УДК 004.75 **DOI:** https://doi.org/10.30837/bi.2026.1(104).09

Яна Данієль[1], Даніїл Максименко[2], Дмитро Панченко[3], Ольга Калиниченко[4], Олексій Турута[5]

[1]ХНУРЕ, м. Харків, Україна, yana.daniiel@nure.ua, ORCID iD: 0000-0002-3895-0744

[2]ХНУРЕ, м. Харків, Україна, daniil.maksymenko@nure.ua, ORCID iD: 0000-0003-3223-5130

[3]ХНУРЕ, м. Харків, Україна, dmytro.panchenko@nure.ua м. Харків, ORCID iD: 0000-0001-5454-5661

[4]ХНУРЕ, м. Харків, Україна, olga.kalynychenko@nure.ua, ORCID iD: 0000-0003-1466-3967

[5]ХНУ ім. Каразіна, м. Харків, Україна, Oleksii.Turuta@karazin.ua, ORCID iD: 0000-0002-0970-8617

НАБІР ДАНИХ УКРАЇНСЬКИХ НОВИН ЯК БЕНЧМАРК ДЛЯ КЛАСИФІКАЦІЇ ТЕКСТІВ

У статті розглянуто задачу створення україномовних наборів даних для класифікації текстів. Запропоновано підхід для створення простого набору даних. Також створено корпус українських новин, придатний для порівняльного оцінювання моделей. На цьому наборі даних проведено бенчмаркінг сучасних моделей на основі трансформерів (mBERT, Slavic BERT, ukr-RoBERTa, ukr-ELECTRA, XLM-R) та базової моделі NB-SVM у різних режимах навчання. Результати показують, що ukr-RoBERTa, ukr-ELECTRA та XLM-R демонструють найвищу якість. XLM-R, як правило, краще працює з об'ємними текстами, тоді як ukr-RoBERTa – з більш короткими послідовностями.

ОБРОБКА УКРАЇНСЬКОЇ МОВИ, КЛАСИФІКАЦІЯ ТЕКСТУ, ТРАНСФОРМЕРИ, ТЕКСТОВИЙ НАБІР ДАНИХ

Y. Daniiel, D. Maksymenko, D. Panchenko, O. Kalynychenko, O.Turuta. Text Classification Benchmark of Ukrainian News Dataset. The problem of the lack of high-quality benchmark datasets for Ukrainian text classification is considered. A framework for creating a simple dataset with minimal annotation effort is proposed. In addition, a corpus of Ukrainian news suitable for comparative model evaluation is created. Using this dataset, a benchmarking study of modern transformer-based models (mBERT, Slavic BERT, ukr-RoBERTa, ukr-ELECTRA, XLM-R) as well as a baseline NB-SVM model is conducted under various training settings. The results show that ukr-RoBERTa, ukr-ELECTRA, and XLM-R achieve the best performance. XLM-R generally performs better on long texts, whereas ukr-RoBERTa is more effective for shorter sequences.

UKRAINIAN LANGUAGE PROCESSING, TEXT CLASSIFICATION, TRANSFORMERS, TEXT DATASET

Вступ

Останнім часом обробка природної мови пережила фазу швидкого розвитку, аналогічну до революції комп'ютерного зору у 2010-х роках. Цей прогрес здебільшого пов'язаний з розвитком архітектур Transformer [2] та BERT [3]. Таких успіхів вдалося досягти переважно завдяки механізму трансферного навчання.

Однак, попереднє навчання таких моделей потребує великої кількості даних та обчислювальної потужності. В результаті більшість найкращих попередньо навчених архітектур існують лише для найпопулярніших мов, таких як англійська, китайська тощо. Для української мови є лише кілька подібних моделей, зокрема ukr-RoBERTa [4] та ukr-ELECTRA [5]. Однак обидві ці моделі не мають належного оцінювання: ukr-ELECTRA орієнтована тільки на POS-теги та завдання NER, тоді як ukr-RoBERTa взагалі не має жодних показників, розрахованих на загальнодоступних наборах даних.

Ця відсутність різноманітних оцінок, яка зазвичай властива науковим роботам з NLP для англійської мови, є результатом недостатньої кількості загальнодоступних, правильно організованих та очищених наборів даних для української мови. Науковцями було зроблено кілька спроб створити еталонний набір даних для найпоширенішого та найпростішого завдання обробки природної мови — класифікації тексту.

В [6] і [7] автори створюють набір даних для аналізу тональності відгуків про готелі, який може бути еталоном для текстових класифікаторів української мови. Вони також відзначають відсутність українських даних та складність їх збору. Наприкінці автори доповнюють свій набір російськими текстами, перекладеними українською за допомогою алгоритмів машинного перекладу.

Інший відомий ресурс з українськими наборами даних [8] пропонує велику колекцію нерозмічених даних, а також набори даних та попередньо навчені моделі для розпізнавання іменованих сутностей (NER). Проте в ньому відсутні українські набори даних для завдань «послідовність до одного».

В цій роботі запропоновано альтернативний підхід до розв'язання задач обробки української мови. Він полягає у використанні багатомовних моделей. Існують два трансформери, навчені різними мовами, включаючи українську: Multilingual BERT [3] та XLM-R [9]. Такі моделі зазвичай навчаються на комбінованому корпусі, що включає тексти на десятках мов (точніше, mBERT і XLM-R навчені на колекції

104 найбільших різномовних наборах даних Вікіпедії). Після цього вони оцінюються за допомогою міжмовних бенчмарків, таких як XNLI [10]. Однак, XNLI не включає української мови, тому ці моделі не тестувалися конкретно на українських даних. Для тестування розробляється загальний фреймфорк, який дозволяє відносно легко збирати велику кількість українських даних та не потребує ручної анотації.

Після цього застосовується методологія для створення набору даних для класифікації новин (хоча її можна застосувати до кількох інших галузей). Цей набір даних використовується для оцінки та порівняння кількох трансформерів з відкритим вихідним кодом, які є доступними та застосовними для української мови. В останньому розділі аналізуються результати та створюються рекомендації щодо вибору потенційної моделі для аналогічних прикладних завдань.

Ми заохочуємо інших дослідників у цій галузі використовувати цей набір даних для подальшої оцінки своїх моделей.

1. Набір даних

Створено набір даних українських новин, зібраних з кількох джерел, перелічених у розділі 2.1. Фреймворк підготовки даних описаний в розділі 1.2. Даний набір даних використовується для бенчмаркінгу різних моделей класифікації тексту. Ці моделі або навчені, або тонко налаштовуються на задачі, описаній у розділі 1.3.

1.1. Джерела даних

Щоб створити досить великий бенчмарк-набір, потрібно зібрати і розмітити величезну кількість текстів. Також, якщо тестувати моделі за різних умов (наприклад, стиль тексту, довжина, поєднання різних мов тощо), потрібно створювати окреме маркування для кожного з обраних налаштувань. На жаль, збір даних є досить дорогим і складним процесом, особливо для української мови, оскільки вона має обмежену кількість джерел даних, а більшість легкодоступних медіаджерел (наприклад, повідомлення в соціальних мережах, новини тощо), містять значні домішки російських та англійськи фрагментів.

Для вирішення цієї проблеми було розроблено конвейер, який дозволяє збирати нерозмічений набір даних для класифікації тексту без особливих зусиль щодо підготовки даних. При необхідності, цей набір даних може розширювати тим самим способом.

Для цього було завантажено, проаналізовано та спеціальним чином підготовлено дані з кількох українських новинних сайтів: BBC News Україна [11], НВ [12], Українська правда [13], Економічна правда [14], Європейська правда [15], Українська правда Життя [16] та Уніан [17]. Розподіл зібраних даних показано на рис. 1. Повний набір необроблених даних складається з 94 994 текстів.

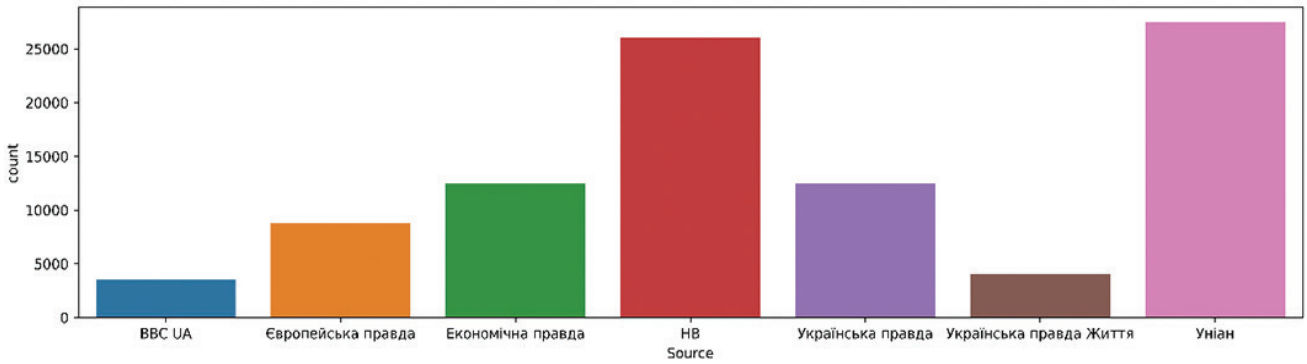

Рис. 1. Розподіл необроблених даних

Після цього використовуємо джерело даних як мету класифікації. Наслідки цього вибору та гіпотези про те, чого модель може навчитися з такої цільової змінної, розглядаються у розділі 2.3.

1.2. Підготовка даних

В необробленому наборі даних існує кілька аспектів, які потребують попередньої обробки та очищення. Навіть проста модель «мішок слів» [19], навчена на цьому отриманому наборі, досягає 0,9 бала за F1-мірою, в той час як більш складні підходи, засновані на глибокому навчанні, показують майже ідеальну точність після кількох епох навчання через численні неявні витоки даних [20].

Для досягнення цих результатів було застосовано наступний конвеєр для підготовки даних:

– Нормалізація пробілів. Початкові та кінцеві символи пробілів були видалені. Послідовності більш ніж одного пробільного символу були стиснуті.

– Видалення заголовка джерела. Усі згадки будь-якої назви джерела у будь-якій граматичній формі (наприклад, «BBC», «Бібісі» або «Служба новин BBC») були замінені спеціальним токеном [ДЖЕРЕЛО] як у назвах статей, так і в текстах. Була використана модифікована версія коректора друкарських помилок Norvig's [21] для усунення неправильного написання заголовків джерел даних.

– Видалення дублікатів. Для кожного кластера дубльованих або схожих текстів залишався лише один екземпляр. Найбільш яскравими прикладами таких кластерів є шаблонні статті про курси валют,

нові випадки коронавірусу в Україні, або новини з фронту, які відрізняються один від одного лише цифрами та дрібними деталями (наприклад, список регіонів червоної зони). Очевидно, що модель могла б запам'ятовувати такі тексти замість вивчення їхньої семантики. Таким чином, усі подібні випадки були розцінені як витік даних.

– Очищення мови. Мова текстів визначалася автоматично за допомогою langdetect [22]. Усі документи неукраїнського походження були видалені з набору даних.

– Витоки шаблонних даних. Було проведено напівавтоматичний пошук типових патернів, що зустрічаються лише у текстах із певного джерела даних, однозначно ідентифікуючи його формою. Усі такі випадки були видалені з набору даних.

Щоб виконати пошук витоку даних шаблону, спочатку ми створили матрицю TF-IDF [23] всіх термінів та найпопулярніших біграм у наборі даних. Потім, ми використовували вибір функції хі-квадрат, щоб знайти 20 найкращих токенів для кожного класу. Всі вибрані токени були перевірені вручну, і деякі типові речення або фрази, що їх містять, були ідентифіковані як витік даних та очищені. Цей процес повторювався кілька разів, поки в топ-20 не потрапив жоден підозрілий токен.

Найбільш яскравими прикладами таких витоків даних шаблону були клікбейтні фрази (наприклад, «Відвідайте наш канал YouTube для більш детальної інформації») та посилання (наприклад, «Зображення надані...»). Приклади таких витоків з контекстом показано рис. 2. Кожен тип витоку даних шаблону був або замаскований токеном [SOURCE], або видалено іншим чином.

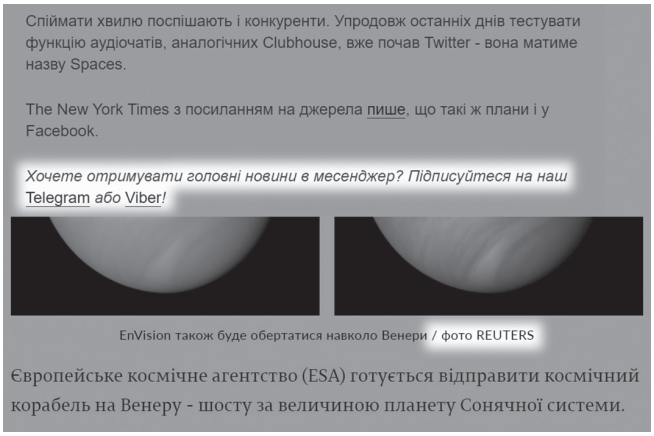

Рис. 2. Приклади витоку даних через шаблонні фрази у статтях BBC та УНІАН відповідно

Хоча ми визнаємо, що такі зміни спотворюють природний розподіл даних, ми вважаємо їх необхідними, щоб зробити задачу репрезентативною для реальних умов, де моделі повинні вивчати складні семантичні зв'язки, а не шукати набір заздалегідь відомих підказок.

Після вищезазначеної попередньої обробки отриманий набір даних складається з 82 554 текстів (близько 12 000 текстів були повністю виключені з різних причин).

Далі цей набір даних поділено на навчальну та тестову вибірки. Повний навчальний набір містить 57789 заголовків та текстів. Тестовий набір складається з 24765 зразків. Підмножини мають аналогічний цільовий розподіл змінних (рис. 3).

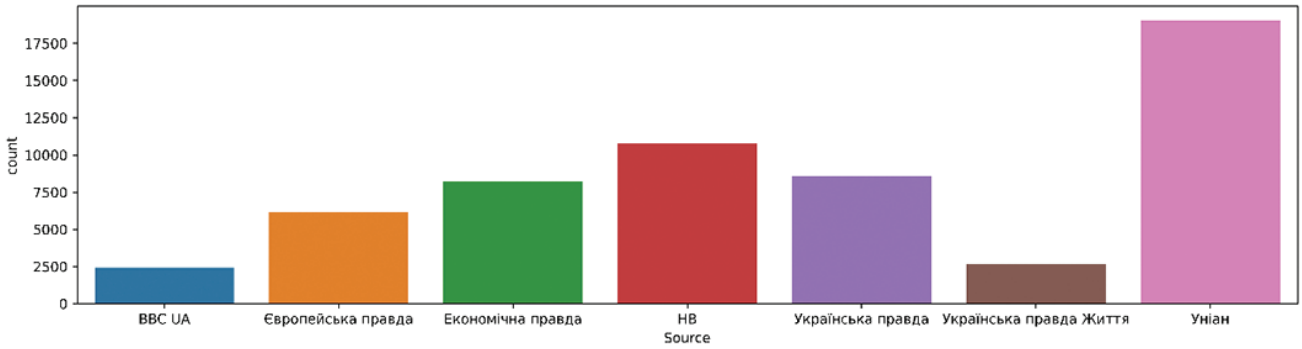

Рис. 3. Розподіл класів навчального набору даних

1.3. Постановка задачі

У якості цілі класифікації використано джерело даних. Таким чином, очікується, що моделі навчатимуться суміші стилістичної класифікації або ідентифікації джерела (оскільки кожне джерело даних має унікальні стилістичні та текстові атрибути), а також тематичному моделюванню (оскільки деякі новинні веб-сайти в нашому наборі даних зосереджені на певних наборах тем, хоча будь-яка велика тема представлена мінімум кількома джерелами).

Ця проблема мультикласової класифікації оцінюється за допомогою макроусередненої оцінки F_1:

$$F_1 = \frac{1}{n}\sum_{i=0}^{n}\frac{2\times Precision_i \times Recall_i}{Precision_i + Recall_i}.$$

З формули видно, що другорядні класи штрафуються в тій же мірі, що і основні, тим самим заставляючи моделі боротися з дисбалансом класів. Набір даних спеціально зроблений незбалансованим, щоб створити більш складну та схожу на реальний світ задачу. Виходячи з цього, пропонується тестувати моделі в двох режимах.

Перший варіант — навчати моделі або на повному навчальному наборі, або на маленькій підмножині, що складається з 8256 зразків. Це дозволяє моделювати продуктивність моделі при обмеженнях навчальних даних: така ситуація може виникнути в реальних додатках, коли вартість маркування даних висока, наприклад, коли для розмітки даних потрібні вузькоспеціалізовані фахівці [18].

Другий варіант — навчати моделі або на повних статтях, або тільки на заголовках. Таким чином, ми можемо визначити продуктивність моделей на текстах різної довжини.

Розподіл довжини тексту та заголовка показано на рис. 4. Тут і далі тексти та заголовки також називаються довгими та короткими текстами відповідно.

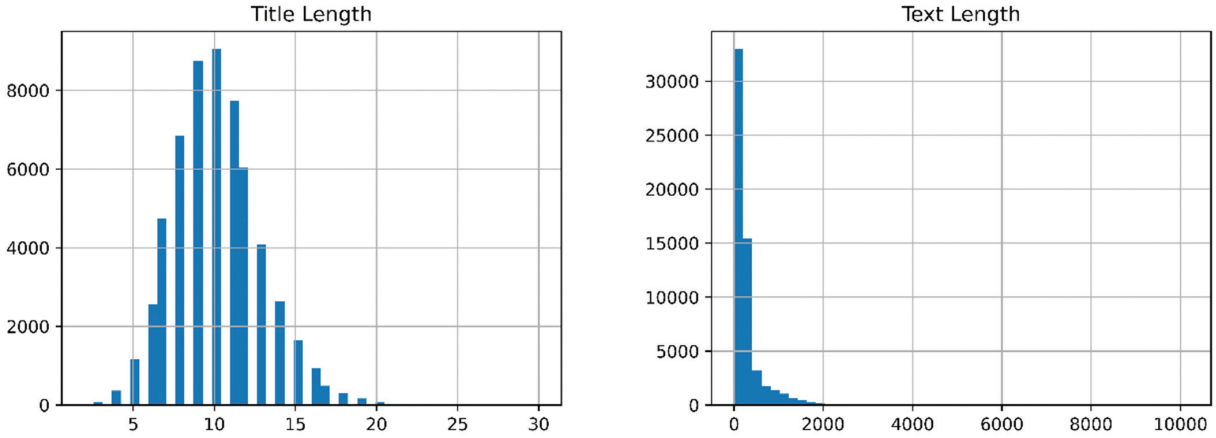

Рис. 4. Архітектура CFFN

Підсумовуючи, можна сказати, що за цих двох умов є чотири навчальні режими. І кожна модель тестується на кожному варіанті.

2. Різновид моделей

В цій роботі тестуються п'ять різних моделей-трансформерів: багатомовну BERT [3], слов'янську BERT [24], ukr-RoBERTa [4], ukr-ELECTRA [5] та XLM-R [9]. Розділи 3.1-5 стисло описують ці моделі.

Також навчається простий, але надійний базовий варіант - NB-SVM на функціях TF-IDF, реалізованих згідно [25]. Його результати можна порвняти з результатами трансформерів при деяких налаштуваннях. Ці результати більш детально обговорюються у розділі 5.

2.1. Багатомовна BERT

Багатомовна BERT (mBERT) є розширенням класичної BERT, навченим на об'єднаних корпусах для 104 мов, включаючи українську. Її можна використовувати для трансферного навчання завдань, в тому числі для класифікації новин. Зазвичай вона показує набагато нижчу продуктивність, ніж моделі, попередньо навчені для кожної окремої мови. Для наших експериментів було спеціально використано необроблену mBERT-базу.

2.2. Слов'янська BERT

Слов'янська BERT є результатом неконтрольованого трансферного навчання від mBERT на об'єднаному корпусі сторінок Вікіпедії, написаних кількома слов'янськими мовами. Хоча цей корпус не включає українську мову, можливо, попереднє навчання на тих самих морфемах може покращити його продуктивність для розглянутого у статті українського набору даних.

2.3 ukr-RoBERTa

ukr-RoBERTa — це версія моделі RoBERTa [26], попередньо навчена спеціально на великомасштабному корпусі, який складається з української Вікіпедії, дедуплікованого набору даних Ukrainian OSCAR [27] та внутрішнього набору даних youscan, зібраного із соціальних мереж. Автори не повідомляють про будь-які результати для цієї моделі в загальнодоступних тестах, хоча й згадують, що вони отримали 2-відсоткове покращення оцінки F1 у своїх внутрішніх наборах даних порівняно з mBERT. Вимірювання продуктивності цієї моделі у загальнодоступному наборі даних має вирішальне значення для її ефективного використання у прикладній науці.

2.4 ukr- ELECTRA

ukr-ELECTRA — це модель на основі архітектури ELECTRA [28], попередньо навчена на сторінках української Вікіпедії та дедуплікованому наборі даних українського OSCAR.

Очікується, що вона повинна працювати краще, ніж укр-RoBERTa, тому що підхід ELECTRA загалом перевершує RoBERTa по більшості завдань. Проте українську версію ELECTRA було попередньо навчено на меншому наборі даних, тому, як показують наші експерименти, вони порівнюються інакше, ніж їхні англійські аналоги.

2.5 XLM-R

XLM-R - це модель на основі RoBERTa, яка навчається так само, як mBERT. XLM-R - єдина модель, яка

має версію попередньо вивчених ваг з відкритим вихідним кодом для версії з великою архітектурою. Щоб знайти найкращу модель класифікації текстів для українських текстів, ми використовуємо цю версію замість базової під час наших експериментів, оскільки очікується, що вона дасть максимальну продуктивність для наступних завдань.

3. Експерименти

Було проведено серію з чотирьох експериментів для кожної моделі:

– малий навчальний набір, навчання лише на заголовках;

– малий навчальний набір, навчання на повних текстах;

– великий навчальний набір, навчання лише на заголовках

– великий навчальний набір, навчання на повних текстах.

Для кожного з цих експериментів проводиться поверхневе налаштування. Ми обираємо планувальник швидкості навчання на невеликій валідаційній підвибірці, відібраній із навчального набору. Після цього ми перенавчаємо модель з найкращими гіперпараметрами на всьому навчальному наборі перед поданням прогнозів.

Щоб порівняти моделі в реальних умовах, замість навчання кожної моделі однаковою кількістю кроків, ми навчаємо кожну модель фіксований час, який дорівнює 24 годинам на одному GPU P100.

Результати бенчмаркінгу наведені в таблиці нижче.

Таблиця 1

Результати бенчмаркінгу використаних моделей

Модель	Короткий текст/ малий навчальний набір	Короткий текст/ великий навчальний набір	Повний текст/ малий навчальний набір	Повний текст/ великий навчальний набір
Базова NB-SVM	0.533	0.790	0.636	0.900
mBERT	0.626	0.853	0.685	0.910
Слов'янська BERT	0.620	0.840	0.708	0.907
ukr-RoBERTa	**0.675**	0.903	**0.745**	0.940
ukr-ELECTRA	0.623	0.909	0.721	0.948
XLM-R	0.624	**0.915**	0.689	**0.950**

4. Результати

Найвища оцінка серед усіх проведених експериментів у XLM-R, навченої на великій версії повнотекстового навчального набору: вона сягає 0,95 F1.

Далі наведено декілька цікавих спостережень:

– і mBERT, і Slavic BERT досить погано працюють із погляду оцінки F1. Хоча це й очікувалося для mBERT, проте, дивно, що Slavic BERT не показала ніякого підвищення точності, працюючи навіть гірше, ніж mBERT, у трьох налаштуваннях з чотирьох.

– ukr-RoBERTa демонструє значне покращення продуктивності порівняно з mBERT (5-6% для коротких текстів та 3-6% для довгих текстів). Вона також показує менший розрив між налаштуваннями короткого та довгого тексту. Ми пов'язуємо це з тим, що вона була навчена на наборі даних, який включає скопійовані повідомлення з соціальних мереж, які, як правило, коротше, ніж інші типи текстів.

– ukr-ELECTRA в середньому показує трохи гірші метрики, будучи менш точною на коротких текстах і трохи точнішою на довгих.

– XLM-R загалом перевершує всі інші моделі на довгих текстах, але водночас демонструє значно нижчу продуктивність на коротких текстах. Варто зазначити, що XLM-R має 24 енкодерні блоки замість 12, тому потребує майже втричі більшої пропускної здатності пам'яті та має вищу затримку порівняно з іншими трансформерними моделями, що брали участь у бенчмаркінгу.

Попри очікування, базова модель NB-SVM демонструє досить високий показник F1 у режимі навчання на великому наборі даних. Вона поступається середній трансформерній моделі лише на 7% у режимі коротких текстів, і показує майже ті самі результати, як mBERT і SlavicBERT у режимі довгих текстів. Ми припускаємо, що це зумовлено тим, що під час навчання моделей на малому наборі даних ефективність підходу трансферного навчання є значно вищою, ніж у випадку великого набору даних.

Ці результати показують, що ukr-RoBERTa може бути кращою моделлю для коротких текстів, тоді як XLM-R або ukr-ELECTRA - найкращий вибір для більш об'ємних. Варто зазначити, що модель NB-SVM, яка не вимагає ні графічного процесора для навчання, ні дорогого обладнання для прогнозування в реальному часі, досягає порівняної продуктивності, якщо набір навчальних даних є досить великим. Результат всього на 5% нижче кращої моделі, а на впровадження та навчання займає п'ятнадцять хвилин, що є прийнятним у багатьох прикладних випадках.

Висновки

У рамках цієї статті розроблено просту та ефективну послідовність дій, яка дозволяє створити набір даних для класифікації тексту з мінімальними витратами.

Використовуючи цей підхід, було створено набір даних для класифікації новин, що складається з майже 60 тисяч навчальних прикладів та дозволяє

проводити бенчмаркінг моделей, використовуючи кілька налаштувань для більш глибокого розуміння плюсів та мінусів моделей.

Набір даних розміщено на платформі Kaggle [1] і є доступним для бенчмаркінгу нових алгоритмів машинного навчання для української мови.

В цій роботі було оцінено кілька наявних моделей з відкритим вихідним кодом на вищезазначеному наборі даних у стандартизованих умовах експерименту. Результати показують, що ukr-RoBERTa та ukr-ELECTRA є найпродуктивнішими моделями середнього розміру, тоді як XLM-R демонструє кращі результати на довгих текстах за відсутності обчислювальних обмежень.

Водночас, NB-SVM демонструє порівняні результати. Це спостереження, а також той факт, що міжмовна модель є однією з найефективніших, означає, що попередньо навченим трансформерним моделям для української мови ще доведеться пройти довгий шлях. Збір великих наборів даних для неконтрольованого попереднього навчання та попереднього навчання великих моделей (наприклад, RoBERTa-large) здаються найбільш перспективними напрямами розвитку.

Список літератури:

[1] https://www.kaggle.com/c/ukrainian-news-classification/

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[4] Vitalii Radchenko. We Trained the Ukrainian Language Model. https://youscan.io/blog/ukrainian-language-model/

[5] Stefan Schweter, Ukrainian ELECTRA model https://github.com/stefanit/ukrainian-electra https://doi.org/10.5281/zenodo.4267880

[6] Babenko, Dmytro. Determining sentiment and important properties of Ukrainian language user reviews : Master Thesis : manuscript rights / Dmytro Babenko ; Supervisor Vsevolod Dyomkin ; Ukrainian Catholic University, Department of Computer Sciences. – Lviv : [s.n.], 2020. – 35 p. : ill.

[7] Babenko, D., & Dyomkin, V. (2019). Determining Sentiment and Important Properties of Ukrainian Language User Reviews. http://ceur-ws.org/Vol-2566/MS-AMLV-2019-paper39-p106.pdf

[8] NER annotation corpus https://lang.org.ua/en/corpora/

[9] Alexis Conneau and Kartikay Khandelwal and Naman Goyal and Vishrav Chaudhary and Guillaume Wenzek and Francisco Guzm´an and Edouard Grave and Myle Ott and Luke Zettlemoyer and Veselin Stoyanov (2019). Unsupervised Cross-lingual Representation Learning at Scale. CoRR, abs/1911.02116.

[10] Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., & Stoyanov, V. (2018). XNLI: Evaluating Cross-lingual Sentence Representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

[11] https://www.bbc.com/ukrainian

[12] https://nv.ua/

[13] https://www.pravda.com.ua/

[14] https://www.epravda.com.ua/

[15] https://www.eurointegration.com.ua/

[16] https://life.pravda.com.ua/

[17] https://www.unian.ua/

[18] Shen, Ying et al. "Improving Medical Short Text Classification with Semantic Expansion Using Word-Cluster Embedding." ArXiv abs/1812.01885 (2018): n. pag.

[19] Zhang, Yin & Jin, Rong & Zhou, Zhi-Hua. (2010). Understanding bag-of-words model: A statistical framework. International Journal of Machine Learning and Cybernetics. 43-52. https://doi.org/10.1007/s13042-010-0001-0

[20] Kaufman, Shachar & Rosset, Saharon & Perlich, Claudia. (2011). Leakage in Data Mining: Formulation, Detection, and Avoidance. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 6. 556-563. https://doi.org/10.1145/2020408.2020496

[21] Peter Norvig. How to Write a Spelling Corrector. url: http://norvig.com/spellcorrect.html.

[22] Shuyo, N. (2010). Language Detection Library for Java.

[23] (2011) TF–IDF. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8.

[24] Arkhipov, A. (2019). Tuning Multilingual Transformers for Language-Specific Named Entity Recognition. In Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing (pp. 89–93). Association for Computational Linguistics.

[25] Wang, C. (2012). Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 90–94). Association for Computational Linguistics.

[26] Yinhan Liu and Myle Ott and Naman Goyal and Jingfei Du and Mandar Joshi and Danqi Chen and Omer Levy and Mike Lewis and Luke Zettlemoyer and Veselin Stoyanov (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR, abs/1907.11692.

[27] Ortiz Su´arez, P., Sagot, B., & Romary, L. (2019). Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In 7thWorkshop on the Challenges in the Management of Large Corpora (CMLC-7). Leibniz-Institut f¨ur Deutsche Sprache.

[28] Kevin Clark, Minh-Thang Luong, Quoc V. Le, & Christopher D. Manning (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In International Conference on Learning Representations

Received (Надійшла) 15.02.2026

Accepted for publication (Прийнята до друку) 05.03.2026

Publication date (Дата публікації) 27.03.2026