УДК 004.912
DOI: https://doi.org/10.17721/1812-5409.2026/1.26

Микола ГЛИБОВЕЦЬ, д-р фіз.-мат. наук, проф.
ORCID ID: 0009-0005-6942-8026
e-mail: glib@ukma.edu.ua
Національний університет "Києво-Могилянська академія", Київ, Україна

Данило ВАНІН, магістр
ORCID ID: 0009-0007-4510-8209
e-mail: danylo.vanin@gmail.com
Національний університет "Києво-Могилянська академія", Київ, Україна

ВИКОРИСТАННЯ ОДНО- ТА БАГАТОМОВНИХ МОДЕЛЕЙ НА БАЗІ BERT ДЛЯ РОЗВ'ЯЗАННЯ ЗАДАЧ АВТОМАТИЧНОЇ ОБРОБКИ ТЕКСТІВ УКРАЇНСЬКОЮ МОВОЮ

Об'єктом дослідження пропонованої статті є одно- та багатомовні моделі на основі BERT. Предметом дослідження – порівняння продуктивності таких моделей на завданнях обробки природної мови (ОПМ) з акцентом на їхньому застосуванні для української мови. Методологічну основу порівняльного аналізу становило використання стандартних підходів до навчання й оцінки моделей. У дослідженні використовували доступні джерела інформації. Експерименти виконувалися лише на відкритих або власноруч донавчених моделях й охоплювали наявні українські бенчмарки.

Результати дослідження свідчать про те, що як одномовні, так і багатомовні моделі на основі BERT можуть бути ефективними для розв'язання завдань ОПМ залежно від конкретної мови, завдання та доступних ресурсів. Хоча одномовні моделі часто перевершують багатомовні в завданнях своєї конкретної мови, багатомовні моделі можуть мати перевагу, коли ресурси для навчання одномовних моделей обмежені.

Результати порівняння демонструють те, що моделі, натреновані "з нуля" на українських текстах або дотреновані на них, загалом мають кращі результати на завданнях ОПМ української мови, ніж багатомовні. Це вдалося показати на завданнях розпізнавання іменованих сутностей, класифікації текстів і заповнення пропусків. Багатомовні моделі демонструють конкурентні або кращі результати у "легших" завданнях, наприклад, у задачі розрізнення значень слів, особливо коли ресурси для повноцінного тренування одномовної моделі обмежені.

Таке зіставлення підтверджує необхідність адаптації або повного тренування україномовних моделей для спеціалізованих задач, водночас залишаючи простір для економічно доцільного використання багатомовних систем у типових сценаріях.

Практична цінність роботи полягає у формуванні узагальненої карти продуктивності сучасних моделей та у підґрунті для створення повномасштабного українського GLUE-подібного бенчмарку, що стимулюватиме подальші дослідження та підвищить точність застосунків ОПМ для української мови.

Ключові слова: ***обробка природної мови, великі мовні моделі, одно- та багатомовні моделі, BERT.***

Класифікація відповідно до AMS 2020: 68T50, 68T07.

Вступ

Обробка природної мови (ОПМ) охоплює безліч завдань, спрямованих на те, щоб інтелектуальні програмні системи могли розуміти, інтерпретувати та генерувати людську мову. Останні дослідження у сфері глибокого машинного навчання, зокрема створення та використання моделей на основні архітектури трансформера (Vaswani et al., 2017; Глибовець, & Бікчентаєв, 2022), як BERT (Devlin et al., 2019), значно розширили можливості для розв'язання задач обробки природньої мови (ОПМ). Та їх використання потребує значних ресурсів.

Досить перспективною альтернативою для мов з обмеженими ресурсами стали багатомовні (multilingual) моделі, що 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. Багатомовні моделі можуть "перевикористовувати" певні "знання" про мови з великими ресурсами на мовах з обмеженими ресурсами.

Нині багато досліджень присвячені вивченню можливостей одномовних і багатомовних моделей. Деякі дослідження, як-от (Virtanen et al., 2019), демонструють, що одномовні моделі перевершують багатомовні на багатьох окремих завданнях для певної мови. Інші дослідження доводять, що багатомовні моделі можуть одночасно бути ефективними для великоресурсних мов і показувати кращі результати, ніж одномовні для тих мов, де ресурси обмежені (напр. у випадку української мови).

Об'єктом дослідження цієї роботи є одно- та багатомовні моделі на основі BERT. Предметом дослідження – порівняння продуктивності таких моделей на окремих типах завдань ОПМ (української мови): розпізнавання іменованих сутностей, розрізнення значень, класифікація текстів, заповнення пропусків і розмічування частин мови. Мета дослідження - оцінити ефективність різних типів моделей BERT для застосування до виконання зазначених завдань. Для досягнення цієї мети було вирішено оцінити: результати одномовних моделей, які були натреновані на переважно українських текстах; адаптованих до української багатомовних моделей (наприклад, за методом WECHSEL); використання "чистих" багатомовних моделей. Зрозуміло, що потрібно було провести і порівняльний аналіз отриманих результатів застосування різних типів моделей та визначити особливості їхнього використання.

Методологічною основою дослідження було використання стандартних підходів до навчання й оцінки моделей. Ми використовували доступні дотреновані на завданні моделі, або базові моделі, які були дотреновані для завдання у межах дослідження. Оцінювали моделі за допомогою таких метрик, як F1-score (для розпізнавання іменованих сутностей) або точність (для розмічування частин мови). Завдання заповнення пропусків оцінювали за допомогою опитування, а також через ручне порівняння й аналіз якості відповідей різних моделей.

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

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

1. Основні означення та результати

Обробка природної мови (ОПМ) передбачає безліч завдань. Серед них можна виділити: класифікація тексту (Text Classification); виявлення спаму (Spam Detection); визначення наміру (Intent Detection); маркування послідовностей (Sequence Labeling); генерація тексту (Text Generation); машинний переклад (Machine Translation); порівняння тексту (Text Comparison); інформаційний пошук (Information Retrieval); аналіз тексту (Text Analysis); розуміння мови (Language Understanding).

Вибір архітектури мовної 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 розв'язано і наївним баєсовим класифікатором. Поширена зараз трансформерна архітектура стала дуже популярною завдяки універсальності та здатності навчатися виявляти зв'язки між елементами тексту, розташованими на різних відстанях. Вони розв'язують проблему згасання важливості контексту та регулярного перерахунку контекстних ваг у рекурентних архітектурах.

Часто завданням на етапі базового тренування є просте передбачення слова в 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.

Незважаючи на те, що BERT уже вміють на загальному рівні розуміти мови (вміння передбачати слова у контексті), усе ж цього недостатньо для більш специфічних завдань. Базова модель може лише передбачати масковані слова, а для складнішого завдання, відповідно, як відповідь на питання чи переказ тексту – модель має бути донавчена на розмічених даних. Замість того, щоб навчати модель "з нуля", процес тонкого налаштування починається з використання вже наявної попередньо навченої моделі, яку адаптують під конкретне завдання. Це дає змогу "перевикористати" знання, отримані з великого неструктурованого набору даних під час попереднього навчання. Додавання нових даних, що близькі до кінцевого завдання, допомагає налаштувати модель для розв'язання конкретних задач.

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

Одномовний BERT (*Bidirectional Encoder Representations from Transformers* – двонаправлене кодування представлень із трансформерів) – це мовна модель, що використовує трансформерну архітектуру. Вона складається із шарів самоуваги (self-attention) і нейронних мереж прямого поширення (Devlin et al., 2019). 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.

Такий підхід допомагає 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 із мінімальними модифікаціями, специфічними для завдання (Devlin et al., 2019). Розроблені багатомовні варіанти BERT, такі як mBERT та XLM-RoBERTa (Shaham et al., 2024), 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 поліпшення роботи на окремих мовах, а особливо для мов з низьким рівнем ресурсів, таких як українська (Liu et al., 2019). Основна ідея полягає в тому, що літери та фрагменти слів часто збігаються у 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. Питання того, які з двох моделей є ефективнішими для конкретної мови, можна визначити лише емпірично. Це дослідження присвячене такому емпіричному порівнянню цих моделей на різних завданнях ОПМ для української. Багатомовні моделі можуть досягати кращих результатів у широкому діапазоні мов з великими ресурсами. Такий "крослінгвістичний трансфер" дає моделям змогу використовувати знання, отримані в одній мові, для інших мов, що особливо корисно для мов з обмеженими ресурсами (Sanh et al., 2019).

Одномовні моделі часто можуть перевершувати свої багатомовні аналоги в деяких завданнях, включно з виявленням мови ворожнечі та генеруванням речень (Pires et al., 2019; Sido et al., 2021). 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 (Ruder, 2022). Важливим також є тип даних, на яких тренувалася модель.

Порівняння результатів одномовних і багатомовних моделей для української мови може вказати, у яких напрямках дослідники повинні приділити більше уваги розробленню якісних україномовних наборів даних і моделей.

Українська мова має багату морфологічну структуру, характерну для синтетичних мов (Haltiuk, & Smywiński-Pohl, 2024). Вона стикається зі значними викликами в області обробки природної мови, насамперед через нестачу даних для тренування. У 2020 р. українську мову визначили як "rising star" – мову, що страждає від нестачі розмічених наборів даних (Joshi et al., 2020). Основною перешкодою для української мови є нестача загальнодоступних, орієнтованих на конкретні завдання корпусів (Haltiuk, & Smywiński-Pohl, 2024). Це заважало розвитку надійних одномовних моделей

для української мови та змушувало дослідників покладатися на крос-мовне трансферне навчання з інших мов (Hamotskyi et al., 2024; Torge et al., 2023).

У представленому дослідженні буде також перевірена гіпотеза, що модель, натренована на корпусі слов'янських мов, може показати кращі результати, ніж одномовна модель для конкретних завдань.

2. Порівняння одно та багатомовних моделей на завданнях обробки 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 дедалі більший 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., 2019) або SuperGLUE (Wang et al., 2020) для англійської чи KLEJ (Rybak et al., 2020) для польської, який би покривав широке коло завдань, наразі немає.

Нам були доступні комплексні бенчмарки: UA-datasets (Panchenko et al., 2022), Eval-UA-tion (Chaplynskyi, & Romanyshyn, 2024). Перша колекція містить набори даних для різних завдань, таких як розмітка частин мови (POS tagging) і класифікація новин (Panchenko et al., 2022). Другий набір охоплює дані для оцінки продуктивності мовних моделей у таких завданнях, як розуміння наративів оповідань (UA-CBT), асоціація статей із їхніми заголовками (UP-Titles) й елементарні мовні завдання (LMES).

Через відсутність готових комплексних бенчмарків для порівняння моделей, у пропонованій 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 і розмічення частин мови.

2.2. Розпізнавання іменованих сутностей (NER)

Розпізнавання іменованих сутностей 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 (Chaplynskyi, & Romanyshyn, 2024), що містить 323 200 слів і 21 993 іменовані сутності. У дослідженні автори також долучили початкові результати для моделі robertalarge (Hugging Face, 2024). У табл. 1 наведено порівняння результатів різних моделей на попередній версії корпусу NER-UK 1.0 (GitHub, 2025). Тут о*б'єднані* результати з (Haltiuk, & Smywiński-Pohl, 2024); Hugging Face, 2024); GitHub, 2025) та власного дослідження. Значення в дужках позначають стандартне відхилення метрики.

Таблиця 1

Результати оцінки моделей на NER-UK 1.0

Тип і 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 1 містить результати моделі, натренованої для української мови методом WECHSEL (Hugging Face, 2025), багатомовної моделі XLM-R та нещодавно опублікованої моделі LiBERTa (Haltiuk, & Smywiński-Pohl, 2024). Остання була натренована з "нуля" на наборах даних української мови. Порівняння показує, що моделі, адаптовані для української мови, ефективніші за багатомовні в розпізнаванні іменованих сутностей.

Як зазначено в (Haltiuk, & Smywiński-Pohl, 2024), 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 містить результати моделі на корпусі слов'янських мов (болгарська, чеська, польська, словенська, російська й українська) (Piskorski et al., 2024). 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 моделях, але все ж гірші, ніж у моделей, адаптованих для української мови.

На бенчмарку WikiANN (Pan et al., 2017), який також передбачає розпізнавання іменованих сутностей, LiBERTa, яка повністю натренована на українській мові, показує гірші результати. Результати порівняння моделей наведені у табл. 2. Значення взято з (Haltiuk, & Smywiński-Pohl, 2024).

Таблиця 2

Результати оцінки моделей на WikiANN

Тип і 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)

Автори зазначають, що такі результати можуть бути спричинені відмінностями між датасетами та навчальними даними для моделі. Модель і токенізатор були натреновані переважно на українських текстах, тоді як багатомовні моделі часто тренуються на даних, зібраних із Wikipedia (Haltiuk, & Smywiński-Pohl, 2024).

Отже, у завданні NER моделі, адаптовані для української мови, показують вищу продуктивність порівняно з багатомовними моделями. Для цього завдання створення та використання моделей, орієнтованих на українську мову, важливе для досягнення кращих результатів. WECHSEL-RoBERTa та LiBERTa, наприклад, 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 розв'язати проблему неоднозначності мови (наявність омонімів "ключ журавлів" – "ключ від дверей"). Оскільки багато слів мають кілька значень, натренована модель намагається визначити правильний сенс слова залежно від контексту. Розв'язок цього завдання може бути частиною ширшого конвеєру з моделей, ціль якого переказати текст чи відповісти на питання про текст.

У дослідженні (Laba et al., 2023) автори порівнюють точність розрізнення значень слів (WSD) для української мови, використовуючи різні стратегії об'єднання та попередньо навчені моделі без додаткового тренування.

Таблиця 3 (значення запозичено з (Laba et al., 2023)) демонструє, що модель paraphrasemultilingual-mpnet-base-v2 (PMMBv2) з (Reimers, & Gurevych, 2019) показує кращі результати WSD, ніж модель RoBERTa, попередньо навчена на українській мові у 2020 р. (Radchenko, 2020). У роботі (Laba et al., 2023) підвищили продуктивність PMMBv2 до 0.779, тонко налаштувавши її на основі датасету зі Словника української мови (СУМ).

Таблиця 3

Середня точність розрізнення значень слів без тонкого налаштування

Тип і 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**

Утім, щоб отримати релевантне порівняння між одномовними та багатомовними моделями в задачі WSD, необхідно протестувати сучаснішу українську модель, таку як LiBERTa (Chaplynskyi, & Romanyshyn, 2024), з додатковим налаштуванням на тому ж наборі даних.

2.4. Класифікація текстів

У дослідженні (Haltiuk, & Smywiński-Pohl, 2024) автори представили модель LiBERTa, що була попередньо навчена з нуля на українських даних. Вони надають результати порівняння її продуктивності на завданні класифікації текстів з Ukranian News Classification benchmark (Panchenko et al., 2022). Цей бенчмарк показує, наскільки точно модель може визначити, до якого новинного джерела належить певна стаття (багатокласова класифікація).

Як видно з табл. 4 (значення запозичено з (Haltiuk, & Smywiński-Pohl, 2024)), модель, попередньо навчена на українських даних, перевершує багатомовну модель XLM-R за продуктивністю. Однак модель, адаптована до української мови за допомогою методу WECHSEL, показує ще кращі результати.

Таблиця 4

Порівняння результатів моделей на задачі класифікації українських новин

Тип і 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)

2.5. Eval-UA-tion

Eval-UA-tion (Hamotskyi et al., 2024) – це спеціальний бенчмарк, розроблений для оцінки продуктивності мовних моделей на українській мові. Він складається з трьох основних завдань: UA-CBT, UP-Titles, LMentry-static-UA (LMES).

Перше 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 і передбачати пропущені слова на основі контексту.

У завданні UP-Titles потрібно зіставити статті з онлайн-газети "Українська Правда" з їхніми 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) – 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-го слова в реченні.

Eval-UA-tion містить базові результати для кожного завдання: показник людських відповідей і випадкової генерації. Мета Eval-UA-tion полягає у всебічній оцінці можливостей мовних моделей для української мови та стимулюванні подальших досліджень у цій сфері. Він був представлений на конференції UNLP 2024 і містить оцінки для різних багатомовних й одномовних моделей. Незважаючи на використання сучасніших моделей замість BERT, порівнянню яких присвячено цю роботу, автори показують результати, що є корисними для порівняння продуктивності багатомовних й одномовних моделей.

Згідно з результатами, представленими на рис. 1, модель SherlockAssistant/Mistral-7B-Instruct-Ukrainian, яка пройшла попереднє налаштування на українських даних, виявилася ефективнішою для завдання UP-Titles (Hamotskyi et al., 2024), ніж багатомовні моделі, що не є OpenAI, та навіть перевершила GPT-3. Хоча GPT-4, багатомовна модель, показала найкращі результати, варто зауважити, що пряме порівняння ускладнюється через значні відмінності у розмірі моделі й обсязі набору даних для попереднього навчання порівняно з іншими моделями.

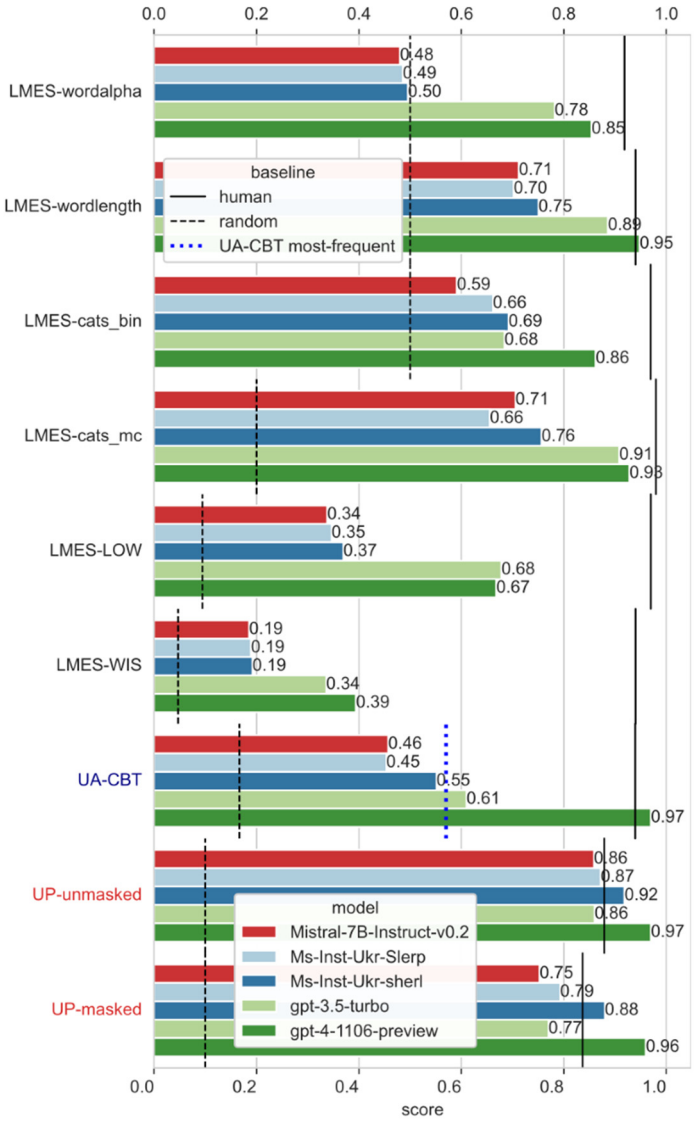

Рис. 1. Результати порівняння моделей на завданнях Eval-UA-tion (Hamotskyi et al., 2024). Позначення: baseline – базова модель; human – оцінка людиною; random – випадковий вибір; UA-CBT most-frequent – найчастіший клас. Вісь X – значення метрики (score)

2.6. Заповнення пропусків (Mask filling)

На етапі попереднього навчання моделей на основі BERT зазвичай 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 нами 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 дає змогу 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 набір питань, що генерувалися моделлю Gemini, а також дописувалися та перевірялися вручну. Для заповнення пропусків на кожному питанні використовували кілька моделей. Оцінювані моделі охоплювали багатомовні моделі, моделі, які були додатково навчені на українських текстах, й 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, що згенерувала конкретну відповідь, була прихована). 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), то для опитування були запрошені люди. Для всіх учасників українська мова є основною мовою спілкування. Усі учасники мали вищу освіту.

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

Результати ручного оцінювання

Багатомовні моделі зазвичай показують гірше знання українського контексту (табл. 6).

Моделі, попередньо навчені або перенесені на українську, часто показують краще знання українського контексту (напр., у табл. 7 модель WECHSEL додала нішеві українські топоніми).

Таблиця 6

Приклад результатів заповнення пропусків різними моделями

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 7

Приклад результату заповнення пропуску моделлю WECHSEL

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, д е можна насолоджуватися природою.

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

Однак усі 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 із науковими фактами (табл. 8, 9).

Таблиця 8

Приклад результатів заповнення пропусків, що вимагають знання фактів, різними моделями

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 9

Приклад результатів заповнення пропусків, що вимагають знання фактів про історію України чи біологію, згенеровані 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** й **відео**, що дозволяє запам'ятовувати **інформацію**.
RoBERTa WECHSEL large	Людська пам'ять функціонує завдяки **науці** й **розуму**, що дозволяє запам'ятовувати **людей**.
YouScan RoBERTa Uk base	Людська пам'ять функціонує завдяки **технології** й **енергії**, що дозволяє запам'ятовувати **інформацію**.
XLM RoBERTa large	Людська пам'ять функціонує завдяки **руху** і **часу**, що дозволяє запам'ятовувати **інформацію**.

Це ще раз підкреслює значну перевагу великих комерційних моделей над меншими безкоштовними моделями з відкритим кодом. Така відмінність у продуктивності, ймовірно, пов'язана з тим, що комерційні моделі мають більшу кількість параметрів і 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, наведені на рис. 2, показують, що моделі, перенесені на українську за допомогою методу WECHSEL, перевершують модель YouScan RoBERTa, попередньо навчену на українській мові. Однак найнижчі бали отримали багатомовні моделі XLM-R.

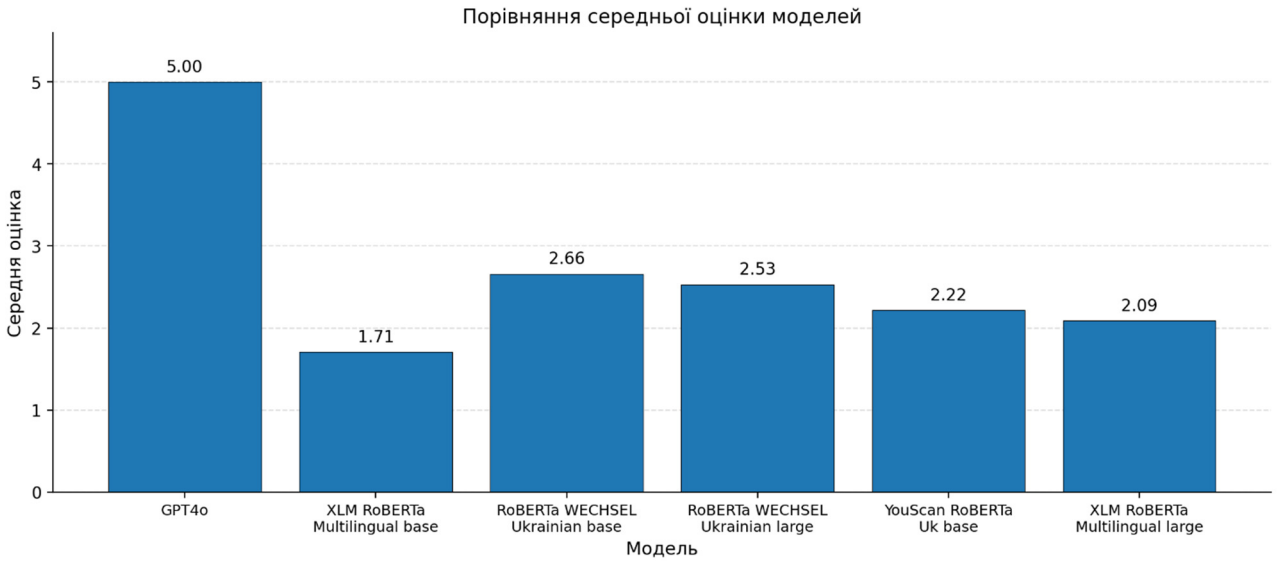

Рис. 2. Результати загального оцінювання моделей (онлайн-опитування)

Водночас жодна модель не досягла якості комерційної моделі GPT4o.

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 (Haltiuk, & Smywiński-Pohl, 2024).

2.7. Розмічування частин мови (POS tagging)

Автори (Haltiuk, & Smywiński-Pohl, 2024) порівнюють продуктивність різних багатомовних й одномовних моделей на українській частині набору даних Universal Dependencies (Nivre et al., 2017). 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 (Haltiuk, & Smywiński-Pohl, 2024), 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 в табл. 10.

Таблиця 10

Результати моделей на задачі розмічування частин мови

Тип і 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)

Отже, це порівняння показує, що для розмічування частин мови різниця у продуктивності між одномовними й багатомовними моделями є мінімальною.

2.8. Результати порівняння

Результати порівняння одно- та багатомовних моделей наведено в табл. 11.

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

Таблиця 11

Порівняння одно- та багатомовних моделей на українських завданнях NLP

Завдання ОПМ для української мови	Порівняння одно- та багатомовних моделей	Додаткові коментарі
Розпізнавання іменованих сутностей (NER)	Одномовна модель тренована переважно 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 EvalUA-tion)	Модель Mistral-7BSherlock дотренована на українських даних показує кращі результати ніж багатомовна модель	Mistral-7B-Sherlock не досягає рівня gpt-41106-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

Водночас багатомовні моделі, дотреновані українською мовою, також можуть бути ефективними для завдань, що не потребують глибокого розуміння специфіки мови. Прикладом таких задач є розрізнення значень. Потенційно відмінність між моделями, натренованими з "нуля" українською та дотренованими українською, може зменшитись, залежно 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, однак уже 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 (Haltiuk, & Smywiński-Pohl, 2024) 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 відмінність потенційно може нівелюватися більше, якщо моделі були адаптовані до української мови.

3. Обмеженість і майбутнє дослідження

Описавши отримані результати, важливо відзначити певні обмеження цього дослідження, які варто враховувати під час інтерпретації та подальшого використання висновків.

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

Для кращого порівняння роботи моделей потрібно мати ресурси для тренування моделі на корпусі слов'янських мов. До виходу LiBERTa (Haltiuk, & Smywiński-Pohl, 2024) існувала потреба також навчити з "нуля" україномовну модель, що, попри спроби, не вдалося зробити в межах поточного дослідження через брак обчислювальних ресурсів.

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 суттєво ускладнює порівняння різних моделей і підходів, оскільки доводиться використовувати окремі набори даних і метрики, які можуть бути не повністю репрезентативними для всіх аспектів мови.

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

Використання результатів конференції UNLP 2024. Конференція UNLP 2024, що відбулася 25 травня 2024 р., стала важливою подією для розвитку ОПМ в Україні. На конференції представили нові моделі, натреновані "з нуля" чи дотреновані на українських текстах (Haltiuk, & Smywiński-Pohl, 2024; Kiulian et al., 2024). Крім того, конференція відзначилася розширенням бенчмарків для оцінки моделей. У межах нашої роботи ми здійснили спробу інтегрувати деякі із цих результатів, проте обмежений час не дав змоги ґрунтовніше використати нові моделі та бенчмарки.

Майбутня робота та напрямки досліджень. Результати дослідження показали, що спеціалізовані україномовні моделі мають значні переваги в більшості завдань з обробки української мови. 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, та появу невеликих комплексних бенчмарків.

Основними напрямками майбутньої роботи є: розробка комплексного бенчмарку, який буде українським аналогом GLUE; розгляд завдань на глибоке розуміння української мови; розширення та диверсифікація набору даних, що використовуються для навчання й оцінки моделей. Це сприятиме підвищенню точності й адекватності моделей.

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 і результатів роботи моделей, а також розробку механізмів контролю та корекції.

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

Дискусія і висновки

В межах представленого дослідження вдалося порівняти результати одно- та багатомовних моделей типу BERT на окремих завданнях обробки української мови. Результати порівняння демонструють те, що моделі, натреновані "з нуля" на українських текстах або дотреновані на них, загалом мають кращі результати, ніж багатомовні. Це вдалося показати на завданнях розпізнавання іменованих сутностей, класифікації текстів та заповнення пропусків. Водночас багатомовні моделі показують наближений або кращий результат на легших завданнях (розрізнення значень слів).

Українські моделі показали кращі можливості захоплювати український контекст і враховувати культурні нюанси, які можуть втрачатися в 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 (Haltiuk, & Smywiński-Pohl, 2024), друга модель уже 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.

Дослідження показало, що вибір між одно- та багатомовними моделями значно залежить від завдання.

У роботі були зібрані та порівняні одно- та багатомовні моделі для різних мов, що додатково підкреслило важливість проведення окремого порівняння для української мови. 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, серед яких і наукові статті та моделі. Деякі з використаних моделей були додатково налаштовані. Для завдань, де потрібно було заповнити пропуски, ми використовували готові попередньо треновані моделі.

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

Процес порівняння моделей значно ускладнювала відсутність комплексного бенчмарку для української мови, який був би аналогом GLUE або KLEJ.

Результати пропонованого дослідження можуть допомогти у визначенні оптимальної моделі для завдань, що були представлені в порівнянні. Крім того, зібрані показники порівняння моделей і 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 допоможе забезпечити їхню ефективність у широкому спектрі завдань обробки мови.

Внесок авторів: Микола Глибовець – концептуалізація, методика; Данило Ванін – програмне забезпечення; Микола Глибовець – аналіз і підготовка теоретичних основ дослідження; Данило Ванін – аналіз результатів, збір і перевірка емпіричних даних, аналіз джерел, підготовка огляду літератури.

Джерела фінансування. Це дослідження не отримало жодного гранту від фінансової установи в державному, комерційному або некомерційному секторах.

Список використаних джерел

Глибовець, А. М., & Бікчентаєв, М. (2022). Програмна система перевірки на плагіат українських текстів. *Наукові записки НаУКМА. Комп'ютерні науки*, 5, 16–25. https://doi.org/10.18523/2617-3808.2022.5.16-25

Chaplynskyi, D., & Romanyshyn, M. (2024). Introducing NER-UK 2.0: A rich corpus of named entities for Ukrainian. In M. Romanyshyn, N. Romanyshyn, A. Hlybovets, & O. Ignatenko (Eds.), *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024* (pp. 23–29). ELRA and ICCL.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

GitHub. (2025). *Lang-Uk/ner-uk.* https://github.com/lang-uk/ner-uk

Haltiuk, M., & Smywiński-Pohl, A. (2024). LiBERTa: Advancing Ukrainian language modeling through pre-training from scratch (Presentation). In M. Romanyshyn, N. Romanyshyn, A. Hlybovets, & O. Ignatenko (Eds.), *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024* (pp. 120–128). ELRA and ICCL.

Hamotskyi, S., Levbarg, A.-I., & Hänig, C. (2024). Eval-UA-tion 1.0: Benchmark for evaluating Ukrainian (large) language models. In M. Romanyshyn, N. Romanyshyn, A. Hlybovets, & O. Ignatenko (Eds.), *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024* (pp. 109–119). ELRA and ICCL.

Hugging Face. (2024). *Uk_ner_web_trf_13class.* https://huggingface.co/dchaplinsky/uk_ner_web_trf_13class

Hugging Face. (2025). *Roberta-large-wechsel-ukrainian.* https://huggingface.co/benjamin/roberta-large-wechsel-ukrainian

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

Kiulian, A., Polishko, A., Khandoga, M., Chubych, O., Connor, J., Ravishankar, R., & Shirawalmath, A. (2024). *From bytes to borsch: Fine-tuning Gemma and Mistral for the Ukrainian language representation.* arXiv preprint arXiv:2404.09138. https://doi.org/10.48550/arXiv.2404.09138

Laba, Y., Mudryi, V., Chaplynskyi, D., Romanyshyn, M., & Dobosevych, O. (2023). Contextual embeddings for Ukrainian: A large language model approach to word sense disambiguation. In M. Romanyshyn (Eds.), *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)* (pp. 11–19). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.unlp-1.2

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach.* arXiv preprint arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692
Nivre, J., Zeman, D., Ginter, F., & Tyers, F. (2017). Universal dependencies. In A. Klementiev & L. Specia (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts* (pp. 1–8). Association for Computational Linguistics. https://aclanthology.org/E17-5001
Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., & Ji, H. (2017). Cross-lingual name tagging and linking for 282 languages. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)* (pp. 1946–1958). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1178
Panchenko, D., Maksymenko, D., Turuta, O., Luzan, M., Tytarenko, S., & Turuta, O. (2022). Ukrainian news corpus as text classification benchmark. In O. Ignatenko, V. Kharchenko, V. Kobets, H. Kravtsov, Y. Tarasich, V. Ermolayev, D. Esteban, V. Yakovyna, & A. Spivakovsky (Eds.), *ICTERI 2021 Workshops. ICTERI 2021. Communications in Computer and Information Science* (Vol. 1635, pp. 550–559). Springer, Cham. https://doi.org/10.1007/978-3-031-14841-5_37
Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4996–5001). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1493
Piskorski, J., Marcińczuk, M., & Yangarber, R. (2024). Cross-lingual named entity corpus for Slavic languages. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 4143–4157). ELRA and ICCL.
Radchenko, V. (2020). *We trained the Ukrainian language model*. Youscan. https://youscan.io/blog/ukrainian-language-model/
Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using Siamese BERT-networks.* arXiv:1908.10084. https://doi.org/10.48550/arXiv.1908.10084
Ruder, S. (2022). *The state of multilingual AI.* ruder.io. https://www.ruder.io/stateof-multilingual-ai/
Rybak, P., Mroczkowski, R., Tracz, J., & Gawlik, I. (2020). KLEJ: Comprehensive benchmark for Polish language understanding. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1191–1201). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.111
Sanh, V., Wolf, T., Ly, J., & Rush, A. M. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv preprint arXiv:1910.01108. https://api.semanticscholar.org/CorpusID:203626972
Shaham, U., Aharoni, R., Rappoport, A., & Goldberg, Y. (2024). *Multilingual instruction tuning with just a pinch of multilinguality*. arXiv preprint arXiv:2401.01854. https://doi.org/10.48550/arXiv.2401.01854
Sido, J., Švec, J., Pecina, P., & Konopík, M. (2021). *Czert: Czech BERT-like model for language representation*. arXiv preprint arXiv:2103.13031. https://doi.org/10.48550/arXiv.2103.13031
Torge, S., Politov, A., Lehmann, C., Saffar, B., & Tao, Z. (2023). Named entity recognition for low-resource languages - Profiting from language families. *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)* (pp. 1–10), Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bsnlp-1.1
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. arXiv:1706.03762. https://doi.org/10.48550/arXiv.1706.03762
Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). *Multilingual is not enough: BERT for Finnish.* arXiv preprint arXiv:1912.07076. https://doi.org/10.48550/arXiv.1912.07076
Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019, May). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019) (poster session)* (pp. 353–355). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5446
Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019*). SuperGLUE: A stickier benchmark for general-purpose language understanding systems.* arXiv preprint arXiv:1905.00537. https://doi.org/10.48550/arXiv.1905.00537

References

Chaplynskyi, D., & Romanyshyn, M. (2024). Introducing NER-UK 2.0: A rich corpus of named entities for Ukrainian. In M. Romanyshyn, N. Romanyshyn, A. Hlybovets, & O. Ignatenko (Eds.), *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024* (pp. 23–29). ELRA and ICCL.
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423
GitHub. (2025). *Lang-Uk/ner-uk.* https://github.com/lang-uk/ner-uk
Haltiuk, M., & Smywiński-Pohl, A. (2024). LiBERTa: Advancing Ukrainian language modeling through pre-training from scratch (Presentation). In M. Romanyshyn, N. Romanyshyn, A. Hlybovets, & O. Ignatenko (Eds.), *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024* (pp. 120–128). ELRA and ICCL.
Hamotskyi, S., Levbarg, A.-I., & Hänig, C. (2024). Eval-UA-tion 1.0: Benchmark for evaluating Ukrainian (large) language models. In M. Romanyshyn, N. Romanyshyn, A. Hlybovets, & O. Ignatenko (Eds.), *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024* (pp. 109–119). ELRA and ICCL.
Hlybovets, A. M., & Bikchentaiev, M. (2022). Software system for checking plagiarism of Ukrainian texts [Prohramna systema perevirky na plahiat ukrainskykh tekstiv]. *Naukovi zapysky NaUKMA. Kompiuterni nauky*, 5, 16–25 [in Ukrainian]. https://doi.org/10.18523/2617-3808.2022.5.16-25
Hugging Face. (2024). *Uk_ner_web_trf_13class.* https://huggingface.co/dchaplinsky/uk_ner_web_trf_13class
Hugging Face. (2025). *Roberta-large-wechsel-ukrainian.* https://huggingface.co/benjamin/roberta-large-wechsel-ukrainian
Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560
Kiulian, A., Polishko, A., Khandoga, M., Chubych, O., Connor, J., Ravishankar, R., & Shirawalmath, A. (2024). *From bytes to borsch: Fine-tuning Gemma and Mistral for the Ukrainian language representation.* arXiv preprint arXiv:2404.09138. https://doi.org/10.48550/arXiv.2404.09138
Laba, Y., Mudryi, V., Chaplynskyi, D., Romanyshyn, M., & Dobosevych, O. (2023). Contextual embeddings for Ukrainian: A large language model approach to word sense disambiguation. In M. Romanyshyn (Eds.), *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)* (pp. 11–19). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.unlp-1.2
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach.* arXiv preprint arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692
Nivre, J., Zeman, D., Ginter, F., & Tyers, F. (2017). Universal dependencies. In A. Klementiev & L. Specia (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts* (pp. 1–8). Association for Computational Linguistics. https://aclanthology.org/E17-5001

ISSN 1728-3817

Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., & Ji, H. (2017). Cross-lingual name tagging and linking for 282 languages. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)* (pp. 1946–1958). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1178

Panchenko, D., Maksymenko, D., Turuta, O., Luzan, M., Tytarenko, S., & Turuta, O. (2022). Ukrainian news corpus as text classification benchmark. In O. Ignatenko, V. Kharchenko, V. Kobets, H. Kravtsov, Y. Tarasich, V. Ermolayev, D. Esteban, V. Yakovyna, & A. Spivakovsky (Eds.), *ICTERI 2021 Workshops. ICTERI 2021. Communications in Computer and Information Science* (Vol. 1635, pp. 550–559). Springer, Cham. https://doi.org/10.1007/978-3-031-14841-5_37

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4996–5001). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1493

Piskorski, J., Marcińczuk, M., & Yangarber, R. (2024). Cross-lingual named entity corpus for Slavic languages. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 4143–4157). ELRA and ICCL.

Radchenko, V. (2020). *We trained the Ukrainian language model*. Youscan. https://youscan.io/blog/ukrainian-language-model/

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using Siamese BERT-networks.* arXiv:1908.10084. https://doi.org/10.48550/arXiv.1908.10084

Ruder, S. (2022). *The state of multilingual AI.* ruder.io. https://www.ruder.io/stateof-multilingual-ai/

Rybak, P., Mroczkowski, R., Tracz, J., & Gawlik, I. (2020). KLEJ: Comprehensive benchmark for Polish language understanding. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1191–1201). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.111

Sanh, V., Wolf, T., Ly, J., & Rush, A. M. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv preprint arXiv:1910.01108. https://api.semanticscholar.org/CorpusID:203626972

Shaham, U., Aharoni, R., Rappoport, A., & Goldberg, Y. (2024). *Multilingual instruction tuning with just a pinch of multilinguality*. arXiv preprint arXiv:2401.01854. https://doi.org/10.48550/arXiv.2401.01854

Sido, J., Švec, J., Pecina, P., & Konopík, M. (2021). *Czert: Czech BERT-like model for language representation*. arXiv preprint arXiv:2103.13031. https://doi.org/10.48550/arXiv.2103.13031

Torge, S., Politov, A., Lehmann, C., Saffar, B., & Tao, Z. (2023). Named entity recognition for low-resource languages – Profiting from language families. *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)* (pp. 1–10), Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bsnlp-1.1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. arXiv:1706.03762. https://doi.org/10.48550/arXiv.1706.03762

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). *Multilingual is not enough: BERT for Finnish.* arXiv preprint arXiv:1912.07076. https://doi.org/10.48550/arXiv.1912.07076

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019, May). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019) (poster session)* (pp. 353–355). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5446

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019*). SuperGLUE: A stickier benchmark for general-purpose language understanding systems.* arXiv preprint arXiv:1905.00537. https://doi.org/10.48550/arXiv.1905.00537

Отримано редакцією журналу / Received: 19.08.25
Прорецензовано / Revised: 24.01.26
Схвалено до друку / Accepted: 16.04.26
Опубліковано / Published: 05.06.26

Mykola GLYBOVETS, DSc (Phys. & Math.), Prof.
ORCID ID: 0009-0005-6942-8026
e-mail: glib@ukma.edu.ua
National University of Kyiv-Mohyla Academy, Kyiv, Ukraine

Danylo VANIN, Master of Arts
ORCID ID: 0009-0007-4510-8209
e-mail: danylo.vanin@gmail.com
National University of Kyiv-Mohyla Academy, Kyiv, Ukraine

THE APPLICATION OF MONOLINGUAL AND MULTILINGUAL BERT-BASED MODELS FOR TEXT AUTOMATION TASKS IN UKRAINIAN

The object of this study is mono- and multilingual models based on the BERT architecture. The research focuses on comparing their performance in natural language processing (NLP) tasks, with an emphasis on their application to the Ukrainian language. The methodological framework follows standard approaches to model training and evaluation, relying on publicly available information sources. Experiments were conducted solely on open-source or self-fine-tuned models and covered all existing Ukrainian benchmarks.

Overall, the findings show that both monolingual and multilingual BERT models can be effective for NLP tasks, depending on the target language, specific task, and available resources. While monolingual models often outperform multilingual ones for tasks in their own language, multilingual models can prove advantageous when resources for training language-specific models are limited.

The comparison demonstrates that models trained from scratch on Ukrainian texts–or further fine-tuned on them–generally achieve higher results on Ukrainian NLP tasks than multilingual counterparts. This advantage was evident in named-entity recognition, text classification, and masked-token prediction. Multilingual models, however, achieve competitive or superior performance in "lighter" tasks such as word-sense disambiguation, especially when full training of a monolingual model is not feasible.

These findings underscore the need to adapt or fully train Ukrainian-language models for specialized tasks while still allowing for the cost-effective use of multilingual systems in everyday scenarios. The practical significance of this work lies in providing a consolidated performance map of modern models and laying the groundwork for a full-scale Ukrainian GLUE-like benchmark, which will stimulate further research and improve the accuracy of Ukrainian NLP applications.

K e y w o r d s : natural language processing, large language models, monolingual and multilingual models, BERT.

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.