УДК 004.89:004.912
DOI: 10.18523/2617-3808.2025.8.76-83

Ванін Д. О.

ВИКОРИСТАННЯ ОДНО- ТА БАГАТОМОВНИХ МОДЕЛЕЙ НА БАЗІ BERT ДЛЯ ВИРІШЕННЯ ЗАДАЧ АВТОМАТИЧНОГО ОБРОБЛЕННЯ ТЕКСТІВ

Об'єктом дослідження цієї статті є одно- та багатомовні моделі на основі BERT. Предметом дослідження було порівняння продуктивності таких моделей на завданнях ОПМ із наголосом на їх застосуванні для української мови. Методологічну основу порівняльного аналізу становило використання стандартних підходів до навчання та оцінки моделей. У дослідженні використовувались доступні джерела інформації.

Загалом результати дослідження свідчать про те, що як одномовні, так і багатомовні моделі на основі BERT можуть бути ефективними для вирішення 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. Проведене порівняння роботи одно- та багатомовних моделей для різних мов додатково підкреслило важливість проведення окремого порівняння їх застосування для української мови.

Проведений аналіз сприятиме створенню комплексного україномовного бенчмарку, що покращить якість моделей і стимулюватиме нові дослідження у галузі ОПМ для української мови, розроблення нових, більш ефективних моделей.

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

Вступ

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

Сучасні моделі зазвичай тренуються у два етапи: базове навчання на великих нерозмічених мовних корпусах і донавчання на менших наборах даних, специфічних для завдання. У результаті базового тренування моделі навчаються «розуміти» мову та передбачати слово, яке найкраще відповідає контексту. Якість результату цього навчання напряму залежить від якості й, найважливіше, розміру навчальних даних (сучасні моделі тренуються на терабайтах нерозмічених даних). Для популярних мов, наприклад англійської чи китайської, великі мовні корпуси відшукати достатньо легко, тимчасом як для мов із 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 із обмеженою кількістю ресурсів для навчання багатомовні моделі одразу показували передові на той час результати. Причина цього — 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 — через схожі слова, абетки, будову речення, ідіоми та запозичення між мовами. Отже, багатомовні моделі можуть перевикористовувати певні «знання» про мови з великими ресурсами на мовах з обмеженими ресурсами.

У науковій літературі багато вивчають можливості одно- і багатомовних моделей. Деякі дослідження демонструють, що одномовні моделі перевершують багатомовні на багатьох завданнях однієї мови. Наприклад, дослідження [17] 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 і показувати кращі результати, ніж одномовні для тих мов, де ресурси обмежені (наприклад у випадку португальської мови).

Об'єктом дослідження цієї статті є одно- та багатомовні моделі на основі BERT. Предметом дослідження було порівняння продуктивності таких моделей на завданнях ОПМ із наголосом на їх застосуванні для української мови. Методологічну основу порівняльного аналізу становило використання стандартних підходів до навчання та оцінки моделей. У дослідженні використовувались доступні джерела інформації.

Дослідження має як наукове, так і практичне значення. З наукового погляду воно допомагає краще розуміти компроміс між використанням одно- та багатомовних моделей і 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 узяти як основу для створення комплексного бенчмарку оцінки моделей для української мови.

1. Основні поняття та моделі ОПМ

Оброблення природної мови (ОПМ) передбачає безліч завдань, спрямованих на те, щоб інтелектуальні програмні системи могли розуміти, інтерпретувати та генерувати людську мову. Серед основних завдань, що вирішуються у сфері ОПМ, виділяють такі:

- класифікація тексту (Text Classification), що охоплює аналіз настроїв (Sentiment Analysis), класифікацію тем (Topic Classification), виявлення спаму (Spam Detection), визначення наміру (Intent Detection);
- маркування послідовностей (Sequence Labeling) з наголосом на розпізнавання іменованих сутностей (Named Entity Recognition, NER), розмічування частин мови (Part-of-Speech Tagging), групування (Chunking, Shallow Parsing), визначення параметрів (Slot Filling);
- генерація мови (Language Generation) з підзадачами генерації тексту (Text Generation), машинного перекладу (Machine Translation), стислого викладу (Summarization), перефразування (Paraphrasing), генерації діалогів (Dialogue Generation);
- порівняння тексту (Text Comparison), що охоплює підзадачі схожості тексту (Text Similarity), визначення перефразування (Paraphrase Identification), оцінки семантичної схожості тексту (Semantic Textual Similarity);
- інформаційний пошук (Information Retrieval);
- аналіз тексту (Text Analysis);
- розуміння мови (Language Understanding).

Моделі оброблення 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 їхню здатність виконувати конкретні завдання в обробленні природної мови.

Першим етапом у тренуванні є збирання даних та попереднє оброблення. Основою для будь-якої моделі ОПМ є великий (терабайти даних) і різноманітний (різні жанри, джерела, формати) набір текстових даних. Такі дані можна отримати з книжок, статей, вебсайтів, 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 їхніх частин (наприклад, символи чи фрагменти слів). Для цього використовуються такі методи, як вкладання слів (Word2Vec, GloVe) або більш складні моделі на основі архітектури трансформерів (BERT, GPT), що перетворюють слова на вектори у багатомірному просторі. Ці вектори захоплюють семантичні зв'язки між словами, які схожі за сенсом і будуть мати високий косинус подібності між векторами-репрезентаціями.

Вибір архітектури моделі залежить від конкретного завдання ОПМ. Наприклад, для нескладних завдань на зразок «послідовність — послідовність» (машинний переклад) можуть бути достатніми рекурентні нейронні мережі (RNN). Завдання класифікації спаму може бути вирішене і наївним баєсовим класифікатором. Поширена зараз трансформерна архітектура стала дуже популярною завдяки універсальності й здатності навчатися виявляти зв'язки між елементами тексту, розташованими на різних відстанях. Вони вирішують проблему згасання важливості контексту й 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 становить модель, поступово змінюються, щоб досягнути глобального чи локального мінімуму. Часто завданням на етапі базового тренування є просте передбачення слова у контексті. Наприклад, модель навчається передбачувати «замасковане» слово у реченні. Для такого завдання не потрібні розмічені дані, і для навчання потенційно можна скористатися будь-яким якісним текстом на цільовій мові.

Процес тонкого налаштування

Попри те, що моделі, такі як BERT, уже 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 складного завдання, як-от відповідь на питання чи переказ тексту, модель має бути донавчена на розмічених даних. Ці дані зазвичай мають значно менший обсяг і 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 дає змогу «перевикористати» знання, отримані з великого неструктурованого набору даних під час попереднього навчання. Додавання нових даних, які близькі до кінцевого завдання, допомагає налаштувати модель для вирішення конкретних задач.

Залежно від завдання може виникнути потреба в незначних змінах архітектури моделі. Наприклад, для аналізу настроїв можна додати до моделі 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 конкретних проблем.

Тонке налаштування передбачає оптимізацію гіперпараметрів, як-от швидкість навчання, розмір пакета даних і кількість епох навчання. Це дозволяє досягти найкращої продуктивності на специфічних для завдання даних і також визначається емпірично.

Одномовний BERT

Одномовний BERT (Bidirectional Encoder Representations from Transformers — двонаправлене кодування представлень із трансформерів) — це мовна модель, яка використовує трансформерну архітектуру. Вона складається з шарів самоуваги (self-attention) і 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 [2].

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 із мінімальними модифікаціями, специфічними для завдання.

Після виходу BERT показав одні з найкращих результатів в 11 завданнях оброблення природної мови [2] і став основою для численних майбутніх моделей, таких як RoBERTa [21], DistilBERT [7] та інші.

Багатомовний BERT

Розроблені багатомовні варіанти BERT, наприклад mBERT і XLM-RoBERTa [24], 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 із низьким рівнем ресурсів, таких як українська [24].

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. Вони демонструють високу продуктивність на мовах із великою кількістю ресурсів (наприклад, англійська, китайська, іспанська), однак їхні результати на мовах із низькими ресурсами

може бути обмежено. Базова причина цього — недостатня репрезентованість розуміння мови у вагах моделі. Чим більшу частку становлять дані певної мови в тренувальному наборі, тим кращими будуть результати моделі для цієї мови.

Хоча моделі BERT досягли значних успіхів на популярних мовах, усе ще існують виклики для мов із недостатніми ресурсами. Обмежена доступність якісних і 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.

2. Багато- та одномовність у мовних моделях

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

Багатомовні моделі, навчені на даних із кількох мов, продемонстрували великі можливості. Вони можуть досягати передових результатів у широкому діапазоні мов, зберігаючи при цьому високу продуктивність у мовах із великими ресурсами [14]. Такий «крос-лінгвістичний трансфер» дозволяє моделям використовувати знання, отримані в одній мові, для інших мов, що особливо корисно для мов з обмеженими ресурсами [24].

Автори [16] показали, що багатомовні моделі можуть досягати наближеної або навіть вищої продуктивності на декількох мовах порівняно з одномовними моделями. Цікаво, що такі результати досягаються, незважаючи на навчання на значно меншій кількості даних у цих конкретних мовах.

Водночас, одномовні моделі часто можуть перегнати свої багатомовні аналоги у деяких завданнях, зокрема у виявленні мови ворожнечі та генеруванні речень [19; 6]. 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 [22]. Важливим також є тип даних, на яких тренувалася модель. Багатомовні моделі часто навчають на даних із Вікіпедії або онлайн-видавництв. Дослідники, які тренують одномовні моделі, можуть приділити більше часу на збирання 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 кожних завданні та мові [13].

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 їхніми одномовними аналогами [13]. 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 [22]. Прикладом є модель Poro 34B із 34 мільярдами параметрів, навчена на фінській, англійській і мовах програмування, яка значно перевершує наявні одномовні моделі для менш репрезентованих мов, таких як фінська [20]. 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 [5]. Це підкреслює необхідність ретельної оцінки та зменшення упереджень у багатомовних моделях.

Зрештою, вибір між одномовними та багатомовними моделями залежить від конкретного завдання, доступних ресурсів і цільових мов. У деяких випадках, як із португальською мовою, обидва типи моделей можуть досягати порівнянних результатів, що робить вибір менш критичним [8]. Однак в інших випадках необхідно ретельно зважити компроміс між крос-лінгвістичним трансфером і спеціалізацією на конкретній мові. Порівняння результатів одномовних та багатомовних моделей для української мови може вказати, у яких напрямах дослідники повинні приділити більше уваги для розроблення якісних україномовних наборів даних і моделей.

3. ОПМ української мови та виклики

Українська мова, що належить до східнослов’янської мовної сім’ї індоєвропейських мов [10], має багату морфологічну структуру, характерну для синтетичних мов [4]. Із понад 40 мільйонами носіїв, вона є однією з найбільш поширених слов’янських мов [4]. Проте українська мова стикається зі

значними викликами в галузі оброблення природної мови, перш за все через нестачу даних для тренування. Через це її класифікують як мову з низьким рівнем ресурсів [4]. У 2020 році українську мову визначили як "rising star" (рис. 1.) — мову, що розвивається з культурною спільнотою, але страждає від нестачі розмічених наборів даних [12]. Українська мова належить до кластера 3 за розподілом мовних ресурсів (кількість розмічених і нерозмічених даних) [23].

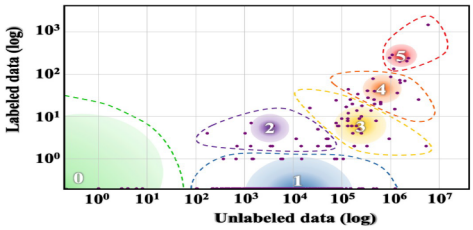

Рис. 1. Графік кластеризації мов [23]

Основною перешкодою для ОПМ української мови є нестача загальнодоступних, орієнтованих на конкретні завдання корпусів [4]. Це історично заважало розвитку надійних одномовних моделей для української мови та змушувало дослідників покладатися на крос-мовне трансферне навчання з інших мов [11]. На відміну від мов із 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 [11].

Хоча українська мова має певні лінгвістичні зв'язки з іншими східнослов'янськими мовами [10], пряме трансферне навчання з цих мов не завжди є оптимальним через відмінності у лексиці, граматиці та культурному контексті. Тому розроблення спеціалізованих моделей і ресурсів ОПМ, адаптованих до унікальних особливостей української мови, є важливим для подолання цих викликів та розкриття повного потенціалу українських ОПМ застосунків.

Навчання моделей на корпусі з декількох подібних мов теоретично може показати кращі результати, ніж навчання на одномовному корпусі, у випадку нестачі одномовних ресурсів. У наступному нашому дослідженні ми плануємо перевірити гіпотезу, що модель, натренована на корпусі слов'янських мов, може показати кращі результати, ніж одномовна модель.

4. Аналіз досліджень порівняння одно- та багатомовних моделей

Декілька досліджень розглядали продуктивність одномовних та багатомовних моделей на основі BERT у різних завданнях оброблення 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) [17] показали, що одномовна фінська модель BERT (FinBERT) перевершує багатомовну модель BERT (mBERT) у різних фінських завданнях ОПМ. Деякі дослідження виявили, що мовно-специфічні моделі BERT перевершують mBERT у розпізнаванні іменованих сутностей (NER) у слов'янських мовах. Velankar та ін. (2022) [25] 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) [18] 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) [8] виявили, що багатомовна модель BERT незначно перевершує одномовні португальські моделі у різних завданнях NLP. Sido та ін. (2021) [6] 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 і трансферного навчання. Yang та ін. (2023) [15] показали, що тримовна модель GPT-3, налаштована на англійських даних, може успішно виконувати інструкції угорською та китайською мовами, що свідчить про потенціал міжмовного транс-

ферного навчання. Vīksna та Skadina (2021) [26] виявили, що незважаючи на доповнення даних, багатомовна модель BERT не покращила показники в розпізнаванні іменованих сутностей у шести слов'янських мовах.

Висновки

Загалом результати дослідження свідчать про те, що як одномовні, так і багатомовні моделі на основі BERT можуть бути ефективними для вирішення 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. Проведене порівняння роботи одно- та багатомовних моделей для різних мов додатково підкреслило важливість проведення окремого порівняння їх застосування для української мови.

Ця робота має на меті зробити внесок у поточну дискусію, порівнюючи продуктивність одномовних і багатомовних моделей на основі BERT для різних завдань ОПМ. 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 [11; 9]. Крім того, конференція відзначилася розширенням бенчмарків для оцінки моделей [12; 3]. У рамках цієї роботи ми здійснили спробу інтегрувати деякі з цих результатів.

Проведений аналіз сприятиме створенню комплексного україномовного бенчмарку, що має покращити якість моделей і стимулюватиме нові дослідження у галузі ОПМ для української мови, розроблення нових, більш ефективних моделей.

У рамках подальшого дослідження вважаємо перспективним порівняти результати одно- та багатомовних моделей типу BERT на завданнях оброблення української мови.

Список літератури

1. Attention is all you need / Ashish Vaswani [et al.] // Proceedings of the 31st international conference on neural information processing systems. — Red Hook, NY, USA, 2017. — Pp. 6000–6010.
2. BERT: pre-training of deep bidirectional transformers for language understanding [Electronic resource] / Jacob 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. — Pp. 4171–4186. — Mode of access: https://aclanthology.org/N19-1423 (date of access: 23.05.2024). — Title from screen.
3. Chaplynskyi D. Introducing NER-UK 2.0: A Rich Corpus of Named Entities for Ukrainian [Electronic resource] / Dmytro Chaplynskyi, Mariana Romanyshyn // Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024, Torino / ed. by M. Romanyshyn et al. — [S. l.], 2024. — Pp. 23–29. — Mode of access : https://aclanthology.org/2024.unlp-1.4 (date of access: 31.05.2024). — Title from screen.
4. Chaplynskyi D. Introducing ubertext 2.0: a corpus of modern Ukrainian at scale [Electronic resource] / Dmytro Chaplynskyi // Proceedings of the second Ukrainian natural language processing workshop (UNLP), Dubrovnik / ed. by M. Romanyshyn. — [S. l.], 2023. — Pp. 1–10. — Mode of access: https://aclanthology.org/2023.unlp-1.1 (date of access: 28.05.2024). — Title from screen.
5. Comparing biases and the impact of multilingual training across multiple languages [Electronic resource] / Sharon Levy [et al.] // Association for computational linguistics. — 2023. — Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. — Pp. 10260–10280. — Mode of access: https://aclanthology.org/2023.emnlp-main.634 (date of access: 23.05.2024). — Title from screen.
6. Czert — czech bert-like model for language representation [Electronic resource] / Jakub Sido [et al.]. — [S. l. : s. n.], 2021. — Mode of access: https://doi.org/10.48550/arXiv.2103.13031 (date of access: 28.05.2024). — Title from screen.
7. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [Electronic resource] / Victor Sanh [et al.] // ArXiv. — 2019. — Abs/1910.01108. — Mode of access: https://api.semanticscholar.org/CorpusID:203626972. — Title from screen.
8. Feijo D. de V. Mono vs multilingual transformer-based models: a comparison across several language tasks [Electronic resource] / Diego de Vargas Feijo, Viviane Pereira Moreira. — [S. l. : s. n.], 2020. — Mode of access: https://doi.org/10.48550/arXiv.2007.09757 (date of access: 28.05.2024). — Title from screen.
9. From bytes to borsch: fine-tuning gemma and mistral for the Ukrainian language representation [Electronic resource] / Artur Kiulian [et al.]. — [S. l. : s. n.], 2024. — Mode of access: https://doi.org/10.48550/arXiv.2404.09138 (date of access: 28.05.2024). — Title from screen.
10. Gomez F. P. A Low-Resource Approach to the Grammatical Error Correction of Ukrainian [Electronic resource] / Frank Palma Gomez, Alla Rozovskaya, Dan Roth // Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), Dubrovnik / ed. by M. Romanyshyn. — [S. l.]. — Pp. 114–120. — Mode of access: https://aclanthology.org/2023.unlp-1.14 (date of access: 28.05.2024). — Title from screen.
11. Haltiuk M. LiBERTa: Advancing Ukrainian Language Modeling through Pre-training from Scratch / Mykola Haltiuk, Aleksander Smywiński-Pohl // Unlp 2024, Turin. — [S. l.], 2024.
12. Hamotskyi S. Eval-UA-tion 1.0: benchmark for evaluating Ukrainian (large) language models [Electronic resource] / Serhii Hamotskyi, Anna-Izabella Levbarg, Christian Hänig // Unlp 2024, Turin. — [S. l.], 2024. — Mode of access: https://hal.science/hal-04534651 (date of access: 28.05.2024). — Title from screen.
13. How good is your tokenizer? On the monolingual performance of multilingual language models [Electronic resource] / Phillip 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.]. — [S. l.], 2021. — Pp. 3118–3135. — Mode of access: https://aclanthology.org/2021.acl-long.243 (date of access: 28.05.2024). — Title from screen.
14. Larger-Scale transformers for multilingual masked language modeling / Naman Goyal [et al.]. — [S. l. : s. n.], 2021.
15. Mono- and multilingual GPT-3 models for hungarian [Electronic resource] / Zijian Yang [et al.]. — [S. l.], 2023. — Pp. 94–104. — Mode of access: https://doi.org/10.1007/978-3-031-40498-6_9 (date of access: 28.05.2024). — Title from screen.
16. Multilingual instruction tuning with just a pinch of multilinguality [Electronic resource] / Uri Shaham [et al.]. — [S. l.], 2024. — Mode of access: https://doi.org/10.48550/arXiv.2401.01854 (date of access: 28.05.2024). — Title from screen.
17. Multilingual is not enough: BERT for Finnish [Electronic resource] / Antti Virtanen [et al.]. — [S. l. : s. n.], 2019. — Mode of access: https://doi.org/10.48550/arXiv.1912.07076 (date of access: 28.05.2024). — Title from screen.
18. Named entity recognition for low-resource languages — profiting from language families [Electronic resource] / Sunna Torge [et al.] // Proceedings of the 9th workshop on slavic natural language processing 2023 (slavicnlp 2023), Dubrovnik / ed. by J. Piskorski [et al.]. — [S. l.]. — Pp. 1–10. — Mode of access: https://aclanthology.org/2023.bsnlp-1.1 (date of access: 28.05.2024). — Title from screen.
19. Pires T. How multilingual is multilingual BERT [Electronic resource] / Telmo Pires, Eva Schlinger, Dan Garrette. — 2019. — Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. — Pp. 4996–5001. — Mode of access: https://aclanthology.org/P19-1493 (date of access: 23.05.2024). — Title from screen.
20. Poro 34B and the blessing of multilinguality [Electronic resource] / Risto Luukkonen [et al.]. — [S. l. : s. n.], 2024. — Mode of access: https://doi.org/10.48550/arXiv.2404.01856 (date of access: 28.05.2024). — Title from screen.
21. RoBERTa: A robustly optimized BERT pretraining approach [Electronic resource] / Yinhan Liu [et al.] // CoRR. — 2019. — Abs/1907.11692. — Mode of access: http://arxiv.org/abs/1907.11692. — Title from screen.
22. Ruder S. The state of multilingual AI [Electronic resource] / Sebastian Ruder // ruder.io. — Mode of access: https://www.ruder.io/state-of-multilingual-ai/ (date of access: 23.05.2024). — Title from screen.
23. The State and Fate of Linguistic Diversity and Inclusion in the NLP World [Electronic resource] / Pratik Joshi [et al.] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics / ed. by D. Jurafsky [et al.]. — [S. l.], 2020. — Pp. 6282–6293. — Mode of access: https://aclanthology.org/2020.acl-main.560 (date of access: 31.05.2024). — Title from screen.
24. Unsupervised cross-lingual representation learning at scale / Alexis Conneau [et al.]. — [S. l. : s. n.], 2020.
25. Velankar A. Mono vs multilingual BERT for hate speech detection and text classification: a case study in marathi / Abhishek Velankar, Hrushikesh Patil, Raviraj Joshi // Artificial neural networks in pattern recognition / ed. by N. El Gayar [et al.]. — Cham, 2023. — Pp. 121–128.
26. Vīksna R. Multilingual slavic named entity recognition [Electronic resource] / Rinalds Vīksna, Inguna Skadina // Proceedings of the 8th workshop on balto-slavic natural language processing, Kiyv. — [S. l.]. — Pp. 93–97. — Mode of access: https://aclanthology.org/2021.bsnlp-1.11 (date of access: 28.05.2024). — Title from screen.

References

Chaplynskyi, D. (2023). Introducing ubertext 2.0: a corpus of modern Ukrainian at scale. In M. Romanyshyn (Ed.), *Proceedings of the second ukrainian natural language processing workshop (UNLP)* (pp. 1–10). Association for Computational Linguistics. https://aclanthology.org/2023.unlp-1.1.

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. https://aclanthology.org/2024.unlp-1.4.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). *Unsupervised cross-lingual representation learning at scale*.

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, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. https://aclanthology.org/N19-1423.

Feijo, D. de V., & Moreira, V. P. (2020). *Mono vs multilingual transformer-based models: a comparison across several language tasks*. https://doi.org/10.48550/arXiv.2007.09757.

Gomez, F. P., Rozovskaya, A., & Roth, D. (N. d.). A Low-Resource Approach to the Grammatical Error Correction of Ukrainian. In M. Romanyshyn (Ed.), *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)* (pp. 114–120). Association for Computational Linguistics. https://aclanthology.org/2023.unlp-1.14.

Goyal, N., Du, J., Ott, M., Anantharaman, G., & Conneau, A. (2021). *Larger-Scale transformers for multilingual masked language modeling*.

Haltiuk, M., & Smywiński-Pohl, A. (2024). LiBERTa: Advancing Ukrainian Language Modeling through Pre-training from Scratch. In *Unlp 2024*.

Hamotskyi, S., Levbarg, A.-I., & Hänig, C. (2024). Eval-UA-tion 1.0: benchmark for evaluating Ukrainian (large) language models. In *Unlp 2024*. https://hal.science/hal-04534651.

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://aclanthology.org/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*. https://doi.org/10.48550/arXiv.2404.09138.

Levy, S., John, N., Liu, L., Vyas, Y., Ma, J., Yoshinari, F., Ballesteros, M., Castelli, V., & Roth, D. (2023). Comparing biases and the impact of multilingual training across multiple languages. In *Association for computational linguistics*, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10260–10280. https://aclanthology.org/2023.emnlp-main.634.

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. *CoRR*, *abs/1907.11692*. http://arxiv.org/abs/1907.11692.

Luukkonen, R., Burdge, J., Zosa, E., Talman, A., Komulainen, V., Hatanpää, V., Sarlin, P., & Pyysalo, S. (2024). *Poro 34B and the blessing of multilinguality*. https://doi.org/10.48550/arXiv.2404.01856.

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://aclanthology.org/P19-1493.

Ruder, S. (2022, 14 листопада). The state of multilingual AI. *ruder.io*. https://www.ruder.io/state-of-multilingual-ai/.

Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., & Gurevych, I. (2021). How good is your tokenizer? On the monolingual performance of multilingual language models. In C. Zong, F. Xia, R. Navigli & W. Li (Eds.), *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)* (pp. 3118–3135). Association for Computational Linguistics. https://aclanthology.org/2021.acl-long.243.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, *abs/1910.01108*. https://api.semanticscholar.org/CorpusID:203626972.

Shaham, U., Herzig, J., Aharoni, R., Szpektor, I., Tsarfaty, R., & Eyal, M. (2024). *Multilingual instruction tuning with just a pinch of multilinguality*. https://doi.org/10.48550/arXiv.2401.01854.

Sido, J., Pražák, O., Přibáň, P., Pašek, J., Seják, M., & Konopík, M. (2021). *Czert — czech bert-like model for language representation*. https://doi.org/10.48550/arXiv.2103.13031/

Torge, S., Politov, A., Lehmann, C., Saffar, B., & Tao, Z. (N. d.). Named entity recognition for low-resource languages - profiting from language families. In J. Piskorski, M. Marcińczuk, P. Nakov, M. Ogrodniczuk, S. Pollak, P. Přibáň, P. Rybak, J. Steinberger & R. Yangarber (Eds.), *Proceedings of the 9th workshop on slavic natural language processing 2023 (slavicnlp 2023)* (pp. 1–10). Association for Computational Linguistics. https://aclanthology.org/2023.bsnlp-1.1.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 6000–6010). Curran Associates Inc.

Velankar, A., Patil, H., & Joshi, R. (2023). Mono vs multilingual BERT for hate speech detection and text classification: a case study in marathi. In N. El Gayar, E. Trentin, M. Ravanelli & H. Abbas (Eds.), *Artificial neural networks in pattern recognition* (pp. 121–128). Springer International Publishing.

Vīksna, R., & Skadina, I. (N. d.). Multilingual slavic named entity recognition. In *Proceedings of the 8th workshop on balto-slavic natural language processing* (pp. 93–97). Association for Computational Linguistics. https://aclanthology.org/2021.bsnlp-1.11.

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

Yang, Z., Laki, L., Váradi, T., & Prószéky, G. (2023). *Mono- and multilingual GPT-3 models for hungarian* (pp. 94–104). https://doi.org/10.1007/978-3-031-40498-6_9.

D. Vanin

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

This article investigates monolingual and multilingual BERT-based Transformer models. Its primary aim is to compare the behaviour of these models on a set of natural-language-processing (NLP) problems, with particular attention to their application to Ukrainian. The analysis is grounded in standard practice: large-scale masked-language pre-training, supervised fine-tuning, and evaluation on public, freely available corpora.

Five representative NLP tasks are examined—document classification, sentiment analysis, named-entity recognition, part-of-speech tagging, and sentence-level semantic similarity—because they cover core linguistic phenomena and underpin most applied pipelines. All checkpoints are trained and tested in identical experimental settings, an approach that is especially important for Ukrainian, which remains a low-resource language.

The results show that both model families are capable of solving the selected tasks, yet each excels under different conditions. Monolingual checkpoints deliver higher accuracy on problems that hinge on fine morpho-syntactic detail, such as handling case endings or varied word order. Multilingual checkpoints, in turn, offer a cost-effective solution when Ukrainian training data are scarce: knowledge transferred from related Slavic languages helps maintain acceptable quality while lowering annotation effort.

By clarifying when each strategy is preferable, the article provides practitioners with a concise decision framework and argues for the creation of a unified open benchmark to track progress. Such infrastructure will raise overall model quality, stimulate new Ukrainian-language research, and accelerate the development of more effective, resource-aware NLP technologies.

Keywords: natural language processing, large language models, monolingual and multilingual models, BERT.

Матеріал надійшов 23.06.2025