. Тренування окремих моделей гарантує, що кожна з них приділить рівний обсяг уваги специфічним словотвірним шаблонам своєї групи. На рисунку 2.2 зображено розподіл частин мови в навчальній вибірці.

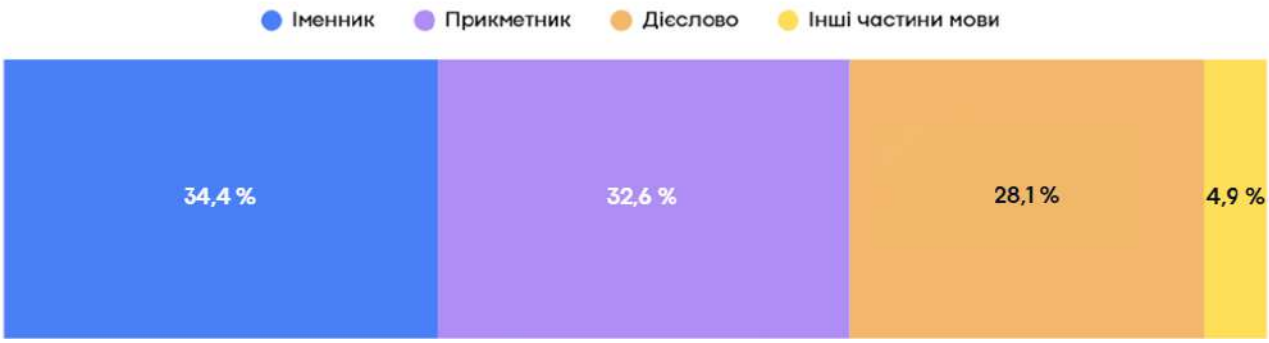

Рисунок 2.2. Частка представлення частин мови в навчальній вибірці

Для реалізації трансформерного підходу використано передтреновану модель українськомовного варіанту архітектури RoBERTa, youscan/ukr-roberta-base, навчену на великому корпусі текстів українською мовою [18]. Вибір цієї моделі обґрунтовано кількома факторами.

По-перше, морфемна сегментація залежить від закономірностей на рівні символів і підслів конкретної мови. Модель, передтреновану на українських текстах, уже закодовано знаннями про типові символьні шаблони, частотні буквосполучення та характерні закінчення.

По-друге, RoBERTa є оптимізованим варіантом BERT з покращеним режимом передтренування: без задачі передбачення наступного речення та з динамічним маскуванням токенів [14]. Ці зміни дають стабільніший та якісніший кодувальник для завдань класифікації токенів. Механізм двосторонньої самоуваги, притаманний архітектурі трансформера-кодувальника, є значущим для морфемної сегментації: класифікуючи поточний символ, модель одночасно бачить увесь контекст слова — і ліворуч, і праворуч.

Для маршрутизації вхідного слова до відповідної спеціалізованої моделі необхідно попередньо встановити його частину мови. Для цього використано морфологічний аналізатор UDPipe [1] — конвеєр обробки природної мови, що підтримує лематизацію, визначення частин мови та синтаксичний аналіз для низки мов, зокрема для української.

UDPipe повертає частину мови у форматі універсальних міток частин мови: NOUN (іменник), ADJ (прикметник), VERB (дієслово) та інші класи. На основі міток запит маршрутизується до однієї з трьох спеціалізованих моделей. Для слів, визначених як інші частини мови (числівники, прислівники, займенники), реалізовано запасну стратегію: запит спрямовується до моделі для іменників як найчисленнішої та найрізноманітнішої за морфемними шаблонами групи.

Загальний набір даних поділено на три підвибірки за частинами мови слів. Для кожного слова визначено частину мови через UDPipe та лематизованої форми. Вибірки для іменників, прикметників та дієслів формувалися незалежно, зі збереженням стратегії поділу за лемами без витоку всередині кожної підвибірки.

Кожна з трьох спеціалізованих моделей побудована за схемою класифікації токенів. Поверх кодувальника RoBERTa додано лінійний класифікаційний шар, що відображає приховані стани кожного символу в десятивимірний простір логітів для семи BIO-класів. Параметри кодувальника налаштовуються разом із класифікаційним шаром.

Для навчання застосовано такі гіперпараметри: 10 епох, розмір пакета даних 32, початкова швидкість навчання 2e-5 із косинусним планувальником та етапом поступового нарощування на перших 10 % кроків, а також спадання ваг на рівні 0,01 як регуляризатор. Для стабілізації процесу навчання застосовано згладжування міток із коефіцієнтом 0,1: замість «жорстких» міток (0 або 1) модель навчається наближатися до «пом’якшених» значень (0,1/(K-1) для хибних класів та 1-0,1 для правильного), що підвищує стійкість до шуму в розмітці.

Модель зберігалася на кожній валідаційній перевірці; за фінальну модель обрано ваги з найкращим значенням точного повнослівного розбору на валідаційній вибірці.

Кожну спеціалізовану модель оцінено на відповідній тестовій підвибірці, ансамблеву систему в цілому — на загальній тестовій вибірці. Як видно на

рисунку 2.3, модель для іменників досягла точності повнослівного розбору 58,59 %, для прикметників — 73,30 %, для дієслів — 79,99 %. Загальна точність ансамблю на об’єднаній тестовій вибірці становить 69,96 %, середньозважена F1-міра — 91,62 %.

Результати ансамблю підтверджують, що передтренована на великому українськомовному корпусі модель RoBERTa «знає» про типові буквосполучення та словозміну, що суттєво прискорює та покращує дотренування на морфемному завданні. Без такої ініціалізації трансформер тієї самої розмірності потребував би більшого навчального корпусу для досягнення порівнянних результатів.

Водночас виявлено обмеження ансамблевого підходу: фрагментація вибірки за частинами мови зменшує обсяг даних для кожної окремої моделі, що обмежує їхню здатність до узагальнення рідкісних словотвірних конструкцій — зокрема, парасинтетичних дієслів із кількома префіксами. Ці недоліки обґрунтовують перехід до єдиної універсальної моделі.

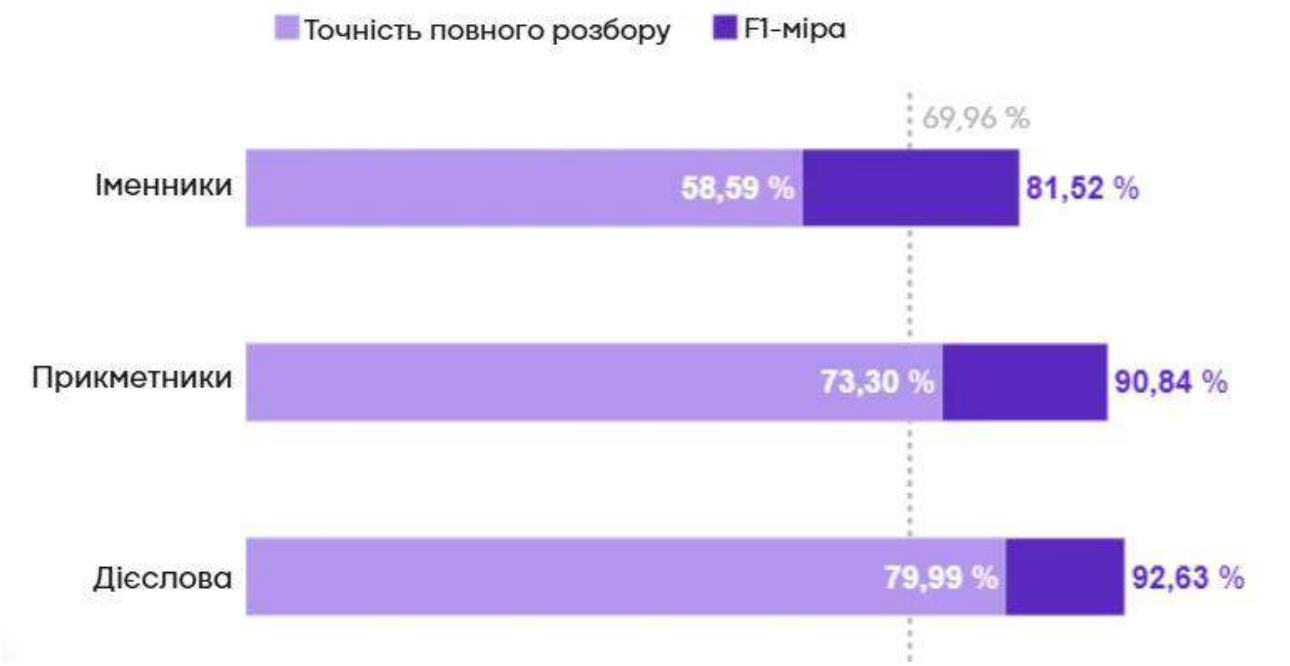

Рисунок 2.3. Результати оцінювання ансамблю спеціалізованих частиномовних трансформерних моделей

2.4. Універсальна трансформерна модель

Аналіз ансамблевого підходу виявив обмеження — втрату частини навчального потенціалу через роздроблення корпусу на три підвибірки. Цей недолік усувається переходом до єдиної моделі, навченої на повному наборі даних без поділу за частинами мови.

Гіпотеза полягає в тому, що трансформерна модель, навчена на всіх частинах мови одночасно, здатна самостійно виявити відповідні шаблони через механізм самоуваги — без підказки щодо граматичного класу слова. Іменникові та дієслівні парадигми словозміни є різними на рівні символів, тому єдина модель із достатнім обсягом навчальних даних зможе навчитися відрізняти їх без зовнішньої маршрутизації. Окрім того, збільшений обсяг навчальних даних для єдиної моделі порівняно з кожною з трьох спеціалізованих ліпше покриває рідкісні словотвірні конструкції.

Відмінністю з попереднім підходом є те, що вхідне слово подається безпосередньо в модель, без попередньої класифікації частини мови. Модель

бачить слово як послідовність символів та через механізм самоуваги формує контекстне уявлення кожного символу з урахуванням усіх інших — від першого до останнього. Такий двосторонній контекст є важливим для морфем із лексичною омонімією: модель, класифікуючи буквосполуку «ник» у середині слова, зважає на весь лівий та правий контекст і визначає, чи є вона суфіксом, як-от у слові «уч-ен-ик» або частиною кореня, як у слові «у-ник-а-ти».

Для навчання єдиної моделі сформовано фінальну збірку даних через об’єднання трьох джерел: початкового корпусу зі Словника афіксальних морфем, розширеного корпусу зі словоформами з ВЕСУМ та додаткових прикладів, отриманих методом активного навчання. Після злиття та повторної фільтрації фінальна вибірка налічує більше ста тисяч унікальних пар «словоформа–BIO-розмітка».

Поділ за лемами без витоків застосовано до об’єднаного набору даних з пропорцією 80 / 10 / 10. Перед навчанням підтверджено, що немає перетину між тестовими лемами та тренувальними — жодна тестова лема не наявна в навчальній частині.

Навчання моделі демонструє стабільне зростання точного повнослівного розбору на валідаційній вибірці від 66,35 % на першій епосі до 79,4 % на четвертій, після якої приріст сповільнюється, як видно на рисунку 2.4. Розрив між навчальною та валідаційною кривими залишається невеликим протягом усього процесу, отже модель не перенавчилась.

Показник правильно визначеної кількості морфем на слово виріс із 75,90 % на першій епосі до 86,13 % на останній.

На тестовій вибірці єдина трансформерна модель досягла точності повнослівного розбору 81,04 %, символьної точності 96,58 % та середньозваженої F1-міри 92,91 %.

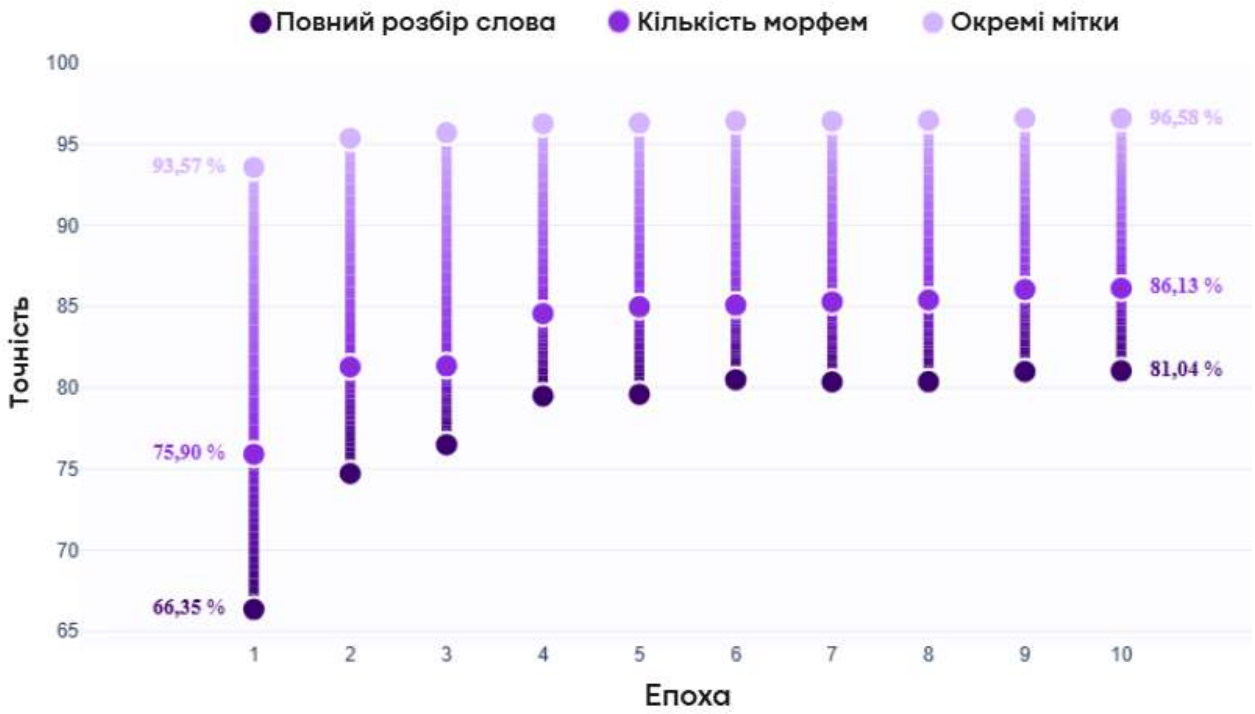

Рисунок 2.4. Результати оцінювання універсальної трансформерної моделі

F1-міра розподілена нерівномірно за типами морфем, що унаочнено на рисунку 2.5. Найвищі значення демонструє клас інтерфіксів із показником 0,973, а також клас префіксів із показником 0,955. Висока точність ідентифікації інтерфіксів обґрунтовується їхньою суворою позиційною обмеженістю (тільки на стику двох кореневих морфем) та вузьким символьним розмаїттям (переважно літери «о» та «е»). Префікси також є закритим класом морфем із чіткою локалізацією на початку словоформи. Натомість найнижчий показник зафіксовано для кореневих морфем із показником 0,917. Найнижча результативність для класу коренів є закономірною, оскільки корінь є найваріативнішою, відкритою частиною слова, що характеризується найбільшим лексичним розмаїттям.

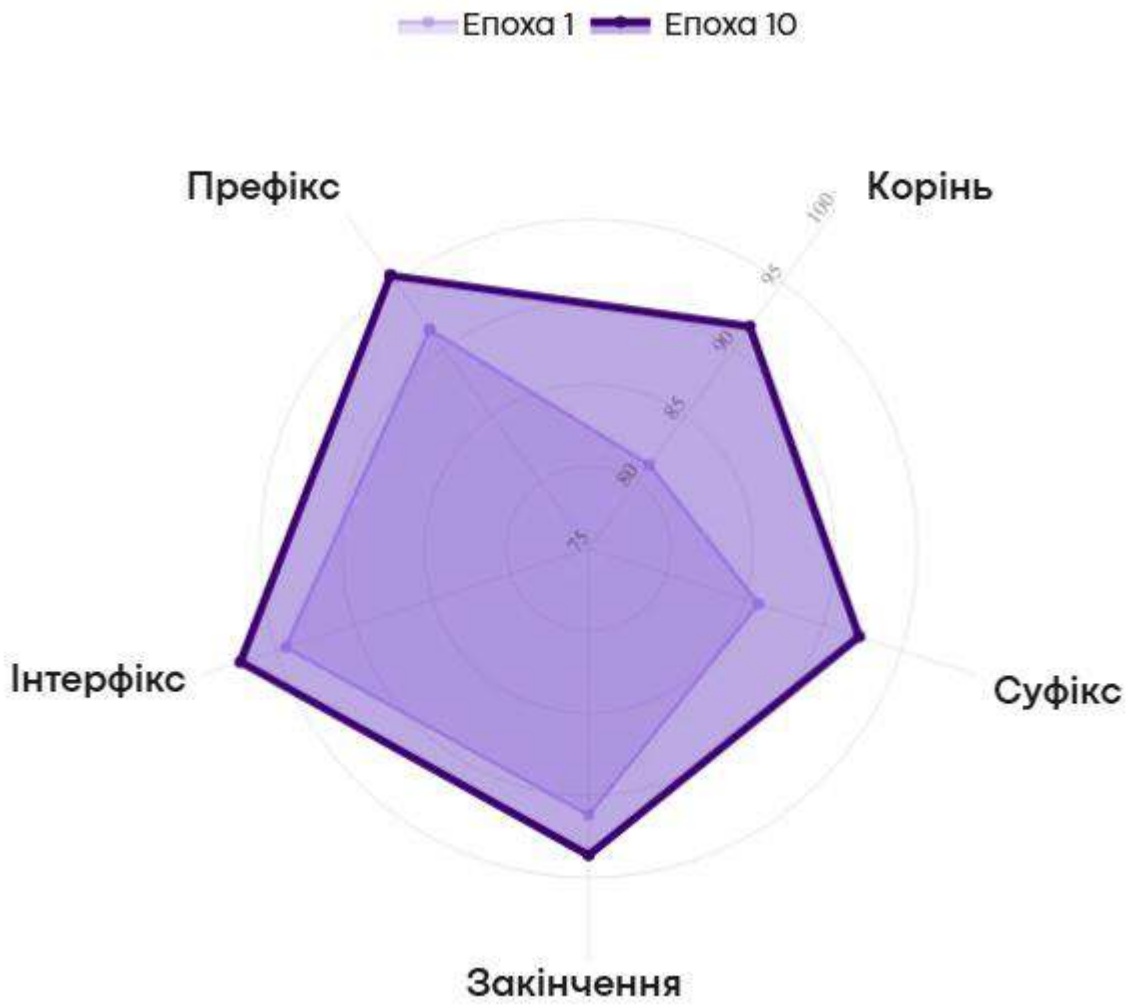

Рисунок 2.5. Зміна F1-міри за типами морфем від першої до десятої епохи

Незважаючи на суттєве покращення метрик якості, нейромережева модель зберігає генерацію некоректних морфем — буквосполук, які не існують у морфемних словниках української мови. Для діагностики цього явища зібрано всі префікси, суфікси та закінчення з передбачень моделі на тестовій вибірці та порівняно із зібраним корпусом афіксальних морфем.

На початку навчання модель генерувала 287 унікальних морфемних буквосполук, яких немає в корпусі. Після десяти епох цей показник знизився до 61. Некоректні морфеми зосереджені переважно серед рідкісних суфіксів та нестандартних закінчень, що виникають через помилки сегментації на межі «суфікс–закінчення». Це спостереження обґрунтовує необхідність верифікації афіксів за корпусом морфем.

2.5. Архітектура гібридної каскадної системи

Проаналізувавши архітектурні підходи, виявлено взаємодоповнювальні переваги та недоліки. Нейромережева модель здатна до узагальнення, але генерує хибні морфеми. Детермінований словниковий підхід із попереднього дослідження [3] гарантує лінгвістичну коректність, але якість розбору невідомої лексики залишається нестабільною через брак механізму узагальнення за межами словника. Пошук за словником для відомих слів забезпечує миттєвий результат, але за визначенням не має охоплення поза відомими прикладами. Гібридна каскадна система об’єднує ці підходи в єдину архітектуру, де результат нейромережевого передбачення послідовно проходить через кілька рівнів контролю якості, а детермінований резервний генератор активується лише в разі виявлення неправильних морфем із недостатньою впевненістю моделі.

Систему реалізовано у вигляді п’яти послідовних блоків, які схематично зображено на рисунку 2.6. На відміну від каскаду де кожен наступний етап замінює попередній, тут нейромережева модель запускається завжди — її передбачення та матриця ймовірностей є спільним входом для всіх наступних блоків. Різниця між блоками полягає не в тому, що вони роблять із вхідним словом, а в тому, яке рішення вони приймають щодо якості нейромережевого передбачення.

Рисунок 2.6. Схема архітектури п’ятиетапної гібридної системи морфемного розбору слова

Перший блок реалізує принцип мемоізації: результати вже виконаних розборів зберігаються та повертаються без повторних обчислень. Пошук за словником побудований на навчальному наборі даних — усі пари «слово–морфемна структура» завантажуються в пам’ять під час ініціалізації системи у вигляді хеш-таблиці з часом доступу O(1).

Цей блок охоплює найчастотнішу загальновживану лексику. Головна перевага — не якість, а швидкість: для слів зі швидкої пам’яті не відбувається жодних нейромережевих обчислень, що є важливим під час розгортання системи в режимі реального часу. Якщо введене слово знайдено в швидкій пам’яті, система повертає збережений результат та завершує оброблення. Якщо ні — запускається нейромережева модель і передбачення передається на другий блок.

Другий блок є основним обчислювальним ядром системи. Вхідне слово подається до навченої трансформерної моделі з класифікаційним шаром. Модель повертає не лише жорсткі мітки класів, а й повну матрицю ймовірностей розміром n × 7, де n — кількість символів у слові, 7 — кількість BIO-класів. Кожен елемент матриці P[i][j] відповідає ймовірності того, що

символ на позиції і належить до класу j. З матриці декодується BIO-послідовність: для кожного символу обирається клас із найвищою ймовірністю. Матриця ймовірностей зберігається в пам’яті та використовується всіма наступними блоками. Нейромережева модель завжди повертає результат — передбачення є для будь-якого вхідного слова. Питання не в наявності результату, а в його надійності, яку оцінюють третій і четвертий блоки.

Третій блок реалізує детермінований контроль якості над виходом нейромережі. Перевіряється відповідність передбаченої BIO-послідовності базовим лінгвістичним правилам структури слова.

У правильному розборі кожного слова присутня хоча б одна морфема типу ROOT. Якщо передбачення не містить жодного кореня — що можливо через надлишкову впевненість моделі в суфіксальних шаблонах для коротких слів — детермінований алгоритм перекласифіковує найдовшу послідовну морфему в ROOT. Відповідно до правил українського словотвору, префікси передують кореню, суфікси та закінчення йдуть після, інтерфікс розміщується між двома коренями в складному слові. Якщо модель передбачила суфікс перед коренем або закінчення перед суфіксом, такі морфеми перекласифіковуються на основі цього правила.

Якщо структурна перевірка не виявила жодних порушень — передбачення моделі вважається структурно коректним і передається на четвертий блок без змін.

Якщо порушення виявлено та виправлено — виправлена послідовність також передається на четвертий блок. В обох випадках четвертий блок отримує структурно валідний розбір і виносить фінальне рішення щодо його лінгвістичної коректності.

Четвертий блок перевіряє лінгвістичну валідність виокремлених афіксів. Для кожної морфеми типу PREF, SUFF, ENDING та INTER буквосполука звіряється з відповідним словником у корпусі афіксальних морфем. Паралельно для кожної морфеми обчислюється середнє значення ймовірностей моделі по її символах — зважена впевненість. Якщо всі афікси є в корпусі або ті, яких

немає, отримують зважену впевненість не нижче за поріг 0,90 — результат вважається надійним і повертається як фінальний. Поріг 0,90 інтерпретується так: модель є впевненою у своєму передбаченні для цих символів, а словникова нестача морфеми може пояснюватися неповнотою корпусу, а не помилкою моделі. Якщо ж виявлено морфему поза корпусом із впевненістю нижче за поріг — це сигнал про галюцинацію, і управління передається п’ятому блоку разом із матрицею ймовірностей.

П’ятий блок активується лише тоді, коли четвертий зафіксував хибну морфему з недостатньою впевненістю моделі. Основою цього рівня є детермінований комбінаторний алгоритм із попереднього дослідження [3], адаптований для використання в гібридній системі. Алгоритм генерує структурних кандидатів через перебір 64-х шаблонів поділу слова на морфеми, що описують різні словотвірні моделі — від найпростіших однокореневих слів до багатоморфемних конструкцій. Для кожного шаблону перевіряються всі можливі способи поділити слово на відповідну кількість частин, і кожна частина звіряється з відповідним словником корпусу.

Якщо декартовий добуток генерує кілька некоректних кандидатів, для ранжування використовується матриця ймовірностей нейромережевої моделі зі другого блоку. Для кожного кандидата обчислюється сума логітів відповідних BIO-класів по всіх символах — кандидат, чий поділ найкраще відповідає розподілу ймовірностей моделі, отримує найвищу оцінку. Так логіти перетворюються з кінцевого рішення на сигнал для ранжування, що усуває їхню головну ваду — схильність до генерації некоректних морфем — зберігаючи лінгвістичну валідність.

Якщо жодного правильного кандидата не знайдено — що можливо для неологізмів або без кореня у словнику — повертається нейромережеве передбачення після структурної корекції третього блоку.

2.6. Висновок до другого розділу

Описано цикл розроблення гібридної системи морфемного аналізу українськомовних слів — від підготовки даних до створення гібридної архітектури.

Сформовано навчальний корпус із пар «словоформа–BIO-розмітка» на основі Словника афіксальних морфем та ВЕСУМ. Реалізовано безвитоковий поділ за лемами, що забезпечує коректне оцінювання генералізаційної здатності моделей.

Навчено та порівняно три архітектурні підходи. Базовий класифікатор на основі випадкового лісу досяг точності повного розбору слова 44,70 % та виявив обмеження локальних методів — неможливість урахувати глобальний контекст слова. Ансамбль трьох спеціалізованих трансформерних моделей із маршрутизацією через UDPipe за частинами мови досяг 69,96 %, проте виявлено залежність від точності маршрутизатора та ефект фрагментації вибірки. Єдина універсальна трансформерна модель без поділу за частинами мови досягла 81,04 % та F1-macro 92,91 %.

На основі результатів навчання спроєктовано п’ятирівневий гібридний каскадний конвеєр: пошук за словником, нейромережева класифікація, перевірка структурної цілісності, верифікація афіксів та комбінаторний генератор.

РОЗДІЛ 3
СТВОРЕННЯ ВЕБЗАСТОСУНКУ ТА АПРОБАЦІЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

Розділ присвячено практичній реалізації результатів дослідження та їхньому оцінюванню. Описано процес вибору архітектурної основи вебзастосунку на підставі порівняльного аналізу трьох розроблених підходів — базового класифікатора на основі випадкового лісу, ансамблю частиномовних трансформерних моделей та універсальної трансформерної нейромережі. Обґрунтовано вибір моделі для інтеграції в гібридну систему та вебзастосунок. Описано архітектуру застосунку, реалізованого на основі гібридного п’ятиетапного конвеєру, та компоненти діагностичної візуалізації, що забезпечують інтерпретабельність роботи нейромережі. Для оцінювання практичної ефективності системи проведено експеримент на спільній тестовій вибірці, у якому порівняно результати повнослівного морфемного розбору п’яти великих мовних моделей, детермінованого комбінаторного алгоритму та розробленої гібридної системи.

3.1. Порівняльний аналіз розроблених підходів та вибір моделі

Навчено та оцінено три архітектурні підходи до завдання морфемної сегментації: базовий класифікатор на основі випадкового лісу із символьними n-грамами, ансамбль частиномовних трансформерних моделей та універсальну трансформерну нейромережу. Для забезпечення коректного порівняння всі три підходи оцінено на ідентичній тестовій вибірці, сформованій без витоку даних — без перетину лем між навчальною та тестовою частинами. За основну метрику обрано точність повнослівного розбору: результат зараховується як правильний лише тоді, коли модель правильно визначила межі та типи всіх морфем слова без жодної помилки. Зведені результати трьох підходів наведено в таблиці 3.1.

Таблиця 3.1. Порівняння метрик якості розроблених підходів на тестовій вибірці

Метод аналізу	**Архітектура**	**Точність повнослівного розбору**	**Середньозважена F1-міра**
Базовий класифікатор	Random Forest	44,70 %	90,79 %
Ансамбль частиномовних моделей	Transformer (POS-Split)	69,96 %	91,62 %
Універсальна модель	Transformer (End-to-End)	81,04 %	92,91 %

Базовий класифікатор на основі випадкового лісу із символьними n-грамами у вікні ±3 досяг точності повнослівного розбору 44,70 % та середньозваженої F1-міри 90,79 %. Висока символьна точність за низькій точності повнослівного розбору пояснюється фундаментальним обмеженням підходу: класифікатор приймає рішення для кожного символу незалежно, без урахування глобального контексту слова. Одна помилка на межі морфем обнуляє результат для всього слова за метрикою повнослівного розбору, тоді як символьна точність залишається високою. Попри нижчі результати, цей підхід виконує значущу методологічну функцію: він демонструє рівень, досяжний без трансферного навчання та передтренованих представлень, і слугує нижньою межею для порівняння.

Ансамбль частиномовних моделей показав результат 69,96 % за середньозваженій F1-мірі 91,62 %. Гіпотеза про те, що спеціалізовані моделі для окремих частин мови перевищать універсальну модель, не підтвердилася. Аналіз причин вказує на два чинники. По-перше, фрагментація навчальної вибірки — кожна з трьох моделей навчалася лише на підмножині даних відповідної частини мови, що зменшило обсяг навчання для кожної з них. По-друге, універсальна модель, навчаючись на всіх частинах мови одночасно,

виявляє спільні морфемні шаблони між ними — наприклад, суфікс «-ник» трапляється в іменниках і похідних прикметниках, і знання про нього, отримане з одного класу, підсилює розпізнавання в іншому. Під час поділу на окремі моделі цей перехресний сигнал втрачається.

Найкращий результат продемонструвала універсальна трансформерна модель: 81,04 % точності повнослівного розбору та 92,91 % середньозваженої F1-міри. Динаміка навчання підтверджує стабільне зростання всіх трьох метрик — точності повнослівного розбору, точності кількості морфем та символьної точності — впродовж 10 епох без ознак перенавчання. Модель навчилася узагальнювати морфемні шаблони поза межами словника: на словах, яких не було в навчальній вибірці, вона демонструє значно вищу точність, аніж детермінований алгоритм із попереднього дослідження.

На підставі результатів порівняльного аналізу для інтеграції в гібридну систему та вебзастосунок обрано універсальну трансформерну модель. Ця модель не лише показала найвищу точність повнослівного розбору серед усіх трьох підходів, а й забезпечує необхідні для гібридного конвеєру характеристики: матрицю ймовірностей для всіх символів слова, що використовується четвертим та п’ятим блоками для оцінювання якості та ранжування кандидатів.

3.2. Створення вебзастосунку

На основі гібридної системи розроблено вебзастосунок для інтерактивного морфемного аналізу українськомовних слів. Під час вибору технологічного стеку визначальними були три критерії: швидкість розгортання, нативна інтеграція з Python-екосистемою машинного навчання та інтерактивні можливості для побудови діагностичних візуалізацій.

Для побудови інтерфейсу обрано фреймворк Streamlit. Streamlit дає змогу створювати вебдодатки безпосередньо в Python-коді без необхідності писати HTML, CSS чи JavaScript [19]. Компоненти відображення, кнопки, поля

введення та графіки декларуються як виклики функцій у тому самому файлі, що містить логіку оброблення. Ця архітектура усуває надлишкові затримки на передавання даних між клієнтською та серверною частинами та спрощує підтримку застосунку.

Основна логіка обробляння реалізована з використанням бібліотеки Transformers від Hugging Face [20] та PyTorch [21] як обчислювальної платформи. Завантаження моделі під час старту застосунку внесено в швидку пам’ять, що гарантує одноразову ініціалізацію ваг незалежно від кількості запитів.

Для побудови всіх діагностичних візуалізацій обрано бібліотеку Plotly [22]. На відміну від статичних графіків Matplotlib [23], Plotly генерує інтерактивні елементи з підтримкою наведення курсору, масштабування та фільтрації [24].

Застосунок структурований за трирівневою схемою. Перший рівень — ініціалізація: під час першого запуску завантажується навчена модель із токенізатором, словникова інформація та корпус афіксальних морфем. Усі три компоненти зберігаються в пам’яті сесії та не перезавантажуються під час наступних запитів.

Другий рівень — обробка запиту. Після введення слова в текстове поле та натискання кнопки виконується весь п’ятиетапний процес. Разом із фінальним результатом розбору зберігається діагностична інформація: активований етап гібридної системи, матриця ймовірностей моделі, оцінка якості четвертого етапу та список перевірених афіксів.

Третій рівень — відображення результатів. Усі компоненти відображення організовано на одній сторінці нижче поля введення вхідного слова.

Перший елемент відображає основний результат — морфемну схему слова в стандартній нотації, прийнятій в українській лінгвістиці, як зображено на рисунку 3.1. Кожну морфему відображено як кольоровий блок із текстом; біля блоку розташовано підрядковий символ-позначку типу морфеми. Кольорове кодування морфем покликане зробити схему візуально читабельною — різні морфеми позначено різними кольорами.

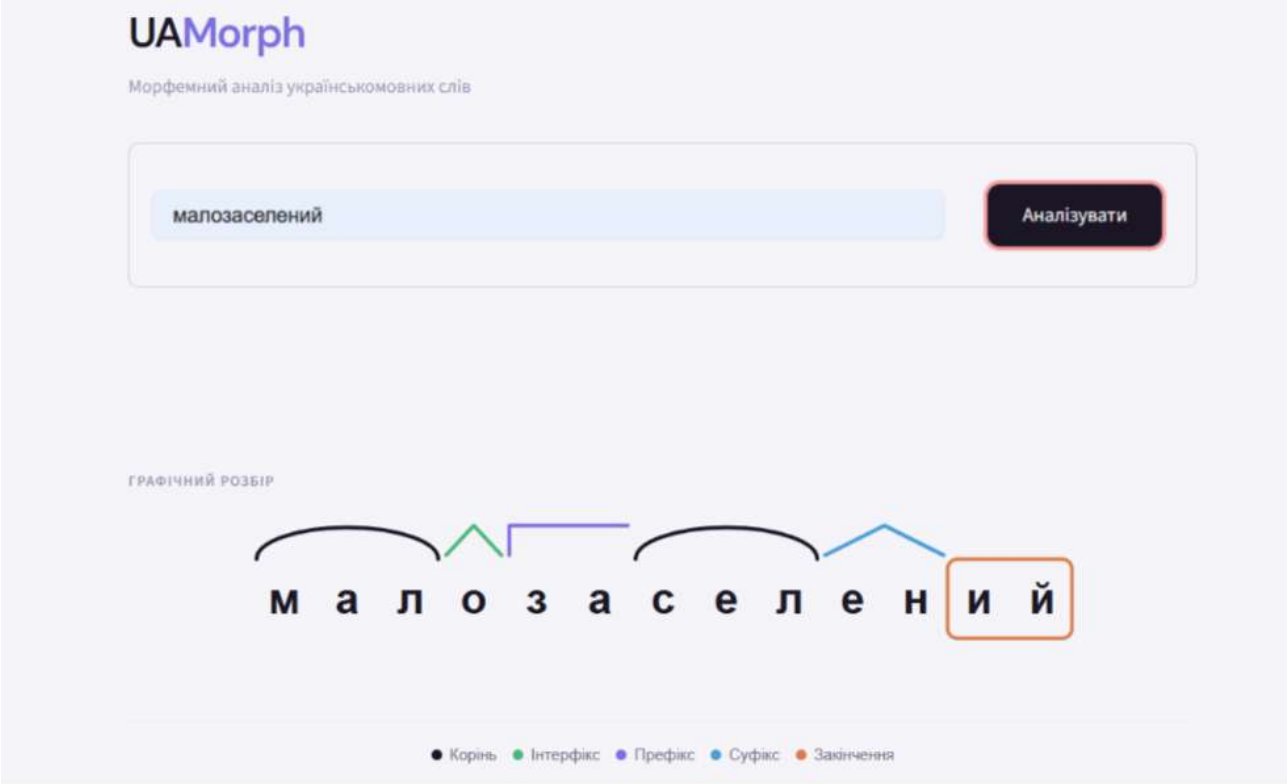

Рисунок 3.1. Результат морфемного розбору введеного слова в стандартній нотації

Нижче від схеми розташовано індикатор активованого етапу системи, що надає користувачу інформацію про те, яким саме механізмом отримано результат, як видно нижче на рисунку 3.2.

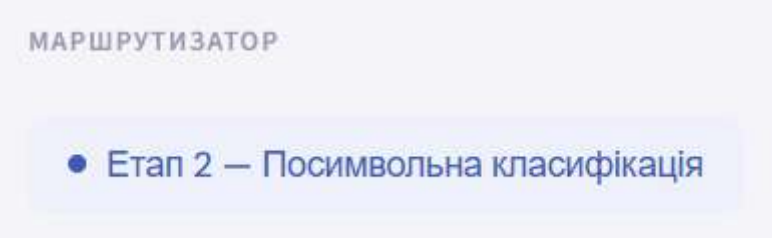

Рисунок 3.2. Індикатор активованого етапу системи

Друга візуалізація містить діагностичні візуалізації нейромережевих передбачень. Перша — стовпчаста діаграма впевненості моделі для кожного символу слова, яку видно на рисунку 3.3. По осі X відображено символи в

порядку їхньої появи в слові, по осі Y — максимальна ймовірність для переможного класу цього символу (тобто ймовірність обраної BIO-марки). Стовпці забарвлено в колір відповідного типу морфеми.

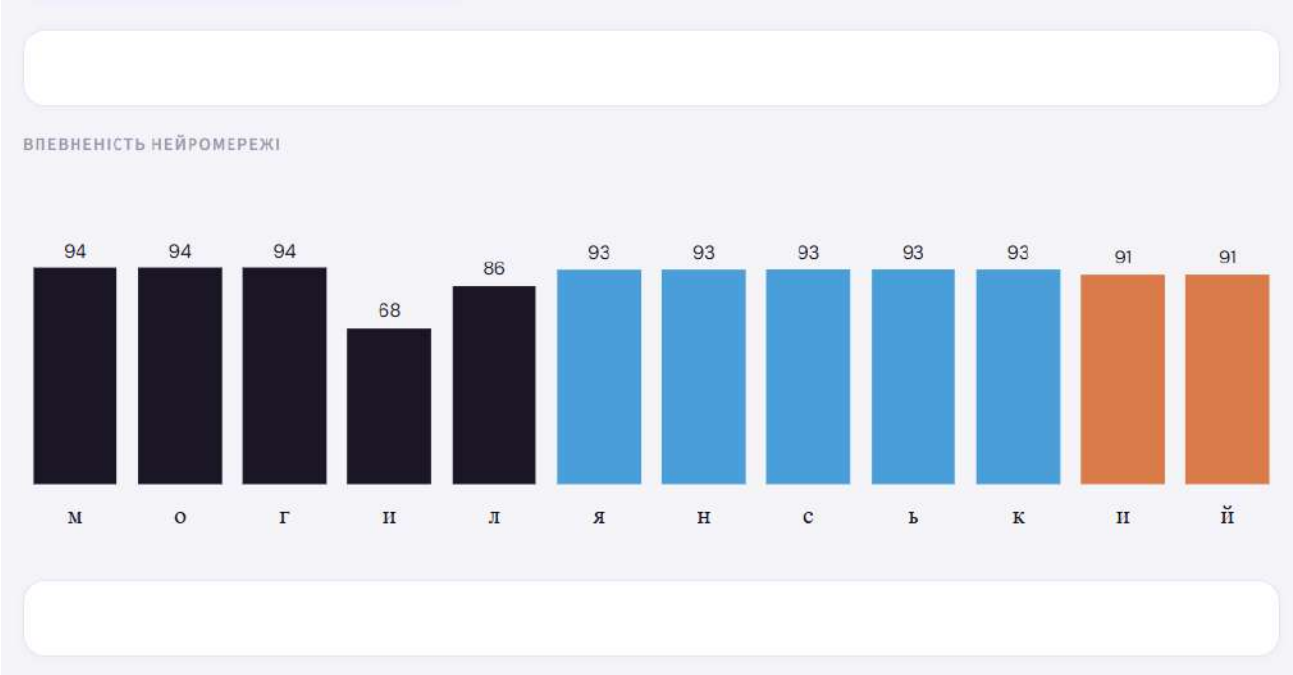

Рисунок 3.3. Впевненість моделі для кожного символу слова

Третя візуалізація — теплова карта ймовірностей, яку зображено на рисунку 3.4. По осі X — символи слова, по осі Y — 5 класів морфем. Кожна клітинка відображає ймовірність P[i][j] у вигляді кольорового кола. Теплова карта дає змогу побачити розподіл невпевненості: символи, для яких модель «коливається» між двома-трьома класами, матимуть одразу кілька забарвлених кіл у рядку, на відміну від символів із чіткою класифікацією.

Рисунок 3.4. Теплова карта ймовірностей передбачень моделі

Четверта візуалізація, що зображена на рисунку 3.5, відображає тривимірну проєкцію активацій нейромережі. Приховані стани останнього шару кодувальника для кожного символу слова редукуються до трьох вимірів. Кожна точка в 3D-просторі відповідає одному символу слова; точки забарвлено в колір відповідного типу морфеми. Просторова близькість точок відображає семантичну схожість контекстуальних уявлень символів: символи однієї морфеми, як правило, утворюють щільні кластери, тоді як символи на межах між морфемами розташовані між кластерами.

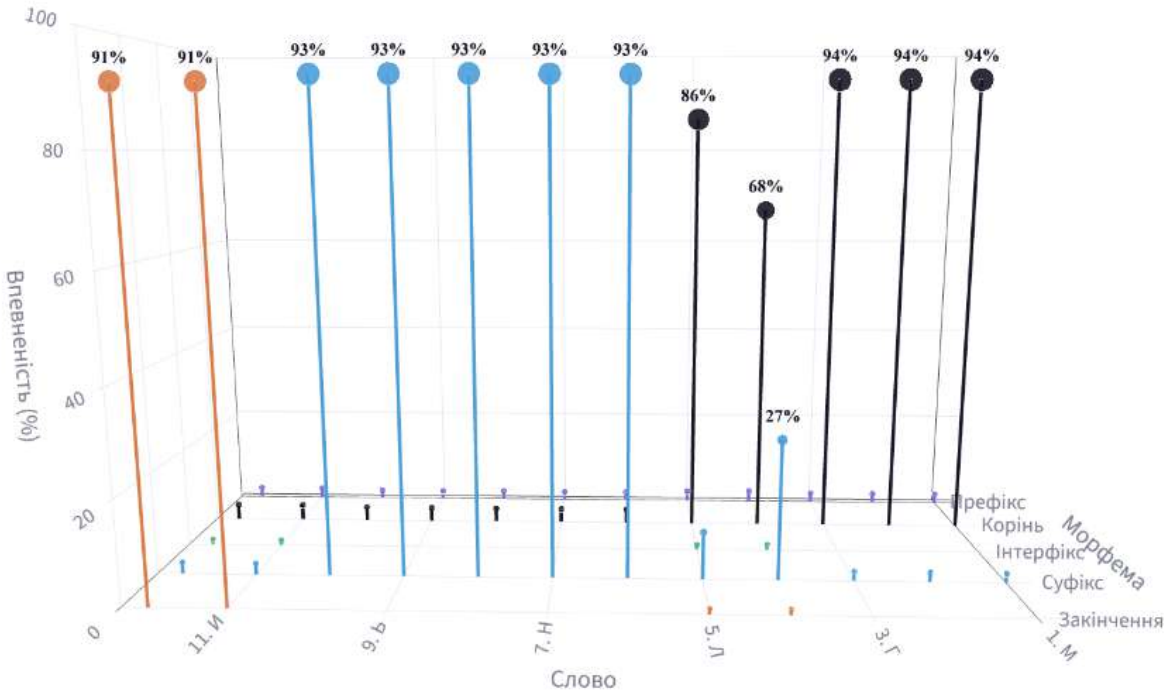

Рисунок 3.5. Тривимірна проєкція активацій нейромережі

П’яте зображення містить санки-діаграму, що відображає маршрут кожного символу від вхідної позиції до відповідного BIO-класу. Лівий стовпець вузлів представляє позиції символів у слові, правий — BIO-класи, що видно нижче на рисунку 3.6.

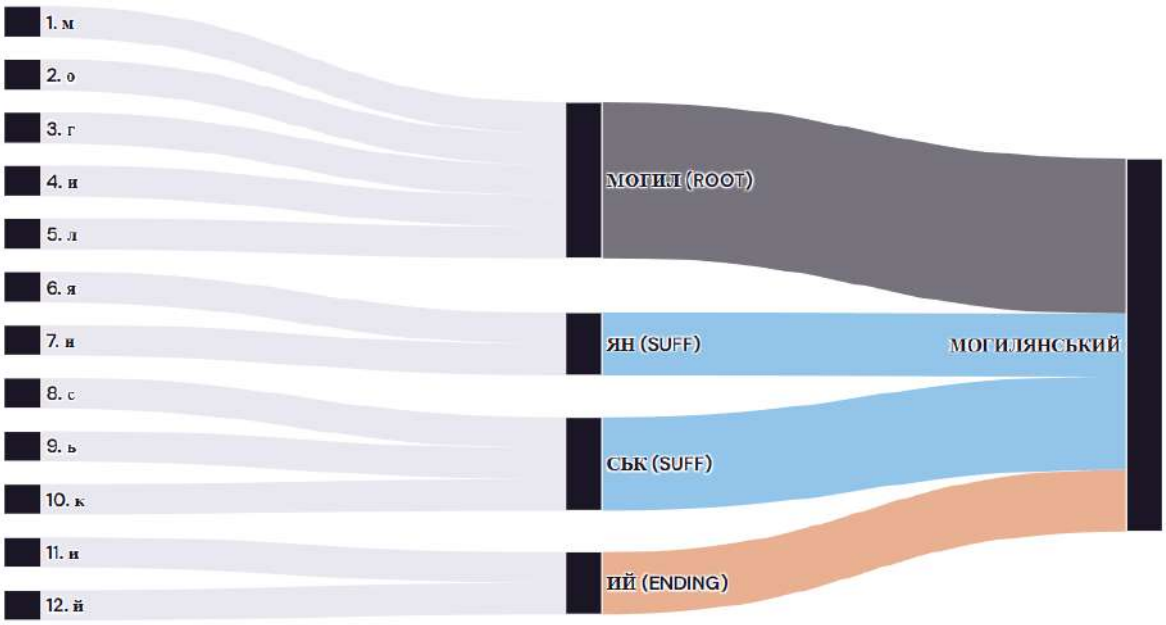

Рисунок 3.6. Санки-діаграма маршруту символів до BIO-класів

Поєднання різних типів візуалізацій забезпечує інтерпретабельність роботи нейромережі: від фінального результату через кількісну оцінку впевненості до просторового та маршрутного розуміння внутрішніх уявлень моделі. Такий рівень візуалізації перетворює застосунок із інструменту використання на інструмент дослідження — він дає змогу виявляти системні слабкості системи та напрями для покращення.

3.3. Апробація отриманих результатів

Для оцінювання практичної ефективності розробленої гібридної системи проведено порівняльний експеримент на спільній тестовій вибірці зі ста українськомовних слів. У вибірці є слова різних частин мови, типів словотвору та морфемної складності: від простих двоморфемних форм до семиморфемних складних слів із кількома префіксами та суфіксами. Вибірку добрано так, щоб рівною мірою представити іменники, прикметники та дієслова, а також забезпечити присутність морфемно омонімічних форм, запозичень та неологізмів — категорій, що традиційно становлять труднощі для автоматичних систем.

Кожну систему перевірено на ідентичному наборі слів. За одиницю оцінювання обрано точність повнослівного розбору: результат вважається правильним лише тоді, коли всі морфеми виокремлено коректно й кожній присвоєно правильний тип. Частковий залік за одну правильно визначену морфему не нараховується.

Порівняно виконання завдання морфемного розбору п’яти великих мовних моделей, комбінаторного алгоритму, розробленого у попередньому дослідженні, та гібридної системи. Для унаочнення на рисунку 3.7 зображено графік порівняння отриманих результатів.

Рисунок 3.7. Порівняння результатів виконання морфемного розбору

Гібридна система досягла точності 94 % — найвищого результату серед усіх порівнюваних підходів. Найближчий конкурент — комбінаторний алгоритм із результатом 54 %, поступаючись на 40 відсоткових пунктів, що підтверджує обмеження детермінованого підходу без нейромережевої частини: алгоритм коректно обробляє слова з поширеними словотвірними шаблонами, але не узагальнює на нетипові форми.

Результати великих мовних моделей виявилися суттєво нижчими за очікування. Grok 4 досяг 2,0 %, Gemini 3.1 Pro правильно розібрав 14,0 % слів, Claude Sonnet 4.6 — 11,0 %, а GPT-5.3 — 6,0 %. Ці показники підтверджують вищезазначене теоретичне обґрунтування: алгоритм підслівного токенування, що є в основі сучасних великих мовних моделей, ділить слово на статистичні блоки, не пов’язані з лінгвістичними межами морфем.

У результаті модель не має прямого доступу до посимвольного складу слова та вимушена реконструювати морфемну структуру через узагальнення на рівні підслів — завдання, для якого ці архітектури не оптимізовані.

3.4. Висновки до третього розділу

Розроблено вебзастосунок на основі гібридної системи, що забезпечує відображення морфемної схеми в стандартній нотації, комплексну візуалізацію впевненості нейромережі, тривимірну проєкцію активацій та санки-діаграму маршрутизації символів, перетворюючи інструмент кінцевого використання на платформу для дослідження та інтерпретації виведення моделі.

Результати апробації підтверджують перевагу гібридного підходу. Точність повного морфемного розбору 94,0 % перевищує результати серед великих мовних моделей на 80,0–92,0 % і в 1,7 раза перевищує результат детермінованого комбінаторного алгоритму. Поєднання трансформерної нейромережі з детермінованою верифікацією є результативнішим, аніж будь-який із цих підходів окремо: нейромережа здатна до узагальнення, механізм перевірки гарантує лінгвістичну коректність результату.

ВИСНОВКИ

Отже, мету роботи — автоматизувати процес морфемної сегментації українськомовних слів через розроблення гібридної системи для визначення структури слова — досягнуто, оскільки всі поставлені завдання виконано: досліджено наявні методи та рішення для морфемного аналізу; сформовано навчальну вибірку з морфемною розміткою; навчено базові класифікатори та нейромережеві моделі; порівняно результати трьох архітектурних підходів; розроблено гібридний алгоритм, що поєднує роботу нейромережі з мовними правилами; створено вебзастосунок для унаочнення процесу аналізу слова; оцінено точність системи на тестовій вибірці.

Кваліфікаційну роботу присвячено розробленню гібридної NLP-системи морфемного аналізу українськомовних слів, що поєднує трансформерну нейромережеву модель посимвольної класифікації з детермінованими лінгвістичними правилами.

У першому розділі проаналізовано специфіку морфологічної будови синтетичних флективних мов та досліджено обмеження наявних підходів до завдання морфемної сегментації. Встановлено, що морфосинтаксичні аналізатори розглядають слово як атомарну одиницю та не здійснюють морфемного поділу; лексикографічні бази функціонують на принципі прямого пошуку та не обробляють невідому лексику; великі мовні моделі через алгоритм підслівного токенування не мають прямого доступу до посимвольного складу слова. Проаналізовано детермінований комбінаторний підхід, розроблений у попередньому дослідженні, та обґрунтовано необхідність переходу до гібридної архітектури.

У другому розділі описано повний цикл розроблення системи. Сформовано навчальний корпус пар «словоформа–BIO-розмітка» на основі «Словника афіксальних морфем» та ВЕСУМ із застосуванням безвитокової стратегії поділу за лемами. Навчено та порівняно три архітектурні підходи. Базовий класифікатор на основі випадкового лісу досяг точності повного розбору слова

44,70 %, виявивши принципове обмеження локальних методів — неможливість урахувати глобальний контекст слова. Ансамбль трьох спеціалізованих трансформерних моделей із маршрутизацією через UDPipe досяг 69,96 % — приріст +25,26 відсоткових пункта — однак виявив залежність від точності маршрутизатора та ефект фрагментації вибірки. Єдина універсальна трансформерна модель без поділу за частинами мови досягла 81,04 % та F1-macro 92,91 %, що підтвердило перевагу навчання на повному корпусі. Спроєктовано п’ятирівневий гібридний каскадний конвеєр: пошук за словником, нейромережева класифікація, перевірка структурної цілісності, верифікація афіксів та комбінаторний генератор із ранжуванням кандидатів за логітами моделі.

У третьому розділі описано розроблення вебзастосунку та проведено апробацію системи. Застосунок реалізовано з використанням Streamlit, Transformers та Plotly; він відображає морфемну схему в шкільній нотації, діаграму впевненості нейромережі, теплову карту ймовірностей, тривимірну проєкцію активацій та санки-діаграму маршрутизації символів. Порівняльний експеримент на вибірці зі 100 слів засвідчив, що гібридна система досягає точності повного морфемного розбору 94,0 %, тоді як результати великих мовних моделей становлять від 2,0 % до 14,0 %, а детермінований комбінаторний алгоритм — 54,0 %. Приріст відносно найкращого альтернативного підходу становить 40 відсоткових пунктів.

Отже, підтверджено, що гібридна архітектура, яка поєднує нейромережеву гнучкість із детермінованою лінгвістичною перевіркою, є результативною. Нейромережева модель здатна до узагальнення невідомої лексики, тоді як верифікаційний механізм гарантує лінгвістичну коректність результату.

Практичним результатом роботи є комплексний вебзастосунок, навчена трансформерна модель, корпус афіксальних морфем і анотована збірка даних — ресурси, придатні для подальшого розвитку інструментів в галузі комп’ютерних наук для обробки природної української мови.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Straka M., Hajic J., Strakova J. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2015. P. 4290–4297. URL: https://aclanthology.org/L16-1680.pdf (date of access: 24.05.2026).
2. Sennrich R., Haddow B., Birch A. Neural Machine Translation of Rare Words with Subword Units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016. P. 1715–1725. URL: https://aclanthology.org/P16-1162.pdf (date of access: 24.05.2026).
3. Кирилін Є. С. NLP: Морфологічно-орфографічний аналіз слова. Київ : НаУКМА, 2024. 55 с.
4. Jurafsky D., Martin J. H. Speech and Language Processing. 3rd ed. draft. 2024. URL: https://web.stanford.edu/~jurafsky/slp3/ (date of access: 24.05.2026).
5. Плющ М. Я. Граматика української мови: У 2 ч. Ч. І. Морфеміка. Словотвір. Морфологія: Підручник. Київ, 2005. 288 с. URL: http://194.44.152.155/elib/local/sk/sk684094.pdf (дата звернення: 24.05.2026)
6. Етнолінгвістика: українська мова: навч. посіб. / П. Ю. Гриценко та ін. Чернівці, 2022. 432 с. URL: https://files.znu.edu.ua/files/Bibliobooks/Inshi74/0054757.pdf (дата звернення: 24.05.2026).
7. Рисін А., Старко В. Великий електронний словник української мови. Великий електронний словник української мови (ВЕСУМ). Вебверсія 6.8.0. 2005-2026. URL: https://vesum.nlp.net.ua/ (дата звернення: 24.05.2026).
8. Starko V., Rysin A. VESUM: A Large Morphological Dictionary of Ukrainian As a Dynamic Tool. 2022. URL: https://ceur-ws.org/Vol-3171/paper8.pdf (date of access: 24.05.2026).

9. Словник афіксальних морфем української мови. Київ, Україна : Ін-т мовознавства ім. О. О. Потебні НАН України, 1998. 429 с.
10. Goldsmith J. Unsupervised Learning of the Morphology of a Natural Language. Computational Linguistics. 2001. Vol. 27, no. 2. P. 153–198. URL: https://doi.org/10.1162/089120101750300490 (date of access: 24.05.2026).
11. Creutz M., Lagus K. Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0. 2005. URL: https://users.ics.aalto.fi/mcreutz/papers/Creutz05tr.pdf (date of access: 24.05.2026).
12. The SIGMORPHON 2016 Shared Task–Morphological Reinflection / R. Cotterell et al. 2016. URL: https://aclanthology.org/W16-2002.pdf (date of access: 24.05.2026).
13. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin et al. Proceedings of NAACL-HLT 2019. 2019. URL: https://aclanthology.org/N19-1423.pdf (date of access: 24.05.2026).
14. RoBERTa: A Robustly Optimized BERT Pretraining Approach / Liu Y. Et al. 2019. URL: https://arxiv.org/pdf/1907.11692 (date of access: 24.05.2026).
15. Attention Is All You Need / A. Vaswani et al. 31st Conference on Neural Information Processing Systems (NIPS 2017). 2017. URL: https://arxiv.org/pdf/1706.03762 (date of access: 24.05.2026).
16. What do you learn from context? Probing for sentence structure in contextualized word representations. / I. Tenney et al. Proceedings of the 2019 International Conference on Learning Representations (ICLR). 2019. URL: https://arxiv.org/pdf/1905.06316 (date of access: 24.05.2026).
17. Breiman L. Random Forests. Machine Learning. 2001. No. 45. P. 5–32. URL: https://link.springer.com/article/10.1023/A:1010933404324 (date of access: 24.05.2026).
18. Radchenko V. youscan/ukr-roberta-base. Hugging Face – The AI community building the future. URL: https://huggingface.co/youscan/ukr-roberta-base (date of access: 24.05.2026).

19. Streamlit Docs. Streamlit documentation. URL: https://docs.streamlit.io/ (date of access: 24.05.2026).
20. Transformers: State-of-the-Art Natural Language Processing / T. Wolf et al. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020. P. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6/ (date of access: 24.05.2026).
21. PyTorch: An Imperative Style, High-Performance Deep Learning Library / A. Paszke et al. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). 2019. URL: https://arxiv.org/abs/1912.01703 (date of access: 24.05.2026).
22. Plotly. Plotly Open Source Graphing Library for Python. URL: https://plotly.com/python/ (date of access: 24.05.2026).
23. Using Matplotlib – Matplotlib 3.9.0 documentation / J. Hunter та ін. Matplotlib – Visualization with Python. URL: https://matplotlib.org/stable/users/index (дата звернення: 24.05.2026).
24. Assessing the Performance of Python Data Visualization Libraries: A Review / A. Lavanya et al. 2023. URL: https://www.ijcert.org/index.php/ijcert/article/view/827 (date of access: 24.05.2026).