Міністерство освіти і науки України
Національний університет «Києво-Могилянська академія»
Факультет інформатики
Кафедра інформатики

Кваліфікаційна робота
освітній ступінь – магістр

на тему: «**ПОРІВНЯННЯ МОНО- ТА БАГАТО-МОВНОЇ МОДЕЛЕЙ НА ОСНОВІ BERT ДЛЯ ВИРІШЕННЯ ЗАДАЧ ОБРОБКИ МОВИ УКРАЇНСЬКОЮ**»

Виконав: студент 2-го року навчання,
Спеціальності
122 Комп'ютерні науки
Ванін Данило Олегович

Керівниця:
Крюкова Галина Віталіївна
доцент, кандидат фіз.-мат. наук

Рецензент:
Марченко Олександр Олександрович,
професор, доктор фіз.-мат. наук

Магістерська робота захищена
з оцінкою ____

Секретар ЕК _____
«____» _____ 20____ р

Київ – 2024

Міністерство освіти і науки України
Національний університет «Києво-Могилянська академія»
Факультет інформатики
Кафедра інформатики

ЗАТВЕРДЖУЮ
Зав. кафедри інформатики
к.ф-м.н., доц. Гороховський С.С
_____
(підпис)
“_____” _____2023 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ
на кваліфікаційну роботу
студенту 2 р.н. магістерської програми Комп’ютерні науки
Ваніну Данилу Олеговичу
Дослідити та порівняти роботу одно- та багатомовних моделей на завданнях обробки української мови

Зміст текстової частини до магістерської роботи:

Зміст
Перелік скорочень та термінів
Вступ
1. Основні поняття
2. Порівняння одно- та багатомовних моделей на завданнях української мови

Висновки
Перелік використаних джерел
Додатки

Дата видачі “_____” _____ 2023 р.

Керівник
Г.В. Крюкова, кандидат фізико-математичних наук, доцент
_____ (підпис)

Завдання отримав
Д.О. Ванін
_____(підпис)

Графік підготовки кваліфікаційної роботи до захисту

Графік узгоджено «_____» _____ 2023 р.

№ з/п	Перелік робіт	Термін	Підпис	Дата	Примітка
1.	Отримання теми кваліфікаційної роботи	10.10.2023			
2.	Ознайомлення з наявною інформацією за темою кваліфікаційної роботи	25.10.2023			
3.	Розробка плану та структури роботи	5.11.2023			
4.	Огляд літератури	15.11.2023			
5.	Дослідження одномовних моделей української мови	01.01.2024			
6.	Дослідження багатомовних моделей	15.01.2024			
7.	Дослідження наявних завдань та наборів даних для української мови	05.02.2024			
8.	Виконання порівняння роботи одно та багатомовних моделей	15.02.2024			
9.	Початок написання текстової частини	01.03.2024			
10.	Подання проміжної версії текстової частини	12.04.2024			
11.	Остаточне завершення написання текстової частини роботи	12.05.2024			
12.	Створення презентації	27.05.2024			
13.	Захист кваліфікаційної роботи	12.06.2024			

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ ТА ТЕРМІНІВ 6
ВСТУП 8
1. ОСНОВНІ ПОНЯТТЯ 11
1.1 Типи завдань з обробки природньої мови 11
1.1.1 Класифікація тексту (Text Classification) 11
1.1.2 Маркування послідовностей (Sequence Labeling) 12
1.1.3 Генерація мови (Language Generation) 12
1.1.4 Порівняння тексту (Text Comparison) 13
1.1.5 Інформаційний пошук (Information Retrieval) 13
1.1.6 Аналіз тексту (Text Analysis) 14
1.1.7 Розуміння мови (Language Understanding) 14
1.1.8 Зміна тексту (Text Modification) 14
1.2 Процес тренування моделей 15
1.2.1 Процес тренування базової моделі 15
1.2.2 Процес тонкого налаштування 16
1.3.1 Одномовний BERT 17
1.3.2 Багатомовний BERT 18
1.3.3 Висновки 19
1.4 Багато та одномовність у мовних моделях 19
1.4.1 Багатомовні моделі: результативність та крос-лінгвістичний трансфер 19
1.4.2 Одномовні моделі: спеціалізація та продуктивність у конкретних завданнях 20
1.4.3 Вплив представлення мови та розміру моделі 20
1.4.4 Вплив багатомовності на упередженість 21
1.4.5 Висновки 21
1.5 ОПМ української мови та виклики 21
1.6 Огляд робіт з порівняння одно- та багатомовних моделей 23
2. ПОРІВНЯННЯ ОДНО ТА БАГАТОМОВНИХ МОДЕЛЕЙ НА ЗАВДАННЯХ УКРАЇНСЬКОЇ МОВИ 25
2.1 Бенчмарки для оцінки результатів 25
2.2 Розпізнавання іменованих сутностей (NER) 26
2.3 Розрізнення значень слів (Word Sense Disambiguation) 28
2.4 Класифікація текстів 29
2.5 Eval-UA-tion 29

2.6 Заповнення пропусків (Mask filling) 31
2.7 Розмічування частин мови (POS tagging) 37
2.8 Висновки 38
3. ОБМЕЖЕННЯ ДОСЛІДЖЕННЯ ТА МАЙБУТНЯ РОБОТА 40
3.1 Обмеження дослідження 40
3.1.1 Обмеженість обчислювальних ресурсів та обсягів даних 40
3.1.2 Відсутність комплексного бенчмарку 40
3.1.3 Використання результатів конференції UNLP 2024 41
3.2 Майбутня робота та напрямки досліджень 41
ВИСНОВКИ 44
ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ 46
ДОДАТКИ 50
Додаток 1. Питання для оцінки моделей на завданні заповнення пропусків (mask filling). Згенеровані за допомогою ШІ 50
Додаток 2. Вебсайт для оцінки результатів заповнення пропусками моделей 55

ПЕРЕЛІК СКОРОЧЕНЬ ТА ТЕРМІНІВ

- **ОПМ (Обробка Природньої Мови)** – (від англ. Natural Language Processing, скороч. NLP) обробка природної мови, галузь штучного інтелекту та лінгвістики, яка займається взаємодією комп'ютерів з людською мовою в її природній формі.
- **ВММ (Великі Мовні Моделі)** – (від англ. Large Language Model, скороч. LLM) моделі, які використовують машинне навчання для обробки та генерації тексту.
- **Тонке налаштування / доналаштування / дотренування** – (англ. Finetuning) процес додаткового тренування моделі на специфічних даних для покращення її продуктивності в конкретних завданнях.
- **Попереднє навчання / базове навчання / навчання з «нуля»** – (англ. Pre-training) початковий етап тренування мовної моделі на великому обсязі текстових даних, метою якого зазвичай є натренувати модель «розуміти» мову (передбачати слова в контексті).
- **Характеристики / ознаки** – (англ. features) індивідуальні атрибути або властивості даних, що використовуються для навчання моделі.
- **Вкладання слів / векторне представлення слів** – (англ. word embedding) техніка представлення слів у вигляді векторів для обробки мовною моделлю.
- **Зворотне поширення** – (англ. backpropagation) метод оптимізації ваг в нейронних мережах шляхом поширення помилки назад через мережу.
- **Пакет / пакет даних** – (англ. batch) група зразків даних, що обробляються разом під час одного кроку тренування моделі.
- **Передові результати** – (англ. State of the Art, скороч. SoTA) найкращі досягнення в певній області або завданні на поточний момент.
- **Набір даних** – датасет, сукупність даних, використаних для тренування або тестування моделей.

- **Упередженість** – (англ. bias) схильність моделі робити систематичні помилки через нерівномірний розподіл тренувальних даних.
- **Розмітка** – (англ. label, labelling) процес надання категорій або міток даним для тренування моделі.
- **Бенчмарк** – (англ. benchmark) стандартизований набір даних та метрики оцінювання, які використовуються для оцінки та порівняння продуктивності різних мовних моделей у виконанні конкретних завдань.
- **Токенізатор** – програмний інструмент, який розбиває текст на окремі одиниці (токени), такі як слова, підслова чи символи, для подальшої обробки мовною моделлю.

ВСТУП

Останні дослідження у сфері глибокого навчання, зокрема створення та використання моделей на основні архітектури трансформера [2], як BERT [1], значно розширили можливості для вирішення задач обробки природньої мови (ОПМ).

Сучасні моделі зазвичай тренуються у два етапи: базове навчання на великих нерозмічених мовних корпусах та донавчання на менших наборах даних, специфічних для завдання (наприклад, класифікація тексту, аналіз насторїв, відповіді на питання). У результаті базового тренування моделі навчаються “розуміти” мову та передбачати слово, яке найкраще підійде у контексті. Якість результату цього навчання напряму залежить від якості та найважливіше розміру (сучасні моделі тренуються на терабайтах нерозмічених даних). У той час, як для популярних мов як англійська чи китайська, великі мовні корпуси відшукати достатньо легко - для мов з низькими ресурсами комплектування набору тренувальних достатнього розміру є відчутною проблемою.

Досить перспективною альтернативою для мов з обмеженими ресурсами стали багатомовні (multilingual) моделі. Ці моделі навчаються на злитих корпусах багатьох мов і показують високі результати на багатьох із них. Для мов з обмеженою кількістю ресурсів для навчання - багатомовні моделі одразу показували передові на той час результати. Причина цьому - крос-лінгвістичний трансфер, або перевикористання знань про одну мову для розуміння іншої. Схожий трансфер помітний у людях, які після вивчення французької мови, можуть набагато швидше вивчати англійську мову - через схожі слова, абетки, будову речення, ідіоми та запозичення між мовами. Таким чином, багатомовні моделі можуть перевикористовувати певні “знання” про мови з великими ресурсами на мовах з обмеженими.

У науковій літературі багато вивчають можливості одномовних та багатомовних моделей. Деякі дослідження демонструють, що одномовні моделі перевершують багатомовні на багатьох завданнях однієї мови. Наприклад, дослідження Virtanen et al. (2019) показало, що одномовна фінська модель BERT перевершила багатомовну модель BERT на різних завданнях фінської мови. Інші дослідження - показують, що багатомовні моделі можуть одночасно бути ефективними для великоресурсних мов та показувати кращі результати, ніж одномовні для тих мов, де ресурси обмежені (наприклад у випадку португальської мови).

Об’єктом дослідження цієї роботи є одно- та багатомовні моделі на основі BERT. Предметом дослідження є порівняння продуктивності таких моделей на завданнях ОПМ для української мови: розпізнавання іменованих сутностей, розрізнення значень, класифікація текстів, заповнення пропусків та розмічування частин мови. Мета дослідження - оцінити ефективність різних типів моделей на наборі задач. Для досягнення цієї мети були сформульовані наступні завдання:

1. Оцінка результатів одномовних моделей, які були натреновані на переважно українських текствах.
2. Оцінка адаптованих до української багатомовних моделей (наприклад, за методом WECHSEL)
3. Оцінка “чистих” багатомовних моделей
4. Порівняння результатів різних типів моделей та визначення особливостей їх використання

Методологічна основа включає використання стандартних підходів до навчання та оцінки моделей. У дослідженні використовувались або доступні дотреновані на завданні моделі, або базові моделі, які були дотреновані для завдання у рамках дослідження. Оцінювали моделі за допомогою таких метрик,

як F1-score (для розпізнавання іменованих сутностей) або точність (для розмічування частин мови). Завдання заповнення пропусків оцінювалось за допомогою опитування, а також через ручне порівняння та аналіз якості відповідей різних моделей. Для оцінки моделей також використовувались українська частина багатомовних бенчмарків або опубліковані українські.

Дослідження має як наукове, так і практичне значення. З наукової точки зору, воно допомагає краще розуміти компроміс між використанням одно та багатомовних моделей та потенційні особливості кожного з підходів. З практичного боку - результати дослідження можна використовувати для зваженого обрання найбільш оптимальної моделі для конкретного завдання. Зібрані результати порівняння та перелік моделей можна також використати, як основу для створення комплексного бенчмарку оцінки моделей для української мови.

1. ОСНОВНІ ПОНЯТТЯ

1.1 Типи завдань з обробки природньої мови

Обробка природної мови (ОПМ) включає в себе безліч завдань, спрямованих на те, щоб комп'ютери могли розуміти, інтерпретувати та генерувати людську мову. Нижче подано перелік основних завдань, що вирішуються у сфері ОПМ, та їхнє основне застосування.

1.1.1 Класифікація тексту (Text Classification)

1. Аналіз настроїв (Sentiment Analysis): Це завдання полягає у визначенні, чи є текст позитивним, негативним або нейтральним. Наприклад, аналіз відгуків клієнтів або повідомлень у соціальних мережах допомагає компаніям зрозуміти, що думають користувачі про їхні продукти чи послуги. Категорії для оцінки можуть розширюватись залежно від задачі, наприклад токсичність (для визначення токсичних коментарів) або правдивість (визначення фейків).
2. Класифікація тем (Topic Classification): Це завдання полягає у розподілі текстів за певними категоріями. Використовується для сортування новинних статей, наукових публікацій або постів у блогах за темами. Окрім тем статей, є також різновид цього завдання, де за текстом модель визначає, якого джерела він стосується (класифікація текстів за новинними виданнями)
3. Виявлення спаму (Spam Detection): Завданням є розпізнавання небажаних повідомлень, таких як спам-листи або коментарі. Окремим підвидом є визначення токсичного чи забороненого тексту.
4. Визначення наміру (Intent Detection): Це завдання допомагає визначити, чого саме хоче користувач від системи, особливо в контексті чат-ботів та віртуальних асистентів. Наприклад, мета відвідування сайта підтримки компанії, яка надає послуги електроенергії.

1.1.2 Маркування послідовностей (Sequence Labeling)

1. Розпізнавання іменованих сутностей (Named Entity Recognition, NER): Це завдання полягає в тому, щоб ідентифікувати власні назви (імена людей, назви міст, організацій тощо в тексті). Моделі також мають визначати якої власної назви чи об’єкта стосуються окремі займенники. Це може використовуватись для побудови баз знань, збагачення пошукових систем або виявлення згадок компанії чи людини в медіа.
2. Розмічування частин мови (Part-of-Speech Tagging): Це завдання полягає в наданні кожному слову в реченні граматичної категорії, наприклад, частини мови.
3. Групування (Chunking, Shallow Parsing): Це завдання полягає у виділенні фраз, наприклад, іменних або дієслівних груп. Це спрощує аналіз тексту без необхідності глибокого синтаксичного аналізу.
4. Визначення параметрів (Slot Filling): Це завдання стосується вилучення конкретної інформації з тексту. Наприклад, у діалогових системах це може бути дата або місце, де користувач хоче забронювати столик у ресторані. Ці моделі вже давно є частиною екосистеми Google чи Apple, які автоматично додають події з листів у календар, або окремих систем підтримки користувачів.

1.1.3 Генерація мови (Language Generation)

1. Генерація тексту (Text Generation): Суть завдання у створенні зв'язного та органічного тексту за запитом. Найчастіше застосовується у створенні контекнту.
2. Машинний переклад (Machine Translation): Це завдання звичайного перекладу між мовами. Моделі, спеціально натреновані на двомовних наборах даних, потенційно можуть показувати кращі результати перекладу, ніж звичайні автоматизовані перекладачі. Це пов’язано з тим, що сучасні моделі краще запам’ятовують контекст, а не перекладають

кожне слово окремо. Вони можуть навіть наводити альтернативні приказки та розуміти діалекти мов.

3. Стислий виклад (Summarization): Це завдання передбачає створення короткого (часто структурованого) переказу довгого тексту. Використовується, щоб отримати зміст із найважливішою інформацією з статті/документу або, як приклад, основні тези із транскрибованого інтерв'ю з клієнтом.
4. Перефразування (Paraphrasing): Це завдання полягає в тому, щоб подати наданий текст іншими словами, зберігаючи оригінальний зміст. Це корисно для спрощення тексту або створення варіантів контенту, наприклад у роботі копірайтерів.
5. Генерація діалогів (Dialogue Generation).

1.1.4 Порівняння тексту (Text Comparison)

1. Схожість тексту (Text Similarity): Це завдання передбачає формування оцінки, наскільки два тексти схожі між собою. Використовується для кластеризації документів, виявлення дублікатів або створення рекомендаційних систем.
2. Визначення перефразування (Paraphrase Identification): Це завдання працює з меншими одиницями тексту, ніж попереднє і визначає наскільки ідентичними є сенс двох фрагментів. Може використовуватись для перевірки на плагіат.
3. Оцінка семантичної схожості тексту (Semantic Textual Similarity): Це завдання оцінює ступінь семантичної схожості між реченнями. Використовується для інформаційного пошуку, завдань пошуку питання-відповіді по базі даних проблем у підтримці.

1.1.5 Інформаційний пошук (Information Retrieval)

1. Відповіді на запитання (Question Answering, QA): Це завдання полягає у пошуку відповідей на запитання. Часто таким моделям дають можливість

надсилати запити у бази знань чи користуватися додатковими інструментами. Моделі, які розв’язують це завдання часто є частиною пошукових систем, віртуальних асистентів та частиною служб підтримки.

2. Витяг інформації (Information Extraction): Це завдання автоматично вилучає структуровану інформацію з неструктурованого тексту. Використовується для баз знань та пошукових систем.
3. Пошук документів (Document Retrieval): Це завдання полягає в пошуку та ранжуванні документів відповідно до запиту користувача.

1.1.6 Аналіз тексту (Text Analysis)

1. Визначення тем (Topic Modeling): Це завдання схоже на завдання стислого викладу тексту. Корисне для створення цифрових картотек або пошукових баз.
2. Виявлення емоцій (Emotion Detection): Завдання полягає у пошуку та класифікації емоцій у тексті. Це особливо корисно для моніторингу соціальних мереж та аналізу відгуків користувачів.

1.1.7 Розуміння мови (Language Understanding)

1. Логічне виведення (Natural Language Inference): Логічне виведення, чи є задана гіпотеза правдивою, хибною відповідно до контексту. Може також повідомляти, якщо в контексті недостатньо даних для перевірки гіпотези.
2. Розуміння прочитаного (Reading Comprehension). Модель адаптована для цього завдання може коректно та повно відповідати на питання щодо наданого тексту.

1.1.8 Зміна тексту (Text Modification)

1. Нормалізація тексту (Text Normalization): Це процес перетворення тексту у стандартний вигляд, наприклад, розширення скорочень або виправлення орфографічних помилок. Це завдання часто розділяють на менші – виправлення лексичних помилок, граматичних помилок, стилістичних помилок.

2. Спрощення тексту (Text Simplification): Це завдання полягає у спрощенні тексту для легшого сприйняття та розуміння. Може використовуватись для адаптації текстів для людей, які менш знайомі зі складною темою академічного тексту чи людей, які не ідеально володіють мовою тексту.
3. Редагування тексту (Text Editing): Це завдання полягає в автоматичному покращенні граматики, стилю або зв'язності тексту. Від нормалізації його відрізняють типом запитів (перевести у інший стиль, додати персонажів, змінити тип оформлення діалогу).

У цій роботі для порівняння одно- та багатомовних моделей будуть обрані ті завдання, для яких існують вже готові набори даних чи бенчмарки українською мовою, є результати оцінки україномовних моделей або публічно доступні моделі, які дотреновані або можна дотренувати для певного завдання.

1.2 Процес тренування моделей

Моделі обробки природної мови — це складні алгоритми, які створені для розуміння, інтерпретації та генерації людської мови. Їх тренування та налаштування відбуваються у кілька етапів, кожен з яких покращує їх здатність виконувати конкретні завдання в обробці природної мови.

1.2.1 Процес тренування базової моделі

Першим етапом у тренуванні є збір даних та попередня обробка. Основою для будь-якої моделі ОПМ є великий (терабайти даних) і різноманітний (різні жанри, джерела, формати) набір текстових даних. Такі дані можна отримати з книг, статей, веб-сайтів, соціальних мереж та інших джерел, що відповідають цільовій задачі. Потім необроблений текст очищують від шуму, виправляють помилки та приводять у формат, придатний для використання моделлю.

Наступним етапом є вилучення характеристик. Моделі ОПМ не можуть безпосередньо працювати з необробленим текстом. Вони покладаються на

числові представлення слів або їх частин (наприклад, символи чи фрагменти слів). Для цього використовуються такі методи, як вкладання слів (Word2Vec, GloVe) або більш складні моделі на основі трансформерів (BERT, GPT), що перетворюють слова у вектори у високорозмірному просторі. Ці вектори захоплюють семантичні зв'язки між словами – слова, які схожі за сенсом будуть мати високий косинус подібності між векторами-репрезентаціями.

Вибір архітектури моделі залежить від конкретного завдання ОПМ. Наприклад, для нескладних завдань типу "послідовність-послідовність" (машинний переклад), можуть бути достатніми рекурентні нейронні мережі (RNN). Завдання класифікації спаму може бути вирішене і наївним баєсивим класифікатором. Поширена зараз трансформерна архітектура стала дуже популярною завдяки універсальності та здатності навчатися виявляти зв’язки між елементами тексту, розташованими на різних відстанях. Вони вирішують проблему згасання важливості контексту та регулярного перерахунку контекстних ваг у рекурентних архітектурах.

Модель тренується за допомогою алгоритму навчання з учителем, наприклад, стохастичного градієнтного спуску. Під час цього ваги функції, яку представляє з себе модель, поступово змінюються, щоб досягнути глобального чи локального мінімума.

Часто завданням на етапі базового тренування є просте передбачення слова у контексті. Наприклад, модель навчається передбачувати «замасковане» слово у реченні. Для цього завдання не потрібні розмічені дані і для навчання потенційно можна скористатися будь-яким якісним текстом на цільовій мові.

1.2.2 Процес тонкого налаштування

Попри те, що моделі, такі як BERT, вже мають загальне розуміння мови (вміння передбачати слова у контексті) – цього недостатньо для більш специфічних завдань. Базова модель може лише передбачати масковані слова, а

відповідно для більш цінного завдання, як відповідь на питання чи переказ тексту – модель має бути донавченою на розмічених даних. Ці дані зазвичай мають значно менший обсяг та більш сфокусовані, ніж оригінальні тренувальні набори. Для тренування моделі, яка відповідає на питання, таким набором даних будуть пари питання-відповідь.

Замість того, щоб навчати модель «з нуля», процес тонкого налаштування починається з використання вже наявної попередньо навченої моделі, яку адаптують під конкретне завдання. Це дозволяє «перевикористати» знання, отримані з великого неструктурованого набору даних під час попереднього навчання. Додавання нових даних, які близьки до кінцевого завдання, допомагає налаштувати модель для вирішення конкретних задач.

В залежності від завдання, може виникнути потреба в незначних змінах архітектури моделі. Наприклад, для аналізу настроїв можна додати класифікаційний шар нейронів, або розморозити деякі шари базової моделі. Такі потреби зазвичай визначають емпірично, хоча у спільноті дослідників часто вже є відпрацьований набір змін, які потрібні для вирішення конкретної проблеми.

Тонке налаштування включає оптимізацію гіперпараметрів, таких як швидкість навчання, розмір пакету даних та кількість епох навчання. Це дозволяє досягти найкращої продуктивності на специфічних для завдання даних і також визначається емпірично.

1.3.1 Одномовний BERT

Одномовний BERT [1] (Bidirectional Encoder Representations from Transformers - двонаправлене кодування представлень з трансформерів) - це мовна модель, яка використовує трансформерну архітектуру. Вона яка складається з шарів само-уваги (self-attention) та нейронних мереж прямого поширення. Модель попередньо навчається на великих одномовних корпусах через передбачення замаскованого токена у нерозміченому тексті. У цьому

методі частина вхідних токенів випадково маскується, а модель навчається прогнозувати їх на основі контексту [1].

Такий підхід дозволяє BERT запам'ятовувати інформацію про мову з неструктурованих даних, як-от значення окремих слів, порядок слів у реченні та граматичні категорії. Завдяки цьому модель можна легко налаштувати для різних завдань з мінімальними модифікаціями, специфічними для завдання [1].

Після виходу BERT показав передові результати в 11 завданнях обробки природної мови [1] і став основою для численних майбутніх моделей, таких як RoBERTa [4], DistilBERT [5] та інші.

1.3.2 Багатомовний BERT

Розроблені багатомовні варіанти BERT, такі як mBERT та XLM-RoBERTa [6], розширюють можливості моделі для роботи з кількома мовами. Ці моделі навчаються на об’єднаних корпусах до 100 мов, використовуючи спільну інформацію між різними мовами для покращення роботи на окремих мовах, а особливо для мов з низьким рівнем ресурсів, таких як українська [6]. Основна ідея полягає в тому, що літери та фрагменти слів часто збігаються в схожих мовах, що спрощує тренування токенізатора, а також багато мов мають подібну семантичну структуру, логіку побудови слів чи навіть спільні слова. Використовуючи багатомовні корпуси, вдається значно збільшити розмір тренувальних даних, а модель може скористатися зі схожості різних мов.

Проте ефективність багатомовних моделей BERT може варіюватися залежно від конкретної мови та завдання. У той час, як вони демонструють високу продуктивність на мовах з великою кількістю ресурсів (наприклад, англійська, китайська, іспанська), їх результати на мовах з низькими ресурсами може бути обмеженою. Базова причина цьому недостатня репрезентованість розуміння мови в вагах моделі. Чим більшу частку складають дані певної мови в тренувальному наборі, тим кращими будуть результати моделі для цієї мови.

Більш детальний огляд та порівняння одно- та багатомовних моделей буде надано у наступних частинах роботи.

1.3.3 Висновки

Хоча моделі BERT досягли значних успіхів на популярних мовах, все ще існують виклики для мов з недостатніми ресурсами. Обмежена доступність якісних та достатніх за обсягом наборів навчальних даних для цих мов може перешкоджати продуктивності як одномовних, так і багатомовних моделей. Питання того, які з двох моделей є більш ефективними для конкретної мови можна визначити лише емпірично. Ця робота присвячена такому емпіричному порівнянню моделей на різних завданнях ОПМ для української.

1.4 Багато та одномовність у мовних моделях

Вибір між багатомовними та одномовними моделями для задач обробки природної мови є складним, оскільки кожен підхід має свої переваги та недоліки.

1.4.1 Багатомовні моделі: результативність та крос-лінгвістичний трансфер

Багатомовні моделі, навчені на даних з кількох мов, продемонстрували вражаючі можливості. Вони можуть досягати передових результатів у широкому діапазоні мов, зберігаючи при цьому високу продуктивність у мовах з великими ресурсами [7]. Такий "крос-лінгвістичний трансфер" дозволяє моделям використовувати знання, отримані в одній мові, для інших мов, що особливо корисно для мов з обмеженими ресурсами [6]. Автори [8] показали, що багатомовні моделі можуть досягати наближену або навіть вищу продуктивність на декількох мовах порівняно з одномовними моделями. Цікаво, що такі результати досягаються незважаючи на навчання на значно меншій кількості даних у цих конкретних мовах.

1.4.2 Одномовні моделі: спеціалізація та продуктивність у конкретних завданнях

У той же час, одномовні моделі часто можуть перегнати свої багатомовні аналоги у деяких завданнях, включаючи виявлення мови ворожнечі та генерування речень [9, 10]. Ця вища продуктивність може бути пов'язана зі здатністю одномовної моделі глибше розуміти нюанси і тонкощі однієї мови, а не розподіляти свої ресурси між кількома мовами [11]. Важливим також є тип даних, на яких тренувалася модель. Багатомовні моделі часто навчають на даних з Вікіпедії або онлайн-видавництв. Дослідники, які тренують одномовні моделі, можуть приділити більше часу на збір більш нішевих наборів даних. Баланс між одномовними та багатомовними моделями досить цікавий. Наприклад у деяких дослідженнях демонструють, що банальна заміна багатомовного токенізатора на одномовний може значно покращити продуктивність майже на кожному завдання та мові [12].

1.4.3 Вплив представлення мови та розміру моделі

Ключовим фактором, що впливає на продуктивність, є рівень представлення мови в моделі. Мови, які добре репрезентовані у словниковому запасі (базовому тренувальному наборі) багатомовної моделі, часто демонструють незначне зниження точності порівняно з їх одномовними аналогами [12]. Проте одномовні моделі зазвичай показують кращі результати на мовах, які займають невеликий відсоток спільного корпусу багатомовної моделі. Збільшення розміру багатомовної моделі може допомогти вирішити цю проблему, дозволяючи моделі виділяти більше потужності на кожну мову [11]. Прикладом є модель Poro 34B з 34 мільярдами параметрів, навчена на фінській, англійській та мовах програмування, яка значно перевершує наявні одномовні моделі для менш репрезентованих мов, таких як фінська [13].

1.4.4 Вплив багатомовності на упередженість

Ще одним важливим фактором є вплив багатомовності на упередженість. Дослідження показують, що багатомовне донавчання іноді може посилювати упередження порівняно з одномовним донавчанням, хоча ефект не завжди є пов’язаним [14]. Це підкреслює необхідність ретельної оцінки та зменшення упереджень у багатомовних моделях.

1.4.5 Висновки

Зрештою, вибір між одномовними та багатомовними моделями залежить від конкретного завдання, доступних ресурсів та цільових мов. У деяких випадках, як з португальською мовою, обидва типи моделей можуть досягати порівнянних результатів, що робить вибір менш критичним [15]. Однак в інших випадках необхідно ретельно зважити компроміс між крос-лінгвістичним трансфером та спеціалізацією на конкретній мові. Порівняння результатів одномовних та багатомовних моделей для української мови може вказати, у яких напрямках дослідники повинні приділити більше уваги для розробки якісних україномовних наборів даних та моделей.

1.5 ОПМ української мови та виклики

Українська мова, що належить до східнослов'янської мовної сім'ї індоєвропейських мов [16], має багату морфологічну структуру, характерну для синтетичних мов [17]. З понад 40 мільйонами носіїв, вона є однією з найбільш поширених слов'янських мов [17]. Проте, незважаючи на широку поширеність, українська мова стикається з значними викликами в області обробки природної мови, перш за все через нестачу даних для тренування. Через це її класифікують як мову з низьким рівнем ресурсів [17]. У 2020 році українську мову визначили як “rising star” (Рис. 1.1) – мову, що розвивається з культурною спільнотою, але страждає від нестачі розмічених наборів даних [18].

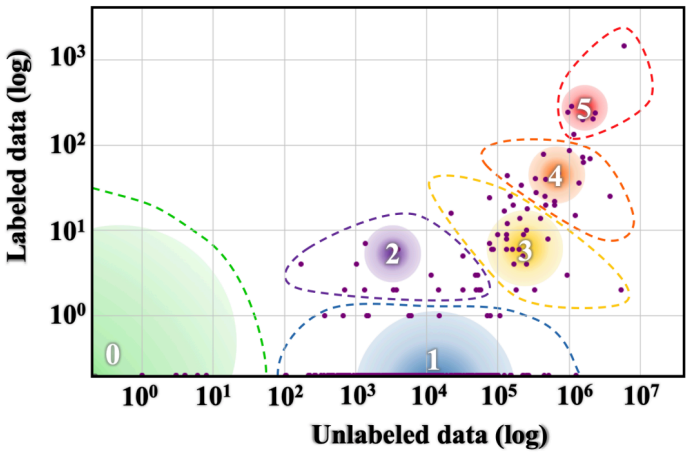

Рисунок 1.1 – Графік кластеризації мов за розподілом мовних ресурсів (кількість розмічених та нерозмічених даних). Українська мова знаходиться у кластері 3 – "rising stars". Взято з [19]

Основною перешкодою для ОПМ української мови є нестача загальнодоступних, орієнтованих на конкретні завдання корпусів [17]. Це історично заважало розвитку надійних одномовних моделей для української мови та змушувало дослідників покладатися на крос-мовне трансферне навчання з інших мов [20]. На відміну від мов з високим рівнем ресурсів, таких як англійська, яка має величезні набори даних та обчислювальні ресурси, українська мова не мала комплексних ресурсів, необхідних для просування досліджень та розробок у галузі ОПМ [20].

Хоча українська мова має певні лінгвістичні зв'язки з іншими східнослов'янськими мовами [16], пряме трансферне навчання з цих мов не завжди є оптимальним через відмінності у лексиці, граматиці та культурному контексті. Тому розробка спеціалізованих моделей та ресурсів ОПМ, адаптованих до унікальних особливостей української мови, є важливою для подолання цих викликів та розкриття повного потенціалу українських ОПМ-застосунків.

В теорії, навчання моделей на корпусі з декількох подібних мов теоретично може показати кращі результати, ніж навчання на одномовному корпусі, у випадку нестачі одномовних ресурсів. У цій роботі буде також перевірена

гіпотеза, що модель, натренована на корпусі слов'янських мов, може показати кращі результати, ніж одномовна модель.

1.6 Огляд робіт з порівняння одно- та багатомовних моделей

Декілька досліджень розглядали продуктивність одномовних та багатомовних моделей на основі BERT у різних завданнях обробки природної мови. Результати цих досліджень були неоднозначними: деякі віддають перевагу одномовним моделям, тоді як інші демонструють ефективність багатомовних моделей.

Наприклад, дослідження, зосереджені на конкретних мовах і мовних сім'ях, виявили переваги одномовних моделей. Virtanen та ін. (2019) [3] показали, що одномовна фінська модель BERT (FinBERT) перевершує багатомовну модель BERT (mBERT) у різних фінських завданнях ОПМ. Подібно, деякі дослідження виявили, що мовно-специфічні моделі BERT перевершують mBERT у розпізнаванні іменованих сутностей (NER) у слов'янських мовах. Velankar та ін. (2022) [21] підтвердили цю тенденцію, показавши, що одномовні моделі Marathi BERT перевершують багатомовні моделі у різних завданнях ОПМ для маратської мови (індоарійська мова, 94 мільйони носіїв).

Ефективність одномовних моделей також спостерігалася при переході у межах мовних родин. Torge та ін. (2023) [22] показали, що одномовні та натреновані на декількох мовах з однієї мовної родини моделі для західнослов'янських мов перевершують великі багатомовні моделі у низхідних (downstream) завданнях.

Однак, інші дослідження також показали ефективність багатомовних моделей. Feijo та Moreira (2020) [15] виявили, що багатомовна модель BERT незначно перевершує одномовні португальські моделі у різних завданнях NLP. Sido та ін. (2021) [10] також виявили, що багатомовна слов'янська модель BERT добре працює у кількох чеських завданнях ОПМ, хоча в деяких з них її перевершила одномовна чеська модель BERT.

Деякі дослідження вивчали ефекти доповнення даних та трансферного навчання. Yang та ін. (2023) [23] показали, що тримовна модель GPT-3, налаштована на англійських даних, може успішно виконувати інструкції угорською та китайською мовами, що свідчить про потенціал міжмовного трансферного навчання. Vīksna та Skadina (2021) [24] виявили, що не зважаючи на доповнення даних багатомовна модель BERT не покращила показники в розпізнаванні іменованих сутностей у шести слов'янських мовах.

В загальному, дослідження свідчать про те, що як одномовні, так і багатомовні моделі на основі BERT можуть бути ефективними для завдань ОПМ залежно від конкретної мови, завдання та доступних ресурсів. Хоча одномовні моделі часто перевершують багатомовні у завданнях своєї конкретної мови, багатомовні моделі можуть мати перевагу, коли ресурси для навчання одномовних моделей обмежені. Ця робота має на меті зробити внесок у поточну дискусію, порівнюючи продуктивність одномовних та багатомовних моделей на основі BERT для різних завдань ОПМ українською мовою.

2. ПОРІВНЯННЯ ОДНО ТА БАГАТОМОВНИХ МОДЕЛЕЙ НА ЗАВДАННЯХ УКРАЇНСЬКОЇ МОВИ

2.1 Бенчмарки для оцінки результатів

Оскільки метою роботи є порівняння результатів моделей на різних завданнях обробки природної мови, важливо визначити потенційні бенчмарки та завдання, які можна використати для такого порівняння.

Попри зростаючий інтерес до ОПМ для української, комплексних бенчмарків для оцінки мовних моделей досі бракує. І хоча для конкретних завдань існують готові набори даних, стандартизованого та повного бенчмарку, подібного до GLUE (Wang et al., 2018) [25] або SuperGLUE (Wang et al., 2019a) [26] для англійської чи KLEJ (Rybak et al., 2020) [27] для польської, який би покривав широке коло завдань, наразі немає.

Доступні комплексні бенчмарки:

- UA-datasets: Ця колекція містить набори даних для різних завдань, таких як розмітка частин мови (POS tagging) та класифікація новин (Panchenko et al., 2022) [28].
- Eval-UA-tion [18]: Представлений у Hamotskyi et al. (2024), цей набір бенчмарків включає дані для оцінки продуктивності мовних моделей у таких завданнях, як розуміння наративів оповідань (UA-CBT), асоціація статей з їх заголовками (UP-Titles) та елементарні мовні завдання (LMES).

Через відсутність готових комплексних бенчмарків для порівняння моделей, в цій роботі використано поєднання результатів на декількох окремих завданнях. Серед них розпізнавання іменованих сутностей, розрізнення значень слів, класифікація текстів, завдання бенчмарку Eval-UA-tion, заповнення пропусків та розмічення частин мови.

2.2 Розпізнавання іменованих сутностей (NER)

Розпізнавання іменованих сутностей є полягає у виявленні та класифікації імен, назв організацій, місць, топографічних назв, тощо у текстах.

Для української мови дослідники нещодавно випустили оновлений корпус NER-UK 2.0 **[29]**, який містить 323 200 слів і 21 993 іменовані сутності. У дослідженні автори також долучили початкові результати для моделі roberta-large [30]. У Таблиці 2.2.1 наведено порівняння результатів різних моделей на попередній версії корпусу NER-UK 1.0 [31].

Тип та мова моделі	**Розмір моделі**	**NER-UK 1.0 F1 оцінка**
RoBERTa WECHSEL Ukrainian	base	90.81 (1.51)
RoBERTa WECHSEL Ukrainian	large	91.24 (1.16)
RoBERTa Scratch Ukrainian	base	89.57 (1.01)
RoBERTa Scratch Ukrainian	large	89.96 (0.89)
ELECTRA Ukrainian	base	90.43 (1.29)
XLM RoBERTa Multilingual	base	90.86 (0.81)
XLM RoBERTa Multilingual	large	90.16 (2.98)
LiBERTa Ukrainian (pretrained)	large	**91.27 (1.22)**
RoBERTa Slavic	large	90.89
uk_ner_web_trf_13class RoBERTa	large	91.3 (?)

Таблиця 2.2.1 – Результати оцінки моделей на NER-UK 1.0 (F1 оцінка). Об'єднані результати з [20], [30], [32] та власного дослідження. Значення в дужках позначають стандартне відхилення метрики.

Таблиця 2.2.1 містить результати моделі, натренованої для української мови методом WECHSEL [33], багатомовної моделі XLM-R та нещодавно опублікованої моделі LiBERTa [20]. Остання була натренована з «нуля» на наборах даних української мови. Порівння показує, що моделі, адаптовані для української мови, ефективніші за багатомовні в розпізнаванні іменованих сутностей.

Як зазначено в [20], цікаво, що друга велика модель XLM-R демонструє гірші результати, ніж усі базові моделі. Вона також має найбільшу варіативність результатів. Це підкреслює необхідність тренування моделей, специфічних для

певної мови, оскільки як WECHSEL-RoBERTa, так і LiBERTa мають меншу варіативність результатів.

Таблиця також включає результати моделі на корпусі слов'янських мов (болгарська, чеська, польська, словенська, російська та українська) [34]. Ця модель була додатково дотренована на частині набору даних у ході дослідження. Можна побачити, що результати моделі, натренованої на слов'янських мовах, кращі, ніж у багатомовних моделей, але все ж гірші, ніж у моделей, адаптованих для української мови.

На бенчмарку WikiANN [35], який також включає розпізнавання іменованих сутностей, LiBERTa, яка повністю натренована на українській мові, показує гірші результати. Результати порівняння моделей наведені у Таблиці 2.2.2.

Тип та мова моделі	**Розмір моделі**	**WikiANN micro f1**
RoBERTa WECHSEL Ukrainian	base	92.98 (0.12)
RoBERTa WECHSEL Ukrainian	large	**93.22 (0.17)**
ELECTRA Ukrainian	base	92.99 (0.11)
XLM RoBERTa Multilingual	base	92.27 (0.09)
XLM RoBERTa Multilingual	large	92.92 (0.19)
LiBERTa Ukrainian (pretrained)	large	92.50 (0.07)

Таблиця 2.2.2 – Результати оцінки моделей на WikiANN. Значення взято з [20]. Значення в дужках позначають стандартне відхилення метрики.

Автори зазначають, що такі результати можуть бути спричинені різницею між датасетами та навчальними даними для моделі [20]. Модель і токенізатор були натреновані переважно на українських текстах, тоді як багатомовні моделі часто тренуються на даних зібраних з Wikipedia [20].

Отже, у завданні NER моделі, адаптовані для української мови, показують вищу продуктивність у порівнянні з багатомовними моделями. Для цього завдання створення та використання моделей, орієнтованих на українську мову, важливе для досягнення кращих результатів. Так WECHSEL-RoBERTa та

LiBERTa показують високу точність і невелику варіативність. Проте, різниця між результатами моделі, навченої на українських даних, та RoBERTa, навченої за методом WECHSEL, незначна. Це може означати, що попереднє тренування на українських даних не дало значної переваги порівняно з методом WECHSEL.

2.3 Розрізнення значень слів (Word Sense Disambiguation)

Розрізнення значень слів (WSD) допомагає вирішити проблему неоднозначності мови (наявність омонімів «ключ журавлів» - «ключ від дверей»). Оскільки багато слів мають кілька значень, натренована модель намагається визначити правильний сенс слова в залежності від контексту. Рішення цього завдання може бути частиною ширшого конвеєру з моделей, ціль якого переказати текст чи відповісти на питання про текст.

У дослідженні [36] автори порівнюють точність розрізнення значень слів (WSD) для української мови, використовуючи різні стратегії об'єднання та попередньо навчені моделі без додаткового тренування. Таблиця 2.3.1 демонструє, що модель paraphrasemultilingual-mpnet-base-v2 (PMMBv2) з [37] показує кращі результати WSD, ніж модель RoBERTa, попередньо навчена на українській мові в 2020 році [38]. У своїй роботі [36] автори покращили продуктивність PMMBv2 до 0.779, тонко налаштувавши її на основі датасету зі словника СУМ (Словник Української Мови).

Тип та мова моделі	**Розмір моделі**	**Середня точність WSD**
mBERT cased multilingual	base	0.602
XLM RoBERTa Multilingual	base	0.529
XLM RoBERTa Multilingual	large	0.547
XLM RoBERTa Ukrainian	base	0.528
RoBERTa Ukrainian	large	0.580
Parahprase MPNET Multilingual	base	**0.735**

Таблиця 2.3.1 – Середня точність розрізнення значень слів без тонкого налаштування. Взято з [36]

Втім, щоб отримати релевантне порівняння між одномовними та багатомовними моделями в задачі WSD, необхідно протестувати сучаснішу українську модель, таку як LiBERTa [20], з додатковим налаштуванням на тому ж наборі даних.

2.4 Класифікація текстів

У дослідженні [20] автори представили модель LiBERTa, яка була попередньо навчена з нуля на українських даних. Вони надають результати порівняння її продуктивності на завданні класифікації текстів з Ukranian News Classification benchmark [28]. Цей бенчмарк показує, наскільки точно модель може визначити, до якого новинного джерела належить певна стаття (багатокласова класифікація).

Як видно з Таблиці 2.4.1, модель, попередньо навчена на українських даних, перевершує багатомовну модель XLM-R за продуктивністю. Однак, модель, адаптована до української мови за допомогою методу WECHSEL [33], показує ще кращі результати.

Тип та мова моделі	**Розмір моделі**	**Класифікація новин macro F1**
RoBERTa WECHSEL Ukrainian	large	**96.48 (0.09)**
XLM RoBERTa Multilingual	large	95.13 (0.49)
LiBERTa Ukrainian (pretrained)	large	95.44 (0.04)

Таблиця 2.4.1 – Порівняння результатів моделей на задачі класифікації українських новин. Значення в дужках позначають стандартне відхилення метрики. Взято з [20]

2.5 Eval-UA-tion

Eval-UA-tion [18] — це спеціальний бенчмарк, розроблений для оцінки продуктивності мовних моделей на українській мові. Він складається з трьох основних завдань:

- UA-CBT: Це завдання створене за аналогією з тестом Children's Book Test. Воно передбачає заповнення пропусків у розповідях правильними словами з наданого набору варіантів. Завдання оцінює здатність моделі розуміти текст та передбачати пропущені слова на основі контексту.
- UP-Titles: У цьому завданні потрібно зіставити статті з онлайн-газети «Українська Правда» з їх правильними заголовками з наданого набору схожих варіантів. Це оцінює здатність моделі зрозуміти основну ідею статті та відрізняти її від схожих тем.
- LMentry-static-UA (LMES): це набір завдань, розроблений на основі бенчмарку LMentry, спеціально для оцінки можливостей мовних моделей. Ці завдання, хоч і прості для людей, часто є складними для штучного інтелекту - наприклад, визначення найдовшого слова з пари або виділення N-того слова у реченні.

Eval-UA-tion містить базові результати для кожного завдання: показник людських відповідей та випадкової генерації. Мета Eval-UA-tion полягає у всебічній оцінці можливостей мовних моделей для української мови та стимулюванні подальших досліджень у цій сфері.

Eval-UA-tion був представлений на конференції UNLP 2024 і включає оцінки для різних багатомовних та одномовних моделей. Незважаючи на використання більш сучасних моделей замість BERT, порівнянню яких присвячена ця робота, автори показують результати, що є корисними для порівняння продуктивності багатомовних та одномовних моделей.

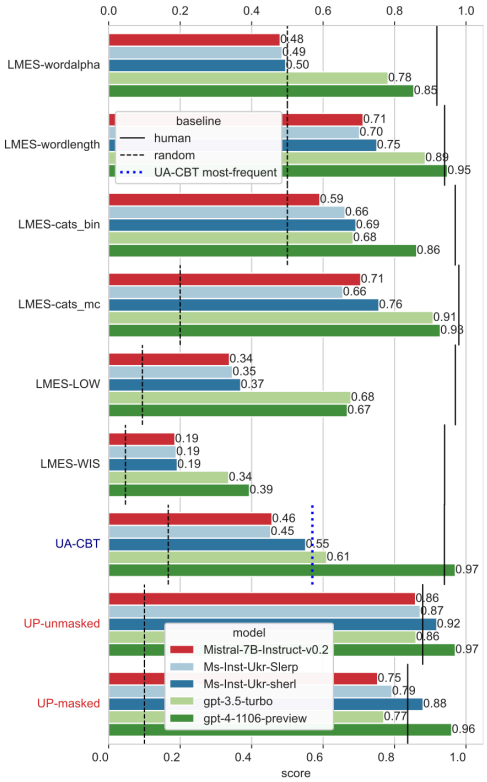

Рисунок 2.5.1 – Результати порівняння моделей на завданнях Eval-UA-tion взяті з [18].

Згідно з результатами, представленими на Рисунку 2.5.1, модель SherlockAssistant/Mistral-7B-Instruct-Ukrainian, яка пройшла попереднє налаштування на українських даних, виявилася більш ефективною для завдання UP-Titles [18], ніж багатомовні моделі, що не є OpenAI, та навіть перевершила GPT-3. Хоча GPT-4, багатомовна модель, показала найкращі результати, слід зазначити, що пряме порівняння ускладнюється через значні відмінності у розмірі моделі та обсязі набору даних для попереднього навчання у порівнянні з іншими моделями.

2.6 Заповнення пропусків (Mask filling)

На етапі попереднього навчання моделей на основі BERT, як правило, їх тренують передбачати випадково замасковані слова у великому корпусі даних без розмітки. Це допомагає мовній моделі вивчати фундаментальні представлення мови або мовні структури, такі як порядок слів у реченнях і

значення слів (через навчання токенізатора та перетворення векторів). Ці попередньо навчені моделі потім донавчаються на специфічних наборах даних для конкретних завдань. Тому оцінювання продуктивності моделі на завданні заповнення пропусків потенційно може показати, наскільки добре модель «розуміє» українську мову.

Для оцінювання моделей створено набір речень із замаскованими словами, що вимагають не лише знання української мови, але й «розуміння» українського контексту (культура, поезія, історія, кухня, традиції). Якість заповнення пропусків різними моделями дозволяє визначити якість даних, використаних під час навчання. Наприклад, відсутність оригінальних українських текстів може призвести до гірших результатів у реченнях, пов'язаних з українською культурою.

Для дослідження якості заповнення пропусків ми створили 97 питань, доступних у Додатку 1. Питання генерувалися моделлю Gemini, а також дописувалися та перевірялися вручну. Для заповнення пропусків на кожному питанні використовували кілька моделей. Оцінювані моделі включали багатомовні моделі, моделі, які були додатково навчені на українських текстах, і одномовні моделі. Проводили два типи оцінювання: ручне оцінювання для висновків і порівняння продуктивності, та загальне оцінювання.

Загальне оцінювання проводили за допомогою вебсайту, де респонденти бачили випадкове запитання з пропуском і відповіді від кількох моделей (інформація про назву моделі, яка згенерувала конкренту відповідь була прихована). Вигляд інтерфейсу вебсайту доступний у Додатку 2. Учасникам опитування пропонували оцінити кожну відповідь за шкалою від 1 до 5 за такими критеріями:

5: Точне та контекстно релевантне заповнення пропуску.

4: Переважно точне, але з незначними помилками.

3: Деяка релевантність, але не зовсім точне.

2: Погана релевантність і точність.

1: Повністю нерелевантне або граматично неправильне.

Оскільки метою опитування була оцінка, наскільки якісно заповнені пропуски моделями, а не результатам якої моделі нададуть перевагу люди (Human Feedback), для опитування було запрошено 5 україномовних людей. Для всіх учасників українська мова є основною мовою спілкування. Всі учасники мають вищу освіту.

Оцінювані моделі включали багатомовний XLM-RoBERTa (базовий і великий) [6], GPT4o (як еталон), RoBERTa WECHSEL український (базовий і великий) [32], та YouScan RoBERTa (базовий, попередньо навчений на українському наборі даних) [38]. Результати ручного та загального оцінювання наведено нижче.

Результати ручного оцінювання:

Багатомовні моделі зазвичай показують нижче знання українського контексту (Таблиця 2.6.1).

Видатні українські письменники, такі як [MASK] і [MASK], зробили великий внесок у літературу. *(питання згенероване ШІ)*	
Назва моделі	Результат заповнення пропусків
XLM RoBERTa base	Видатні українські письменники, такі як **Шевченко** і **Шевченка**, зробили великий внесок у літературу.
RoBERTa WECHSEL base	Видатні українські письменники, такі як **Франко** і **Шевченко**, зробили великий внесок у літературу.
YouScan RoBERTa Uk base	Видатні українські письменники, такі як **Шевченко** і **інші**, зробили великий внесок у літературу.

Таблиця 2.6.1 – Приклад результатів заповнення пропусків різними моделями.

Моделі, попередньо навчені або перенесені на українську, часто показують краще знання українського контексту (наприкал, у Таблиці 2.6.2 модель WECHSEL додала нішеві українські топоніми).

Українські Карпати славляться своїми курортами, такими як [MASK] і [MASK], де можна насолоджуватися [MASK]. *(питання згенероване ШІ)*	
Назва моделі	Результат заповнення пропусків
RoBERTa WECHSEL large	Українські Карпати славляться своїми курортами, такими як Трускавець і Яремче, де можна насолоджуватися природою.

Таблиця 2.6.2 – Приклад результату заповнення пропуску моделлю WECHSEL.

Моделі, попередньо навчені на українських даних, також зазвичай генерують більш несподівані відповіді порівняно з іншими моделями. Вони також більш послідовні з закінченнями слів, що є важким завданням для багатомовних моделей, оскільки українська мова є морфологічно багатою.

Однак всі моделі послідовно показують брак якості міркувань і знання українського культурного контексту порівняно з набагато більшою та зрілою моделлю GPT4o. Відповіді схильні бути більш загальними, особливо для запитань, пов'язаних з науковими фактами. (Таблиця 2.6.3, Таблиця 2.6.4).

Софія Київська - це [MASK], яка була побудована у [MASK] році і є прикладом [MASK] архітектури. *(питання згенероване ШІ)*	
Назва моделі	Результат заповнення пропусків
GPT4	Софія Київська - це **собор**, яка була побудована у **1037** році і є прикладом **візантійської** архітектури.
XLM RoBERTa base	Софія Київська - це **церква**, яка була побудована у **1905** році і є прикладом **сучасної** архітектури.
RoBERTa WECHSEL base	Софія Київська - це **церква**, яка була побудована у **2007** році і є прикладом **української** архітектури.
RoBERTa WECHSEL large	Софія Київська - це **будівля**, яка була побудована у **1991** році і є прикладом **сучасної** архітектури.
YouScan RoBERTa Uk base	Софія Київська - це **церква**, яка була побудована у **2001** році і є прикладом **народної** архітектури.

XLM RoBERTa large	Софія Київська - це **церква**, яка була побудована у **1929** році і є прикладом **української** архітектури.

Таблиця 2.6.3 – Приклад результатів заповнення пропусків, які вимагають знання фактів, різними моделями.

Людська пам'ять функціонує завдяки [MASK] і [MASK], що дозволяє запам'ятовувати [MASK]. *(питання згенероване ШІ)*	
Назва моделі	Результат заповнення пропусків
GPT4	Людська пам'ять функціонує завдяки **мозковій діяльності** і **нейронним зв'язкам**, що дозволяє запам'ятовувати **інформацію**.
XLM RoBERTa base	Людська пам'ять функціонує завдяки **енергії** і **GPS**, що дозволяє запам'ятовувати **інформацію**.
RoBERTa WECHSEL base	Людська пам'ять функціонує завдяки **мозку** і **відео**, що дозволяє запам'ятовувати **інформацію**.
RoBERTa WECHSEL large	Людська пам'ять функціонує завдяки **науці** і **розуму**, що дозволяє запам'ятовувати **людей**.
YouScan RoBERTa Uk base	Людська пам'ять функціонує завдяки **технології** і **енергії**, що дозволяє запам'ятовувати **інформацію**.
XLM RoBERTa large	Людська пам'ять функціонує завдяки **руху** і **часу**, що дозволяє запам'ятовувати **інформацію**.

Таблиця 2.6.4. Приклад результатів заповнення пропусків, які вимагають знання фактів про історію України чи біологію, згенеровані різними моделями.

Це ще раз підкреслює значну перевагу великих комерційних моделей над меншими безкоштовними моделями з відкритим кодом. Така різниця у продуктивності, ймовірно, пов’язана з тим, що комерційні моделі мають більшу кількість параметрів та навчаються на значно ширших наборах даних.

Результати загального оцінювання:

Під час загального оцінювання всі відповіді моделей оцінювали від 5 до 1. GPT4o використовували як еталон. Результати загального оцінювання на Рисунку 2.6.5 показують, що моделі, перенесені на українську за допомогою методу

WECHSEL, перевершують модель YouScan RoBERTa, попередньо навчену на українській. Однак найнижчі бали отримали багатомовні моделі XLM-R.

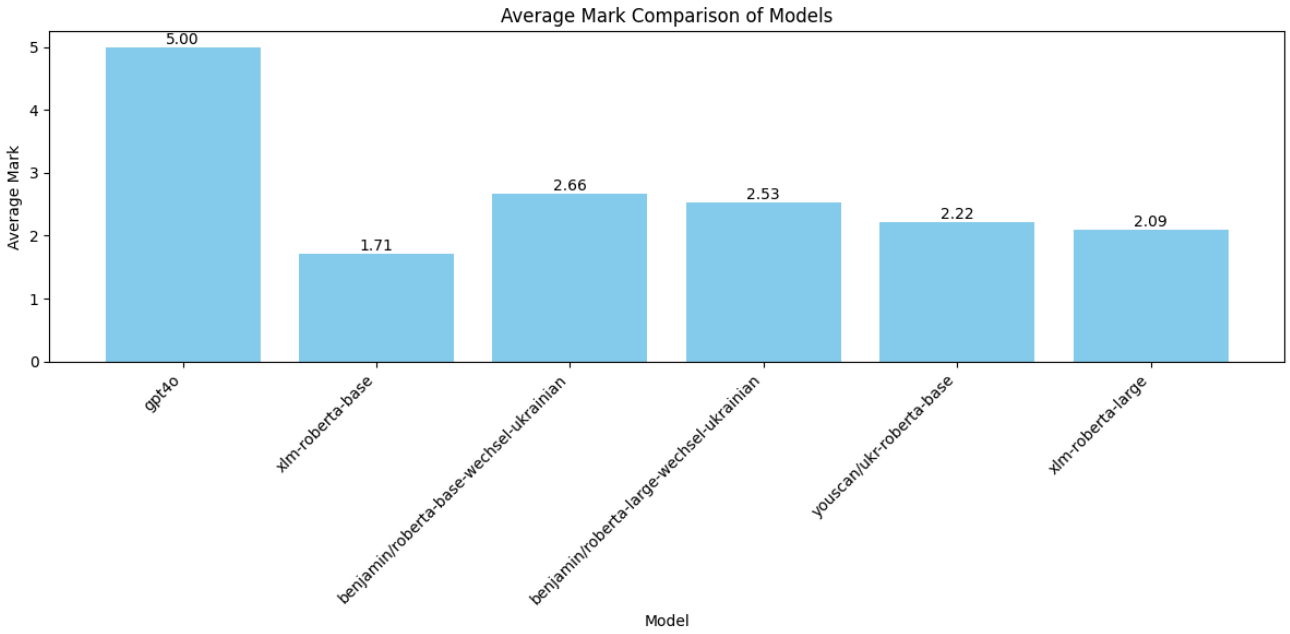

Рисунок 2.6.5 – Результати порівняння середнього балу моделей на завданні заповнення пропусків. Можливі оцінки в діапазоні 1-5

Водночас жодна модель не досягла якості комерційної моделі GPT4o, як показано на Рисунку 2.6.6.

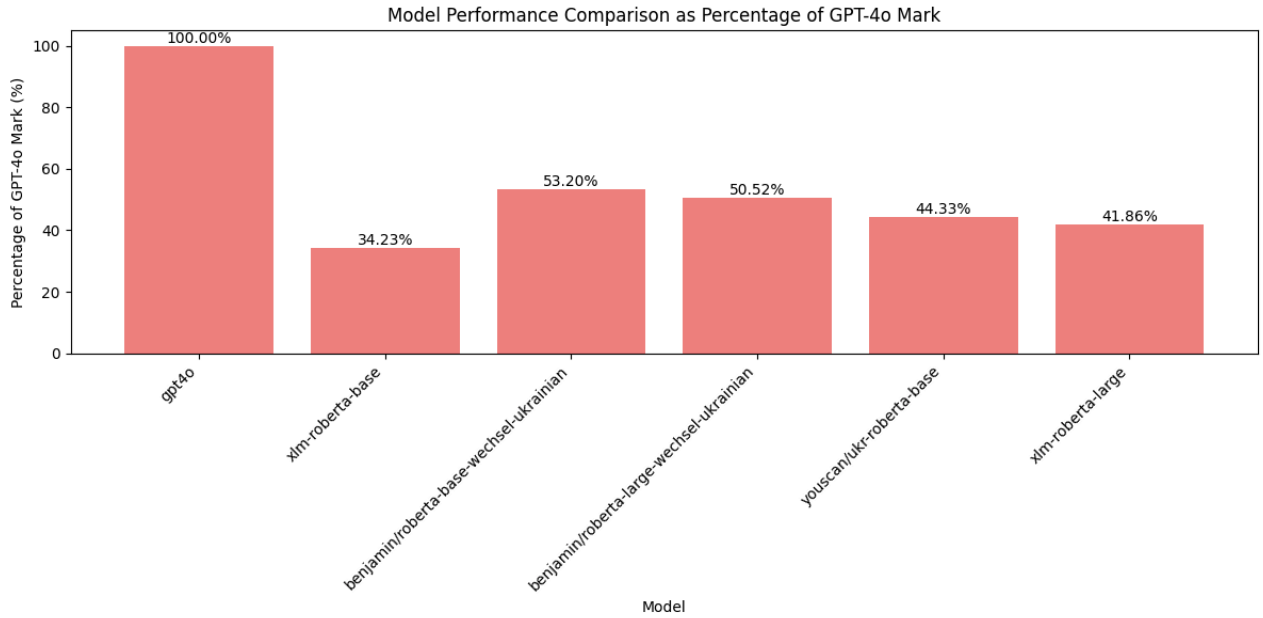

Рисунок 2.6.6 – Результати порівняння моделей з рівнем GPT4 на завданні заповнення пропусків від 0 до 100%.

Для завдання заповнення пропусків моделі, попередньо навчені або перенесені на українську, зазвичай працюють краще, ніж багатомовні моделі. Проте, адаптовані до української мови за допомогою методу WECHSEL багатомовні моделі, несподівано перевершують моделі, попередньо навчені виключно на українських текстах. Можливим поясненням цього може бути те, що моделі WECHSEL отримують перевагу від глибшого розуміння мови та величезних масивів даних, що використовуються для навчання базових моделей. Однак залишається незрозумілим, чи покажуть вони вищі результати, ніж остання українська модель LiBERTa, друга версія якої показала кращу продуктивність, ніж WECHSEL на інших завданнях [41].

2.7 Розмічування частин мови (POS tagging)

У [20] автори порівнюють продуктивність різних багатомовних та одномовних моделей на українській частині набору даних Universal Dependencies [39]. Розмічування частин мови (POS tagging) — це процес призначення граматичної категорії (або "тегу") кожному слову в тексті. Ці теги вказують на частину мови слова, наприклад, іменник, дієслово, прикметник, прислівник, займенник, прийменник, сполучник або вигук.

Згідно з [20], продуктивність усіх моделей у цьому завданні є досить високою та схожою, що може свідчити про те, що завдання POS tagging є відносно простим для української мови, а різниця в результатах може бути випадковою. Результати порівняння наведені у Таблиці 2.7.1.

Тип та мова моделі	**Розмір моделі**	**UD POS (точність)**
RoBERTa WECHSEL Ukrainian	base	98.57 (0.03)
RoBERTa WECHSEL Ukrainian	large	98.74 (0.06)
ELECTRA Ukrainian	base	98.59 (0.06)
XLM RoBERTa Multilingual	base	98.45 (0.07)
XLM RoBERTa Multilingual	large	**98.71 (0.04)**
LiBERTa Ukrainian (pretrained)	large	98.62 (0.08)

Таблиця 2.7.1 – Результати моделей на задачі розмічування частин мови, взято з [20].

Таким чином, це порівняння показує, що для розмічування частин мови різниця у продуктивності між одномовними та багатомовними моделями є мінімальною.

2.8 Висновки

Результати порівняння одно- та багатомовних моделей наведено в Таблиці 2.8.1.

Завдання ОПМ для української мови	**Порівняння одно- та багатомовних моделей**	**Додаткові коментарі**
Розпізнавання іменованих сутностей (NER)	Одномовна модель тренована в основному на українських текстах показує найкращий результат	
Розрізнення значень слів (Word Sense Disambiguation)	Багатомовна модель дотренована українською показує кращий результат	Є потреба порівняти з результатом більш сучасної української моделі LiBERTa v2
Класифікація текстів	Модель адаптована до української методом WECHSEL показує кращі результати, ніж українська LiBERTa	Друга версія моделі LiBERTa перевершує WECHSEL, але результати ще не опубліковані
UA-CBT (частина Eval-UA-tion)	Модель Mistral-7B-Sherlock дотренована на українських даних показує кращі результати ніж багатомовна модель.	Mistral-7B-Sherlock не досягає рівня gpt-4-1106-preview.
UP-Titles		
LMentry-static-UA (LMES):		
Заповнення пропусків	Модель адаптована до української методом WECHSEL показує кращі результати, ніж українська YouScan/RoBERTa	Жодна модель не досягає рівня GPT4.
Розмічування частин мови (POS Tagging)	Незначна різниця у результатах одно- та багатомовних моделей	Можливо пов’язано з легкістю задачі

Таблиця 2.8.1 – Порівняння одно- та багатомовних моделей на українських завданнях NLP

В загальному, адаптовані до української одномовні моделі часто демонструють кращі результати у спеціалізованих завданнях. Прикладом цього є показники порівняння моделей для розпізнавання іменованих сутностей та класифікації текстів. Це часто пов’язано з тим, що моделі, спеціально треновані на українських даних, краще розуміють контекст та особливості мови, а також орієнтуються в контексті української культури, від чого можуть страждати багатомовні моделі.

Водночас багатомовні моделі, дотреновані українською мовою, також можуть бути ефективними для завдань, які не потребують глибокого розуміння специфіки мови. Прикладом таких задач є розрізнення значень. Потенційно різниця між моделями натренованими з «нуля» українською та дотренованими українською може зменшитись, в залежності від якості та кількості тренувальних даних. Для деяких простих задач, як розмічування частин мови, різниця між різними типами моделей незначна. У практиці для таких задач вибір правильної моделі може не особливо впливати на ефективність.

Одномовна модель, попередньо навчена на українських даних, хоча і не перевершує WECHSEL модель для української мови, вже показує кращі результати, ніж попередні повністю українські моделі. Це може свідчити про те, що час, коли модель, натренована лише на українських даних, показуватиме кращі результати, ніж адаптована модель, вже близько. (Згідно з виступом автора [20] на конференції UNLP, друга версія моделі LiBERTa вже показує кращі результати, ніж WECHSEL модель.)

Отже, вибір між одномовними та багатомовними моделями залежить від конкретного завдання: для більш спеціалізованих і складних задач доцільніше використовувати одномовні моделі, тоді як для загальніших задач багатомовні можуть бути аналогічно ефективними. Ця різниця потенційно може нівелюватися більше, якщо моделі були адаптовані до української мови.

3. ОБМЕЖЕННЯ ДОСЛІДЖЕННЯ ТА МАЙБУТНЯ РОБОТА

3.1 Обмеження дослідження

Незважаючи на отримані результати, важливо відзначити певні обмеження цього дослідження, які слід враховувати при інтерпретації та подальшому використанні висновків.

3.1.1 Обмеженість обчислювальних ресурсів та обсягів даних

Повноцінне навчання «з нуля» ВММ, таких як BERT, потребує значних обчислювальних потужностей та величезних обсягів даних. Попередня оцінка вартості та часу навчання моделі (64 GPU протягом 4 діб за ціною $4.5/год) складає приблизно $7000. Окрім цього, доступні корпуси українських текстів обмежені (оригінальний набір даних BERT містить 3.3 мільярда слів). Через ці обмеження в цьому дослідженні переважно використовували тонке налаштуванню вже навчених моделей. Хоча воно і дозволяє досягти непоганих результатів з меншими витратами грошей та часових ресурсів, це не дозволило повністю оцінити потенціал моделі, оскільки вона вже адаптована до певних даних на етапі базового тренування та містить упередження стомсовно цих даних.

Для кращого порівняння роботи моделей потрібно мати ресурси для тренування моделі на корпусі слов'янських мов. До виходу LiBERTa [20] існувала потреба також навчити з «нуля» україномовну модель, що попри спроби, не вдалось зробити у межах поточного дослідження через брак обчислювальних ресурсів.

3.1.2 Відсутність комплексного бенчмарку

На відміну від англійської мови, для якої існує стандартизований набір бенчмарків GLUE, що охоплює різні задачі ОПМ, для української мови такого інструменту бракує. Це суттево ускладнює порівняння різних моделей та підходів, оскільки доводиться використовувати окремі набори даних та метрики, які можуть бути не повністю репрезентативними для всіх аспектів мови.

Відсутність універсального стандарту оцінки робить порівняння різних моделей менш об’єктивним та ускладнює процес перевірки для кожного завдання.

3.1.3 Використання результатів конференції UNLP 2024

Конференція UNLP 2024, що відбулася 25 травня 2024 року, стала важливою подією для розвитку ОПМ в Україні. На конференції представили нові моделі, натреновані «з нуля» чи дотреновані на українських текстах [20, 40]. Крім того, конференція відзначилася розширенням бенчмарків для оцінки моделей [18, 29]. У рамках цієї роботи ми здійснили спробу інтегрувати деякі з цих результатів, проте обмежений час не дозволив більш ґрунтовно використати нові моделі та бенчмарки.

3.2 Майбутня робота та напрямки досліджень

Результати дослідження показали, що спеціалізовані україномовні моделі мають значні переваги у більшості завдань з обробки природної мови для української. Однак, з'явилася нагальна потреба в створенні комплексного україномовного бенчмарку, аналогічного до GLUE або KLEJ, для всебічної оцінки моделей за різними завданнями. Ця потреба стає ще актуальнішою з огляду на активний розвиток моделей, натренованих переважно на українських даних [20], та появу невеликих комплексних бенчмарків [18].

Напрямки майбутньої роботи:

- **Розробка комплексного бенчмарку:** Основним пріоритетом є створення великого комплексного бенчмарку, який буде українським аналогом GLUE. Він має включати широкий спектр завдань ОПМ, таких як розуміння тексту, аналіз настроїв, відповіді на запитання, перефразування та інші.
- **Завдання з глибоким розумінням:** Акцент буде зроблено на завданнях, які вимагають від моделей не просто обробки тексту, а й глибокого

розуміння української мови, її культурних, історичних та соціальних особливостей. Такий підхід дозволить оцінити здатність моделей вирішувати складні завдання, враховуючи специфіку українського середовища.

- **Розширення набору даних:** Важливим кроком є розширення та диверсифікація набору даних, що використовуються для навчання та оцінки моделей. Це передбачає збір текстів різних жанрів, стилів, тематик, а також створення спеціалізованих корпусів для конкретних завдань. Розширення бази даних сприятиме підвищенню точності та адекватності моделей.
- **Врахування етичних аспектів:** При розробці бенчмарку та моделей необхідно приділяти значну увагу етичним аспектам, таким як упередженість, дискримінація та дезінформація. Важливо забезпечити, щоб моделі були справедливими, прозорими та не завдавали шкоди суспільству. Це передбачає ретельний аналіз даних, алгоритмів та результатів роботи моделей, а також розробку механізмів контролю та корекції.

Очікувані результати:

Створення комплексного україномовного бенчмарку сприятиме:

- **Покращенню якості моделей:** Забезпечить більш точну та всебічну оцінку продуктивності моделей, що дозволить виявити їх сильні та слабкі сторони.
- **Стимулюванню досліджень:** Підтримає активізацію досліджень у галузі ОПМ для української мови та сприятиме розробці нових, більш ефективних моделей.

Загалом, реалізація запропонованих напрямків досліджень має значний потенціал для розвитку української ОПМ-спільноти та створення потужних інструментів для вирішення різноманітних завдань у різних сферах життя.

ВИСНОВКИ

В рамках магістерської роботи вдалося порівняти результати одно та багатомовних моделей типу BERT на завданнях обробки української мови. Результати порівняння демонструють те, що моделі, натреновані “з нуля” на українських текстах або дотреновані на них, в загальному мають кращі результати, ніж багатомовні. Це вдалося показати на завданнях розпізнавання іменованих сутностей, класифікації текстів та заповнення пропусків. У той же час, багатомовні моделі показують наближений або кращий результат на більш легких завданнях (розрізнення значень слів).

Українські моделі показали кращі можливості захоплювати український контекст та враховувати культурні нюанси, які можуть втрачатися у багатомовних моделей. Хоча модель LiBERTa натренована “з нуля” на українських наборах даних показала гірші результати, ніж просто перенесена на українську методом WECHSEL багатомовна модель - згідно з виступом авторів LiBERTa [41], друга модель вже перевершує WECHSEL-RoBERTa. Це може позначати, що кількість нерозмічених якісних даних українською доходить до того моменту, коли треновані лише на них моделі вже можуть показати кращі результати, ніж адаптовані багатомовні.

В загальному, дослідження показало, що вибір між одно та багатомовними моделями сильно залежить від завдання. На простих завданнях різниця між ефективністю моделей незначна, але на завданнях, які потребують кращого знання мови чи культурного контексту українські моделі будуть більш продуктивними.

У ході роботи були зібрані порівняння роботи одно та багатомовних моделей для різних мов, що додатково підкреслило важливість проведення окремого порівняння для української. Ефективність певного підходу залежить від мови, завдання та доступних ресурсів. Наприклад, одномовна фінська модель

BERT перевершила багатомовну mBERT, але результати порівняння для португальської мови були кардинально протилежними.

Для порівняння одно та багатомовних моделей були зібрані результати з різних джерел, включаючи наукові статті та моделі. Деякі з використаних моделей були додатково налаштовані. Для завдань, де потрібно було заповнити пропуски, ми використовували готові претреновані моделі.

Більшість моделей, використаних у дослідженні, не досягають рівня комерційних моделей, як ChatGPT4 через обмеженість кількості параметрів моделі, обсягу даних для тренування та можливостей отримувати зворотній зв’язок щодо роботи моделі.

Процес порівняння моделей значно ускладнювала відсутність комплексного бенчмарку для української мови, який був би аналогом GLUE або KLEJ. Розробці такого бенчмарку планується присвятити майбутню роботу.

Результати цього дослідження можуть допомогти у визначенні оптимальної моделі для завдань, які були представлені у порівнянні. Крім того, зібрані показники порівняння моделей та перелік завдань можуть слугувати основою для розробки комплексного бенчмарку для української мови. Загалом, результати роботи також підкреслюють важливість розробки та вдосконалення одномовних моделей для української мови, що дозволить забезпечити їхню ефективність у широкому спектрі завдань обробки мови.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. BERT: pre-training of deep bidirectional transformers for language understanding / J. Devlin et al. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) / ed. by J. Burstein, C. Doran, T. Solorio. Minneapolis, Minnesota, 2019. P. 4171–4186. URL: https://aclanthology.org/N19-1423 (дата звернення: 23.05.2024).
2. Attention is all you need / A. Vaswani et al. Proceedings of the 31st international conference on neural information processing systems. Red Hook, NY, USA, 2017. P. 6000–6010.
3. Multilingual is not enough: BERT for Finnish / A. Virtanen та ін. 2019. URL: https://doi.org/10.48550/arXiv.1912.07076 (дата звернення: 28.05.2024).
4. RoBERTa: A robustly optimized BERT pretraining approach / Y. Liu et al. CoRR. 2019. Abs/1907.11692. URL: http://arxiv.org/abs/1907.11692.
5. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter / V. Sanh et al. ArXiv. 2019. Abs/1910.01108. URL: https://api.semanticscholar.org/CorpusID:203626972.
6. Unsupervised cross-lingual representation learning at scale / A. Conneau et al. 2020.
7. Larger-Scale transformers for multilingual masked language modeling / N. Goyal et al. 2021.
8. Multilingual instruction tuning with just a pinch of multilinguality / U. Shaham et al. 2024. URL: https://doi.org/10.48550/arXiv.2401.01854 (дата звернення: 28.05.2024).
9. Pires T., Schlinger E., Garrette D. How multilingual is multilingual BERT?. Proceedings of the 57th annual meeting of the association for computational linguistics, Florence / ed. by A. Korhonen, D. Traum, L. Màrquez. 2019. P. 4996–5001. URL: https://aclanthology.org/P19-1493 (дата звернення: 28.05.2024).
10. Czert - czech bert-like model for language representation / J. Sido et al. 2021. URL: https://doi.org/10.48550/arXiv.2103.13031 (дата звернення: 28.05.2024).
11. Ruder S. The state of multilingual AI. ruder.io. URL: https://www.ruder.io/state-of-multilingual-ai/ (дата звернення: 23.05.2024).
12. How good is your tokenizer? On the monolingual performance of multilingual language models / P. Rust et al. Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers) / ed. by C. Zong et al. 2021. P. 3118–3135. URL: https://aclanthology.org/2021.acl-long.243 (дата звернення: 28.05.2024).

13. Poro 34B and the blessing of multilinguality / R. Luukkonen et al. 2024. URL: https://doi.org/10.48550/arXiv.2404.01856 (дата звернення: 28.05.2024).
14. Comparing biases and the impact of multilingual training across multiple languages / S. Levy et al. Association for computational linguistics. 2023. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. P. 10260–10280. URL: https://aclanthology.org/2023.emnlp-main.634 (дата звернення: 23.05.2024).
15. Feijo D. d. V., Moreira V. P. Mono vs multilingual transformer-based models: a comparison across several language tasks. 2020. URL: https://doi.org/10.48550/arXiv.2007.09757 (дата звернення: 28.05.2024).
16. Gomez F. P., Rozovskaya A., Roth D. A Low-Resource Approach to the Grammatical Error Correction of Ukrainian. Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), Dubrovnik / ed. by M. Romanyshyn. P. 114–120. URL: https://aclanthology.org/2023.unlp-1.14 (дата звернення: 28.05.2024).
17. Chaplynskyi D. Introducing ubertext 2.0: a corpus of modern Ukrainian at scale. Proceedings of the second ukrainian natural language processing workshop (UNLP), Dubrovnik / ed. by M. Romanyshyn. 2023. P. 1–10. URL: https://aclanthology.org/2023.unlp-1.1 (дата звернення: 28.05.2024).
18. Hamotskyi S., Levbarg A.-I., Hänig C. Eval-UA-tion 1.0: benchmark for evaluating Ukrainian (large) language models. Unlp 2024, Turin. 2024. URL: https://hal.science/hal-04534651 (дата звернення: 28.05.2024).
19. The State and Fate of Linguistic Diversity and Inclusion in the NLP World / P. Joshi et al. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics / ed. by D. Jurafsky et al. 2020. P. 6282–6293. URL: https://aclanthology.org/2020.acl-main.560 (дата звернення: 31.05.2024).
20. Haltiuk M., Smywiński-Pohl A. LiBERTa: Advancing Ukrainian Language Modeling through Pre-training from Scratch. Unlp 2024, Turin. 2024.
21. Velankar A., Patil H., Joshi R. Mono vs multilingual BERT for hate speech detection and text classification: a case study in marathi. Artificial neural networks in pattern recognition / ed. by N. El Gayar et al. Cham, 2023. P. 121-128.
22. Named entity recognition for low-resource languages - profiting from language families / S. Torge et al. Proceedings of the 9th workshop on slavic natural language processing 2023 (slavicnlp 2023), Dubrovnik / ed. by J. Piskorski et al. P. 1–10. URL: https://aclanthology.org/2023.bsnlp-1.1 (дата звернення: 28.05.2024).
23. Mono- and multilingual GPT-3 models for Hungarian / Z. Yang et al. 2023. P. 94–104. URL: https://doi.org/10.1007/978-3-031-40498-6_9 (дата звернення: 28.05.2024).
24. Vīksna R., Skadina I. Multilingual slavic named entity recognition. Proceedings of the 8th workshop on balto-slavic natural language processing, Kiyv. P. 93–

97. URL: https://aclanthology.org/2021.bsnlp-1.11 (дата звернення: 28.05.2024).
25. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding / A. Wang et al. 2019.
26. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems / A. Wang et al. 2020. URL: https://doi.org/10.48550/arXiv.1905.00537 (дата звернення: 28.05.2024).
27. KLEJ: comprehensive benchmark for Polish language understanding / P. Rybak et al. Proceedings of the 58th annual meeting of the association for computational linguistics / ed. by D. Jurafsky et al. 2020. P. 1191–1201. URL: https://aclanthology.org/2020.acl-main.111 (дата звернення: 28.05.2024).
28. Ukrainian News Corpus as Text Classification Benchmark / D. Panchenko et al. 2022. P. 550–559. URL: https://doi.org/10.1007/978-3-031-14841-5_37.
29. Chaplynskyi D., Romanyshyn M. Introducing NER-UK 2.0: A Rich Corpus of Named Entities for Ukrainian. Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024, м. Torino / ред.: M. Romanyshyn та ін. 2024. С. 23–29. URL: https://aclanthology.org/2024.unlp-1.4 (дата звернення: 31.05.2024).
30. https://huggingface.co/dchaplinsky/uk_ner_web_trf_13class. URL: https://huggingface.co/dchaplinsky/uk_ner_web_trf_13class (дата звернення: 31.05.2024).
31. URL: https://github.com/lang-uk/ner-uk/ (дата звернення: 31.05.2024).
32. https://huggingface.co/benjamin/roberta-large-wechsel-ukrainian. URL: https://huggingface.co/benjamin/roberta-large-wechsel-ukrainian (дата звернення: 31.05.2024).
33. Minixhofer B., Paischer F., Rekabsaz N. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies / ed. by M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz. P. 3992–4006. URL: https://aclanthology.org/2022.naacl-main.293 (дата звернення: 31.05.2024).
34. Piskorski J., Marcinczuk M., Yangarber R. Cross-lingual Named Entity Corpus for Slavic Languages. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) / ed. by N. Calzolari et al. P. 4143–4157. URL: https://aclanthology.org/2024.lrec-main.369 (дата звернення: 31.05.2024).
35. Cross-lingual Name Tagging and Linking for 282 Languages / X. Pan et al. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) / ed. by R. Barzilay, K. Min-Yen. Vancouver, Canada, 2017. P. 1946–1958. URL: https://aclanthology.org/P17-1178 (дата звернення: 31.05.2024).

36. Contextual Embeddings for Ukrainian: A Large Language Model Approach to Word Sense Disambiguation / Y. Laba et al. Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), Dubrovnik / ed. by M. Romanyshyn. 2023. P. 11–19. URL: https://aclanthology.org/2023.unlp-1.2 (дата звернення: 31.05.2024).
37. Reimers N., Gurevych I. Sentence-BERT: sentence embeddings using siamese bert-networks. Conference on empirical methods in natural language processing. 2019. URL: https://api.semanticscholar.org/CorpusID:201646309 (дата звернення: 31.05.2024).
38. Radchenko V. We Trained the Ukrainian Language Model. URL: https://youscan.io/blog/ukrainian-language-model/ (дата звернення: 31.05.2024).
39. Universal Dependencies / J. Nivre et al. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts, Valencia / ed. by A. Klementiev, L. Specia. 2017. URL: https://aclanthology.org/E17-5001 (дата звернення: 31.05.2024).
40. From bytes to borsch: fine-tuning gemma and mistral for the Ukrainian language representation / A. Kiulian et al. 2024. URL: https://doi.org/10.48550/arXiv.2404.09138 (дата звернення: 28.05.2024).
41. Haltiuk M., Smywiński-Pohl A. LREC-COLING UNLP 2024 Presentation. URL: https://unlp.org.ua/wp-content/uploads/2024/06/haltiuk-pohla.pdf (дата звернення: 31.05.2024).

ДОДАТКИ

Додаток 1. Питання для оцінки моделей на завданні заповнення пропусків (mask filling). Згенеровані за допомогою ШІ

1. Столиця України, Київ, відома своїми історичними та культурними особливостями. Наприклад, у Києві знаходяться такі відомі пам'ятки, як [MASK] і [MASK].
2. Одного разу відважний козачок вирушив у чарівний ліс, де він зустрів [MASK] і [MASK].
3. Під час збору врожаю в українському селі люди зазвичай займаються [MASK] і [MASK]. Це важливий час для всіх жителів.
4. Українська мова відіграє важливу роль у сучасному світі, оскільки вона [MASK] і [MASK].
5. Моя кохана, ти як [MASK], що розцвітає навесні, як [MASK], що зігріває душу.
6. Національні свята України включають [MASK] і [MASK]. Вони пов'язані з такими традиціями, як [MASK] і [MASK].
7. Через 100 років Україна стане [MASK] і [MASK]. Люди будуть використовувати [MASK] і [MASK] для повсякденного життя.
8. Подорожуючи по найбільших містах України, варто відвідати [MASK] у Києві та [MASK] у Львові.
9. Рецепт борщу включає такі інгредієнти, як [MASK] і [MASK]. Процес приготування починається з [MASK] і закінчується [MASK].
10. Відомий український діяч [MASK] зробив великий внесок у розвиток країни, зокрема в галузі [MASK] і [MASK].
11. Перший день в університеті був незабутнім. Я зустрів [MASK] і [MASK], а також відвідав [MASK].
12. Ідеальний день на березі моря включає [MASK] і [MASK]. Це дає відчуття [MASK] і [MASK].
13. Моя улюблена книга - це [MASK], тому що [MASK].
14. Мій улюблений український рецепт - це [MASK]. Для його приготування потрібно [MASK] і [MASK].
15. Останній фільм, який я дивився, був про [MASK]. Він мені сподобався, бо [MASK].
16. Моє хобі - це [MASK]. Я займаюся ним, коли [MASK], і це приносить мені [MASK].
17. У нашій родині традиції святкування Різдва включають [MASK] і [MASK].
18. Лист моєму майбутньому я через 10 років: Я сподіваюся, що ти досяг [MASK] і [MASK]. Пам'ятай про [MASK] і [MASK].
19. Найкращі моменти мого життя пов'язані з [MASK] і [MASK]. Це було [MASK] і [MASK].
20. Якби я мав мільйон доларів, я б витратив його на [MASK] і [MASK].

21. Моя улюблена подорож була до [MASK], де я бачив [MASK] і [MASK].
22. Найважливіше у дружбі для мене - це [MASK] і [MASK].
23. Моя улюблена пора року - це [MASK], тому що [MASK].
24. Мій улюблений фільм або серіал - це [MASK]. Він мені подобається через [MASK].
25. Мій ідеальний дім - це місце, де є [MASK] і [MASK].
26. Якщо б я потрапив на безлюдний острів, я б взяв із собою [MASK], [MASK] і [MASK].
27. Одного разу у фантастичному світі відбулася подія: [MASK] і [MASK]. Це було незабутньо.
28. Якби я міг змінити своє життя, я б [MASK] і [MASK].
29. Моя ідеальна робота - це [MASK], де я можу [MASK] і [MASK].
30. Моя улюблена українська страва - це [MASK]. Вона смачна, бо [MASK].
31. Мій найбільший страх - це [MASK]. Це мене лякає, бо [MASK].
32. Я б порадив прочитати кожному книгу [MASK], тому що [MASK].
33. Для мене успіх означає [MASK] і [MASK].
34. Мій ранок вихідного дня починається з [MASK] і [MASK].
35. Моя улюблена музика - це [MASK]. Мені подобаються виконавці [MASK], бо [MASK].
36. Я мрію відвідати [MASK], тому що [MASK].
37. Свій вільний час я зазвичай проводжу, займаючись [MASK] і [MASK].
38. У дитинстві я любив займатися [MASK] і [MASK]. Це приносило мені [MASK].
39. Найбільш пам'ятний подарунок, який я отримував - це [MASK]. Це було [MASK], бо [MASK].
40. Через п'ять років я бачу себе [MASK], де я [MASK] і [MASK].
41. Сонячні батареї працюють завдяки [MASK]. Вони перетворюють [MASK] на [MASK].
42. Інтернет функціонує за допомогою [MASK] і [MASK], що дозволяє обмінюватися інформацією між [MASK].
43. Історія Київської Русі починається з [MASK] і включає події, як [MASK] і [MASK].
44. Видатні українські письменники, такі як [MASK] і [MASK], зробили великий внесок у літературу.
45. Людська пам'ять функціонує завдяки [MASK] і [MASK], що дозволяє запам'ятовувати [MASK].
46. Концепція еволюції полягає в [MASK] і [MASK], що дозволяє видам адаптуватися до [MASK].
47. ДНК - це [MASK], що відповідає за [MASK] і [MASK] в організмі.
48. Культурні традиції українців включають [MASK] і [MASK], які є важливою частиною нашої спадщини.

49. Економіка ринкового типу працює за принципами [MASK] і [MASK], що дозволяє [MASK].
50. Штучний інтелект застосовується в [MASK] і [MASK], що дозволяє [MASK].
51. Принципи демократичної системи управління включають [MASK] і [MASK], що забезпечують [MASK].
52. Основні етапи в історії України включають [MASK] і [MASK], які визначили [MASK].
53. Вакцини діють, стимулюючи [MASK] і [MASK], що забезпечує [MASK].
54. Тарас Шевченко зробив великий внесок у літературу, створивши [MASK] і [MASK], які стали [MASK].
55. Хімічні зміни відрізняються від фізичних тим, що [MASK], а фізичні зміни [MASK].
56. Процес перетворення енергії в електростанції включає [MASK] і [MASK], що дозволяє отримувати [MASK].
57. Чорні діри мають такі властивості, як [MASK] і [MASK], що робить їх [MASK].
58. Нервова система людини функціонує завдяки [MASK] і [MASK], що дозволяє [MASK].
59. Фактори, що впливають на зміну клімату, включають [MASK] і [MASK], які призводять до [MASK].
60. Історія козацтва включає [MASK] і [MASK], що визначило [MASK].
61. Блокчейн-технологія працює за принципами [MASK] і [MASK], що забезпечує [MASK].
62. Нанотехнології застосовуються в [MASK] і [MASK], що дозволяє [MASK].
63. Процеси, що відбуваються в земній корі під час землетрусів, включають [MASK] і [MASK], що призводить до [MASK].
64. Ферменти в організмі виконують роль [MASK], що дозволяє [MASK] і [MASK].
65. Українські народні ремесла включають [MASK] і [MASK], які є важливою частиною культури.
66. Міжнародна торгівля функціонує завдяки [MASK] і [MASK], що дозволяє [MASK].
67. Квантова фізика включає такі положення, як [MASK] і [MASK], що дозволяє розуміти [MASK].
71. Українські національні костюми відрізняються своєю красою та унікальністю. Наприклад, у Західній Україні популярні [MASK], а на Сході [MASK].
72. Карпати - це прекрасний гірський регіон України, де можна знайти [MASK] і [MASK]. Цей регіон відомий своїми [MASK] і [MASK].
73. Українська кухня багата на традиційні страви, такі як [MASK] і [MASK]. Особливо смачні вони, коли приготовлені за старовинними рецептами.

74. Микола Гоголь відомий своїми творами, такими як [MASK] і [MASK]. Він народився в [MASK] і багато писав про [MASK].
75. Чорне море є важливою частиною української географії. Воно омиває береги [MASK] і [MASK], де розташовані такі міста, як [MASK] і [MASK].
76. Літературний процес в Україні завжди був багатим і різноманітним. Наприклад, у XIX столітті популярними були [MASK] і [MASK], а у XX столітті [MASK] і [MASK].
77. Українські народні пісні відомі своєю мелодійністю і глибиною змісту. Наприклад, пісня [MASK] розповідає про [MASK], а пісня [MASK] - про [MASK].
78. Видатний український художник Казимир Малевич створив такі відомі роботи, як [MASK] і [MASK]. Його творчість вплинула на розвиток [MASK].
79. В українському фольклорі часто зустрічаються персонажі, як [MASK] і [MASK], які уособлюють [MASK].
80. Українські Карпати славляться своїми курортами, такими як [MASK] і [MASK], де можна насолоджуватися [MASK].
81. Богдан Хмельницький був відомим українським гетьманом, який досяг [MASK] у боротьбі за незалежність України.
82. Софія Київська - це [MASK], яка була побудована у [MASK] році і є прикладом [MASK] архітектури.
83. Українські вишиванки мають різноманітні орнаменти, які символізують [MASK] і [MASK]. У різних регіонах України можна знайти [MASK] вишивки.
84. Гуцули - це етнографічна група, яка мешкає у Карпатах і відома своїми [MASK]. Їх культура включає [MASK].
85. Укрзалізниця - це [MASK], яка забезпечує [MASK] між містами України. Її мережа включає [MASK].
86. Українські замки, такі як [MASK] і [MASK], є важливими історичними пам'ятками, які розповідають про [MASK].
87. Майдан Незалежності в Києві є символом [MASK]. Він відомий своїми [MASK].
88. Традиційні українські музичні інструменти включають [MASK]. Вони використовуються для виконання [MASK].
89. Українська писанка - це [MASK], яка символізує [MASK]. Вона розмальована за допомогою [MASK].
90. Володимир Великий був князем Київської Русі, який прийняв християнство у [MASK] році, що вплинуло на [MASK].
91. Олесь Гончар - це відомий український письменник, який написав такі твори, як [MASK].
92. Українські картини-пейзажі часто зображують [MASK], які передають красу [MASK].
93. Рушник у традиційній українській культурі символізує [MASK]. Він використовується в [MASK].

94. Українські легенди і міфи часто включають персонажів, таких як [MASK], які уособлюють [MASK].
95. Українські ярмарки відомі своєю атмосферою, де можна знайти [MASK]. Вони проводяться у [MASK].
96. Національний музей історії України в Києві зберігає артефакти, такі як [MASK], що розповідають про [MASK].
97. Олександр Довженко - видатний український кінорежисер, відомий своїми фільмами, такими як [MASK].
98. Сучасна українська музика включає такі жанри, як [MASK], які виконуються такими артистами, як [MASK].
99. Традиційні українські танці включають [MASK]. Вони виконуються на [MASK].
100. Українські обрядові свята включають [MASK], які відзначаються такими традиціями, як [MASK].

Додаток 2. Вебсайт для оцінки результатів заповнення пропусками моделей

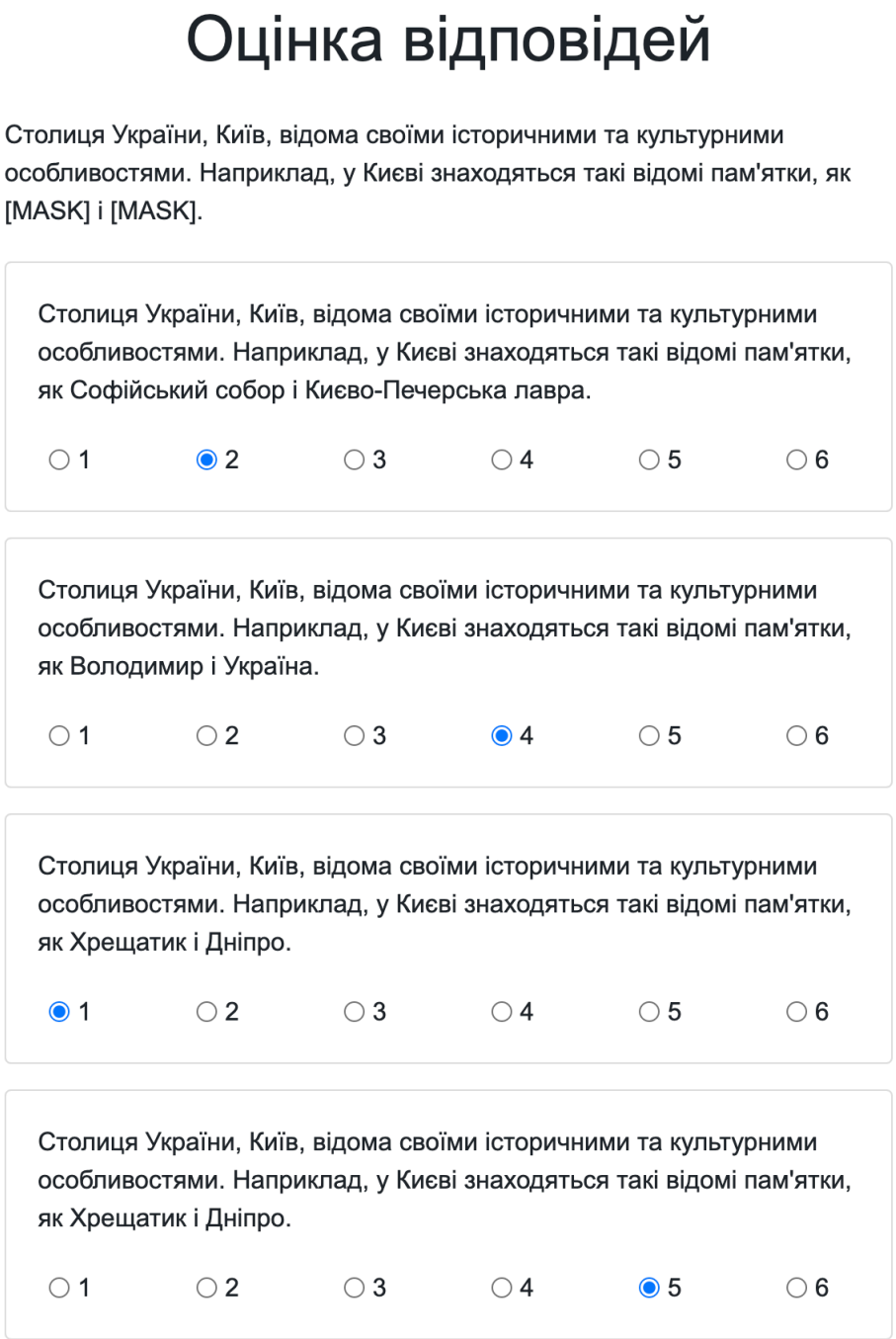

Рисунок 3. Додаток 2. Фрагмент інтерфейсу вебсайту для оцінки результатів заповнення пропусків.