УДК 004.8
DOI https://doi.org/10.26661/2786-6254-2025-2-06

ГІБРИДНА СИСТЕМА ВИЯВЛЕННЯ ТЕРМІНІВ В УКРАЇНСЬКОМОВНИХ ТЕКСТАХ ДЛЯ ЗАБЕЗПЕЧЕННЯ КОГНІТИВНОЇ ДОСТУПНОСТІ

Савіцький Р. С.
аспірант,
старший викладач кафедри інженерії програмного забезпечення
Державний університет «Житомирська політехніка»
вул. Чуднівська, 103, Житомир, Україна
orcid.org/0000-0001-9804-3604
roman.savitskyi@gmail.com

Ключові слова: *машинне навчання, виявлення термінів, українська мова, когнітивна доступність, BERT, термінологія.*

Когнітивна доступність є критично важливою для забезпечення рівного доступу до інформації, особливо для користувачів із дислексією, розладами аутистичного спектра та віковими когнітивними змінами. Значна насиченість наукових і технічних текстів спеціалізованою термінологією, що може становити до половини всієї лексики, часто стає суттєвою перешкодою для розуміння змісту. Морфологічна складність української мови, що поєднує багатство відмінкових форм і активне використання запозичень, ускладнює автоматизоване виявлення термінів і потребує розроблення спеціалізованих методів.

У статті представлено комплексне дослідження ефективності сучасних алгоритмів машинного навчання для автоматичного виявлення термінологічної лексики в українських текстах. Порівняння охоплює глибинні моделі на базі трансформерів, класичні алгоритми з лінгвістичним інжинірингом ознак і словникові рішення, засновані на правилах, інтегровані в гібридній ансамблевій системі з адаптивним розподілом ваг. Експерименти проведено на спеціально створеному корпусі з 537 фрагментів тексту з BIO-розміткою, який містить понад три тисячі токенів і характеризується високою часткою термінології. Валідація за допомогою стратифікованої п’ятикратної крос-перевірки засвідчила перевагу ансамблевого підходу, який досяг F1-метрики 0,847 і точності 0,903, перевищивши результати окремих моделей, зокрема fine-tuned BERT (F1 = 0,835), випадковий ліс (F1 = 0,774) та словникового пошуку (F1 = 0.744)

Аналіз важливості ознак виявив, що ключовими індикаторами термінів є довжина слова та співвідношення голосних, а інтеграція цих характеристик із контекстуальним моделюванням забезпечує оптимальний баланс між точністю та швидкістю. Запропонована система здатна обробляти понад дві тисячі токенів на секунду, що дозволяє використовувати її в режимі реального часу для веб- та мобільних застосунків, спрямованих на підвищення когнітивної доступності контенту. Розроблена модель і відкритий корпус розміщені на платформі Hugging Face Hub та можуть стати базою для подальших досліджень і впровадження інструментів автоматичного спрощення текстів, орієнтованих на користувачів з особливими потребами.

MACHINE LEARNING APPROACHES FOR DETECTING TERMS IN UKRAINIAN TEXTS

Savitskyi R. S.
Senior Lecturer at the Department of Software Engineering
Zhytomyr Polytechnic State University
Chudnivska str., 103, Zhytomyr, Ukraine
orcid.org/0000-0001-9804-3604
roman.savitskyi@gmail.com

Key words: *machine learning, term detection, Ukrainian language, cognitive accessibility, BERT, terminology.*

Cognitive accessibility of texts is a critically important aspect of information inclusion, particularly for individuals with dyslexia, autism spectrum disorders, and age-related cognitive changes. Complex terminology poses significant barriers to information access, especially in scientific and technical documents where term concentration can reach 30–50%. Ukrainian, as an inflected language with a developed morphological system, presents additional challenges for automated detection of terminological lexicon through the abundance of case forms and active borrowing of international terms. This article presents a comprehensive study of four machine learning approaches for term detection in Ukrainian texts: fine-tuning of the youscan/ukr-roberta-base transformer model with additional classification layer for token classification, random forest classification with 14 specialized linguistic features adapted for Ukrainian morphology (word length, vowel ratio for «аеєиіїоуюя», morphological endings), dictionary-based deterministic approach using 16,796 classified lexical units with morphological normalization through Stanza NLP pipeline, and a hybrid ensemble system with dynamic weight adaptation based on weighted voting mechanism. Experiments were conducted on a specially created annotated corpus of 537 Ukrainian text samples with BIO term markup, containing 3,173 tokens with 52,76% terminological lexicon, validated through 5-fold stratified cross-validation. Results demonstrate that the ensemble approach achieves the best performance (F1-score = 0,847, Accuracy = 0,903), outperforming BERT fine-tuning (F1 = 0,835), random forest (F1 = 0,774), and dictionary search (F1 = 0,744). Feature importance analysis reveals that word length and vowel ratio are the most powerful discriminators for Ukrainian terminology. The system provides processing speed of 2,156 tokens per second, sufficient for real-time cognitive accessibility applications in web browsers and mobile applications. The developed RomanSavitskyi/ukr-term-detection model and annotated corpus are published on Hugging Face Hub for further research on Ukrainian terminological lexicon and development of accessibility tools for digital inclusion.

Вступ. Рівний доступ до інформації неможливий без урахування когнітивної доступності. За оцінками Всесвітньої організації охорони здоров'я, приблизно 15% населення світу має різні форми когнітивних порушень, як-от дислексія, розлади аутистичного спектра, вікові когнітивні зміни тощо [1]. Складна термінологія, довгі речення і абстрактні концепції створюють бар'єри для доступу до інформації. Особливо гостро проблема проявляється в науковій і технічній літературі, де концентрація спеціалізованої термінології може досягати 30–50% від загального словника тексту. Дослідження показують, що навіть одиничні незрозумілі терміни можуть суттєво знижувати розуміння всього тексту, створюючи ефект каскадного когнітивного навантаження [2].

Нині в українському цифровому просторі практично відсутні спеціалізовані інструменти для автоматизованого виявлення та спрощення термінології. Наявні міжнародні рішення не враховують морфологічних особливостей української мови, а також зростання вимог до цифрової доступності робить розроблення інструментів не лише технологічно важливою, але й соціально необхідною проблемою.

Автоматизація процесу виявлення термінів може суттєво покращити доступність широкого спектра контенту, а ефективні алгоритми дозволили б не лише ідентифікувати проблемні еле-

менти тексту, але й покращити якість тексту, надаючи авторам і редакторам інструменти для створення когнітивно доступних матеріалів.

У контексті описаних вище проблем, дане дослідження спрямоване на вирішення наукових питань, що визначають ефективність автоматизованого виявлення термінів в україномовних текстах. Тому метою даного дослідження є обґрунтування і опис гібридної системи виявлення термінів в україномовних текстах для забезпечення когнітивної доступності.

Для досягнення поставленої мети необхідно вирішити комплекс взаємопов'язаних завдань. Першочерговим завданням є створення анотованого корпусу українських текстів з розміткою термінологічної лексики та формування системи словникових ресурсів для забезпечення детерміністичного підходу до виявлення термінів. Наступним є адаптація та дослідження ефективності трансформерної моделі youscan/ukr-roberta-base для класифікації термінів в українських текстах з урахуванням морфологічної складності мови. Окрім того, необхідно розробити систему лінгвістичних ознак, специфічних для української термінології, та дослідити ефективність класифікатора випадкового лісу. Окремим завданням є реалізація словникового методу виявлення термінів з морфологічною нормалізацією через Stanza NLP, а також побудова гібридної системи з динамічним розподілом ваг, що поєднує переваги трансформерного, статистичного та словникового підходів. **Огляд літератури.** Традиційні формули читабельності, як-от індекс читабельності Флеша – Кінкейда й індекс Фога (Gunning Fog index), базуються переважно на поверхневих характеристиках тексту, як-от довжина речень, кількість складів у словах і частота вживання лексики [3]. Проте сучасні дослідження демонструють обмеженість таких підходів, особливо для морфологічно складних мов. Так, дослідження [4] показує, що когнітивне навантаження тексту значною мірою визначається не лише синтаксичною структурою, але й семантичною складністю окремих лексичних одиниць. Автори встановили, що навіть поодинокі незрозумілі терміни можуть створювати каскадний ефект зниження розуміння, особливо в читачів з когнітивними особливостями [4].

Одним із ключових викликів у виявленні термінологічної лексики в українських текстах є поєднання флективної природи мови з морфологічною варіативністю. На особливу увагу заслуговує дослідження [5], в якому аналізується ефективність моделей машинного навчання, як-от випадковий ліс, метод опорних векторів (SVM) та двоспрямованих кодувальних представлень із трасформерів (Bidirectional Encoder Representations from Transformers (BERT)), для завдань дезінформаційної детекції в українських текстах. Автори виявили, що навіть базові морфологічні перетворення значно підвищують точність класифікації, а застосування попередньо навченої трансформерної моделі забезпечує найкращі результати в умовах обмеженого корпусу [5].

У роботі [6] було запропоновано модель на основі рекурентної нейронної мережі (RNN) для виявлення образливої лексики в україномовному сегменті інтернету. Незважаючи на тематичну специфіку, автори продемонстрували важливість токенізації, контекстної сегментації та граматичної нормалізації. Запропонований ними sequence labelling підхід базується на BIO-розмітці та може бути адаптований для класифікації термінів [6].

Дослідження [7] охоплює методи машинного навчання для класифікації емоційного забарвлення тексту. Автори застосували частоту терміна, оберненого до частоти документа (TF-IDF), логістичну регресію та метод опорних векторів до корпусу українських текстів із високою семантичною складністю. Зокрема, було визначено вплив довжини слова, частоти голосних і морфологічних закінчень на точність класифікації. Ці ознаки є ключовими в термінодетекції, що робить результати дослідження релевантними для побудови ефективних лексичних класифікаторів [7].

У роботі [8] запропоновано гібридний підхід до авторської атрибуції текстів, що поєднує семантичний аналіз із класичними методами машинного навчання. Особливу увагу приділено побудові ансамблевих моделей, що враховують як частотні, так і синтаксичні характеристики тексту. Автори показали, що поєднання векторизації на основі граматичної анотації (POS-міток) із морфологічними фільтрами дозволяє значно підвищити точність класифікації [8].

Автори [9] досліджують проблему виявлення граматичних помилок у текстах українською мовою з використанням машинного навчання. Автори використовують метод опорних векторів, випадковий ліс і системи, засновані на правилах, а також оцінюють вагу ознак, як-от довжина слова, морфеми, співвідношення голосних. Саме ці ознаки показали ефективність і в завданні виявлення термінів, підтверджуючи їхню релевантність у побудові моделей пояснення [9].

Автори [10] дослідили великий міжмовний корпус проєкту ParlaMint II, де для лінгвістичної анотації кількох корпусів (серед яких і український) застосовували різні конвеєри NLP, зокрема Stanza NLP pipeline.

Проте нині практично відсутні спеціалізовані дослідження автоматичного виявлення термінів для української мови. Наявні роботи фокусуються переважно на розпізнавання іменованих сутнос-

тей і загальних NLP завданнях, залишаючи прогалину в термінологічній екстракції.

Методологія. Було створено анотований корпус для тренування моделей, де основний набір даних для навчання алгоритмів виявлення термінів складається з 537 вручну анотованих прикладів українського тексту, організованих у форматі BIO-розмітки (Beginning – Inside – Outside) та збережено у JSON структурі:

```
{
 "id": "sample_001",
 "text": "Машинне навчання використовує алгоритми",
  "tokens": ["Машинне", "навчання", "використовує", "алгоритми"],
   "labels": ["B-TERM", "I-TERM", "O", "B-TERM"],
 "domain": "IT"
}
```

Фрагменти тексту відібрані з наукових статей і освітніх матеріалів у домені інженерії програмного забезпечення, математики, фізики та технічних наук. Критеріями відбору були наявність спеціалізованої термінології, розмір фрагмента – 3–21 токен і збалансоване представлення предметних областей.

Анотування корпусу виконувалося автором з подальшою верифікацією фахівцем з обчислювальної лінгвістики на вибірці 20% корпусу (107 фрагментів). Терміном уважалася лексична одиниця, що позначає спеціалізоване поняття предметної області та потребує пояснення для нефахівця. Токенізація виконувалася за допомогою wordpiece tokenizer з моделі youscan/ukr-roberta-base для забезпечення сумісності із трансформерним підходом.

Загальний обсяг набору даних становить 3,173 токени, з яких 1,674 (52,76%) позначені як терміни, що свідчить про високу концентрацію термінологічної лексики у відібраних текстах.

Структурні характеристики набору даних демонструють придатність для навчання моделей машинного навчання. Середня довжина речення становить 5,91 токена з варіацією від 3 до 21 токена, що відповідає типовим фрагментам наукових та технічних текстів. Приблизно 94,6% прикладів (508 із 537) містять принаймні один термін, що забезпечує збалансоване представлення класу з наявністю термінів у навчальному наборі.

Тематичний аналіз виявлених термінів показує різноманітність предметних областей: інформаційні технології (“javaScript”, “docker”, “NLP”), математика та фізика («рівняння Ейнштейна», «теорема Піфагора»), методології розробки програмного забезпечення (“scrum”, “agile”), кібербезпека («криптографічний захист інформації», «кібератака»). Оскільки корпус охоплює тексти з різних доменів, це дозволяє протестувати здатність моделей до узагальнення.

Для підтримки підходів, заснованих на правилах і забезпеченні необхідного рівня порівнянь, було створено комплексну систему словникових ресурсів, що налічує 16,796 класифікованих лексичних одиниць. Процес формування словників здійснювався через агрегацію існуючих лексикографічних ресурсів та їх адаптацію для завдання виявлення термінів. Найбільшими компонентами є словник архаїчної лексики (7,427 записів) та лайки (6,609 записів), що свідчить про фокус системи на виявленні когнітивно складної лексики.

Словникові ресурси організовані за принципом пріоритетної заміни, де кожен термін супроводжується альтернативними варіантами різного рівня складності. Морфологічна обробка словникових одиниць здійснюється через Stanza NLP pipeline для української мови, що забезпечує нормалізацію відмінкових форм і числових варіацій. Технічна реалізація пошуку використовує хешування для первинного точного збігу з подальшою морфологічною нормалізацією для обробки флективних варіантів.

Основою рішення є тонке налаштування (fine-tuning) трансформерних моделей для виявлення термінів. Модель є адаптацією попередньо навченої архітектури на базі трансформерів для специфічного завдання класифікації токенів (token classification). За базову модель було обрано youscan/ukr-roberta-base, що є спеціалізованою версію RoBERTa для української мови, що демонструє високу якість на завданнях розуміння української мови [5].

Процес тонкого налаштування навчання проводиться з використанням гіперпараметрів, як-от 3 епохи, batch size 8, warmup steps 500, зниження ваги 0,01 та планування швидкості навчання з автоматичною оптимізацією. Розподіл набору даних становить 80% для навчання та 20% для перевірки з ранньою зупинкою на основі точності валідації.

Критичним компонентом імплементації є точна післяобробка результатів моделі. Система включає фільтрацію за довжиною слів (3–50 символів), видалення власних назв через евристичний аналіз капіталізації та морфологічних закінчень, а також фільтрування за спеціально заданим списком для 128 категорій загальновживаних слів. Додатково застосовується адаптивне порогове значення, де поріг класифікації змінюється залежно від наявності термінологічних ознак у слові.

Альтернативний підхід – випадковий ліс з лінгвістичними ознаками, а саме створення ансамблю дерев рішень зі спеціально розробленими лінгвістичними ознаками, адаптованими для особливостей української мови. Система feature

extraction базується на 14 основних характеристиках слів, що охоплюють фонетичні, морфологічні та структурні властивості.

Основні групи ознак включають метричні (довжина слова, кількість складів), фонетичні (співвідношення голосних до), структурні (повторювані символи, наявність цифр та спеціальних знаків) та позиційні (початкова/кінцева голосна) ознаки. Додатково враховуються морфологічні індикатори термінів: технічні префікси («авто», «мікро», «кібер»), суфікси («логія», «графія», «скопія»), характерні для термінологічної лексики словотворчі елементи.

Імплементація використовує scikit-learn класифікатор випадкового лісу з оптимізованими параметрами, як-от кількість дерев 100, максимальна глибина 10, мінімальне розбиття термінів 5, що допомагає запобігти перенавчанню на відносно невеликому наборі даних. Важливою особливістю є динамічна адаптація ознак під контекст виявлених термінів, що дозволяє покращити точність для рідкісних термінологічних категорій.

Словниковий метод базується на точному збігу з використанням комплексної системи словникових ресурсів і морфологічної форм. Основу становить спеціалізований словник із 675 складних термінів і 1,274 абревіатур, доповнений системою морфологічного аналізу через Stanza NLP pipeline для української мови.

Алгоритм включає декілька етапів обробки, починаючи від первинного пошуку за хешем, продовжуючи морфологічною нормалізацією для обробки відмінкових форм і числових варіацій, закінчуючи багаторівневою фільтрацією. Особливістю імплементації є використання оцінки впевненості (confidence scoring), де кожний збіг оцінюється за точністю відповідності (точний збіг = 1,0, морфологічні варіанти = 0,8–0,9) та контекстною валідністю через аналіз сусідніх слів. Такий підхід забезпечує високу точність за збереження прийнятного рівня виявлення для добре представлених у словнику категорій термінів.

Фінальна система поєднує переваги всіх описаних методів через weighted ensemble architecture з динамічним розподілом ваг залежно від характеристик конкретного випадку. Базові ваги становлять BERT fine-tuned модель (0,5), випадковий ліс з лінгвістичними ознаками (0,3), словниковий підхід (0,2).

Особливістю системи є адаптивні ваги, де ваги компонентів корегуються залежно від характеристик тексту, оскільки для технічних текстів підвищується вага BERT моделі, для загальної лексики – словникового підходу. Такий механізм забезпечує міцність системи до різних типів контенту та покращує загальну якість класифікації термінів.

Дослідження базувалося на трансформерній архітектурі youscan/ukr-roberta-base, попередньо навченій на українськомовному корпусі обсягом 7,8 гігабайта. Архітектурне рішення передбачало інтеграцію спеціалізованого класифікаційного модуля для розпізнавання іменованих сутностей на рівні токенів із застосуванням трикласової схеми маркування (B-TERM, I-TERM, O).

Процедура донавчання здійснювалася з коефіцієнтом навчання 2×10^{-5} та розміром пакету 16 зразків. Максимальна довжина вхідної послідовності була обмежена 512 токенами. Навчальний процес тривав протягом 10 епох із реалізацією механізму дострокового припинення за досягнення мінімального значення функції втрат на валідаційній вибірці.

Класифікатор для випадкового лісу базується на 14 спеціалізованих лінгвістичних ознаках, адаптованих для української морфології, як-от довжина слова, кількість повторюваних символів, максимальна довжина повтору, кількість голосних, кількість приголосних, співвідношення голосних, кількість цифр і спеціальних символів, позиційні характеристики, кількість сленґових патернів, категоріальні ознаки довжини та співвідношення символів.

Оцінювання проводилося за стандартними метриками для завдань sequence labeling: Precision, Recall та F1-score для кожного класу (B-TERM, I-TERM, O) з мікро- та макроусередненням. Додатково вимірювалися швидкість інференсу (токенів/секунда), розмір моделі та споживання пам’яті для оцінювання практичної застосовності.

Перевірка здійснювалася через 5-fold стратифіковану крос-валідацію з підтримкою балансу класів у кожному згині. Статистична значущість різниць між моделями перевірялася за допомогою парних t-тестів з поправкою Боферонні для множинних порівнянь [11].

Результати та аналіз. Результати демонструють значні відмінності у продуктивності методів залежно від архітектурних особливостей та підходів до обробки української термінології. Найкращі результати за комплексним критерієм F1-score демонструє ансамблевий підхід (0,847), що ефективно поєднує сильні сторони всіх базових методів. BERT fine-tuning показує другий результат (0,835), демонструє потужність контекстуального розуміння для термінологічної класифікації. Випадковий ліс з лінгвістичними ознаками досягає F1-score 0,774, що є прийнятним результатом для інтерпретованої моделі зі швидким припущенням (див. табл. 1).

Словниковий підхід характеризується найвищою точністю (0,886), що зумовлено детерміністичною природою точного зіставлення. Проте рівень виявлення (0,643) обмежений покриттям

Таблиця 1

Загальна продуктивність алгоритмів виявлення термінів

Метод	Precision	Recall	F1-Score	Accuracy	Швидкість (токенів/с)	Розмір моделі
BERT Fine-tuning	0,847	0,823	0,835	0,891	1,247	467 MB
Випадковий ліс	0,792	0,756	0,774	0,834	8,934	12 MB
Словниковий пошук	0,886	0,643	0,744	0,812	12,450	3,2 MB
Ансамблевий підхід	0,861	0,834	0,847	0,903	2,156	482 MB

словникових ресурсів, особливо для нових термінів і морфологічних варіантів поза словником. Саме такий дисбаланс типовий для систем, заснованих на правилах.

Словниковий підхід демонструє найвищу швидкість обробки (12,450 токенів/с) через операції простого пошуку та відсутність складних обчислень. Випадковий ліс показує другий результат (8,934 токенів/с) завдяки ефективності деревоподібної архітектури для висновків. BERT fine-tuning, незважаючи на найбільший розмір моделі, забезпечує прийнятну швидкість (1,247 токенів/с) для застосування в режимі реального часу. Ансамблевий підхід показує проміжну швидкість (2,156 токенів/с), що є компромісом між точністю та продуктивністю.

Парні t-тести з поправкою Бонферонні підтверджують статистичну значущість різниць між усіма методами за F1-score [11]. Найбільші відмінності спостерігаються між підходами, створеними на базі BERT та словниковою системою, особливо за рівнем виявлення (recall) метрикою.

Варто зазначити, що оцінка важливості терміна для токена t обчислюється за формулою лінійної комбінації прогнозів базових класифікаторів:

$$W(t) = w_1 \times P_1(t) + w_2 \times P_2(t) + w_3 \times P_3(t),$$

де $P_1(t) = P\_BERT(t)$ – імовірність класифікації токена як терміна трансформерною моделлю; $P_2(t) = P\_RF(t)$ – імовірність класифікатора випадкового лісу з лінгвістичними ознаками; $P_3(t) = P\_DICT(t)$ – бінарний індикатор словникового збігу; w_i – нормалізовані вагові коефіцієнти ($w_1 + w_2 + w_3 = 1$).

Базові ваги встановлюються згідно із продуктивністю кожного методу на валідаційному наборі: $w_1 = 0,5$, $w_2 = 0,3$, $w_3 = 0,2$. Значення відображають емпірично встановлену ієрархію точності, а саме BERT fine-tuning демонструє найвищу контекстуальну чутливість, випадковий ліс забезпечує стабільність через лінгвістичні ознаки, а словниковий підхід надає високу precision для відомих термінів.

Детальний аналіз важливості 14 лінгвістичних ознак для класифікатора випадкового лісу виявляє специфічні характеристики української термінологічної лексики. У дослідженні було доведено важливість для кожної ознаки, проте розглянемо деякі з них. Довжина слова (важливість = 0,234) виявилася найпотужнішим дискримінатором для української термінології. Аналіз розподілу показує, що українські терміни мають бімодальний розподіл довжини, зокрема короткі абревіатури (3–5 символів) та довгі складні терміни (8–15 символів). Середня довжина термінів (9,4 символу) значно перевищує загальновживані слова (6,1 символу). Співвідношення голосних (0,189) відображає фонетичні особливості української наукової термінології. Українські терміни часто мають низьке співвідношення голосних (0,42), особливо це помітно для технічних термінів на кшталт «криптографія», «алгоритм», «протокол». Співвідношення букв (0,156) ефективно відрізняє справжні терміни від абревіатур і символьних позначень. Чисті терміни, що складаються лише з літер, мають вищий пріоритет порівняно зі змішаними конструкціями, що містять цифри чи спеціальні символи.

Кореляційний аналіз виявив сильні зв'язки між пов'язаними ознаками, оскільки довжина слова корелює з кількістю приголосних ($r = 0,84$) і голосних ($r = 0,78$). Особливо ефективними виявилися комбінації таких ознак, як довжина та співвідношення голосних для технічних термінів, наявність цифр і спеціальні символи для формальних позначень, морфологічні закінчення для класичних наукових термінів. Використання українського вокалізму «аеєиіїоуюя» для розрахунку співвідношення голосних виявилося критично важливим. Експерименти з латинським набором голосних показали зниження F1-score на 0,043, що підтверджує необхідність мовно-специфічної адаптації лінгвістичних ознак.

Порівняння з іншими дослідженнями показує, що специфіка української мови створює додаткові виклики для виявлення термінів [4]. Створена гібридна система досягла F1-метрики 0,847, тоді як ансамблева модель авторів для англійської мови показала 0,388, що пояснюється не лише різницею в постановці завдань (виявлення термінів проти загального спрощення), але й морфологічною складністю української мови, яка потребувала інтеграції Stanza NLP pipeline та підвищила точність словникового підходу на 23%.

Варто більш детально проаналізувати помилки класифікації на тестовому наборі. Категоризація помилок здійснювалася через ручну перевірку випадково вибраних 200 неправильно класифікованих токенів зі збалансованим вибірковим дослідженням по всіх методах (див. табл. 2).

Таблиця 2

Розподіл типів помилок за методами

Тип помилки	BERT	Випадковий ліс	Словниковий	Ансамблевий
Хибнопозитивний результат				
Загальновживані IT-слова	23%	31%	8%	19%
Власні назви компаній	18%	12%	2%	11%
Абстрактні поняття	15%	22%	3%	12%
Загальні наукові слова	12%	18%	5%	9%
Хибнонегативний результат				
Багатословні терміни	35%	42%	58%	31%
Англомовні терміни	28%	38%	72%	25%
Морфологічні варіанти	22%	15%	45%	18%
Нові/рідкісні терміни	15%	5%	38%	26%

Аналіз хибнопозитивного результату показав, що загальновживані IT-слова становлять найбільшу категорію помилок для традиційних методів. Слова на кшталт «програма», «система», «документ», «інформація», мають морфологічні характеристики термінів (довжина, закінчення), але є занадто загальними для класифікації як специфічна термінологія. Випадковий ліс особливо схильний до таких помилок (31%) через фокус на поверхневих лінгвістичних ознаках без семантичного розуміння. Власні назви компаній (компанії “Microsoft”, “Google”, “Apple”) помилково класифікуються як терміни через їх появу в технічних контекстах. BERT показує вищий рівень таких помилок (18%) через контекстуальні асоціації, тоді як словниковий підхід ефективно фільтрує через спеціалізований стоп-список слів (2%).

Абстрактні поняття («метод», «процес», «результат») створюють виклик для всіх методів через двозначність, адже вони можуть бути як загальновживаними словами, так і термінами, залежно від контексту. Випадковий ліс демонструє найвищий рівень таких помилок (22%) через неможливість врахувати семантичний контекст.

Аналіз хибнонегативного результату показує, що багатослівні терміни («машинне навчання», «штучний інтелект», «обробка природної мови») становлять найбільший виклик для всіх методів. Словниковий підхід показує найгірші результати (58%) через обмежене покриття багатослівних конструкцій у словниках. BERT демонструє найкращі результати (35%) завдяки контекстуальному розумінню зв’язків між словами.

Морфологічні варіанти термінів («криптографічний», «алгоритмічний», «програмістський») пропускаються через недосконаліть морфологічної нормалізації. Випадковий ліс показує найкращі результати (15%) завдяки лінгвістичним ознакам, що враховують морфологічні патерни.

Особливу категорію становлять контекстуально-залежні помилки, де одне слово може бути терміном в одному контексті та загальним словом в іншому. Наприклад, «мережа» як комп’ютерний термін проти «мережа» у загальному понятті зв’язків. BERT показує найкращу здатність розрізняти такі випадки завдяки механізмам узгодження, але все ще демонструє 12% помилок у цій категорії. Традиційні методи практично неспроможні вирішувати такі неоднозначності без додаткових можливостей аналізу контексту.

Висновки. Порівняльний аналіз чотирьох підходів дозволив виділити фундаментальні компроміси між точністю, інтерпретованістю та обчислювальною ефективністю для виявлення української термінології. Отримані результати підтверджують, що вибір оптимального методу залежить від специфічних вимог застосування та доступних ресурсів.

Унаслідок чого була обґрунтована та розроблена гібридна система виявлення термінів в українськомовних текстах для забезпечення когнітивної доступності. Дослідження роботи якої дозволило виявити специфічні особливості визначення термінів в українській мові. Морфологічна варіативність значно ускладнила точне зіставлення зі словниками. Використання Stanza NLP pipeline для морфологічної нормалізації покращило рівень виявлення для словникового підходу на 23%, але все ще залишає значні прогалини для рідкісних форм та неологізмів.

До перспектив майбутніх досліджень відносимо розроблення ієрархічних схем багаторівневого класифікатору, що дозволить краще враховувати рівні термінологічної складності, а також удосконалення контекстуального моделювання шляхом інтеграції контексту на рівні документа для більш точного визначення термінів залежно від контексту.

ЛІТЕРАТУРА

1. Hartley S. World Report on Disability (WHO). 2011. DOI: 10.13140/RG.2.1.4993.8644
2. Zha X., Wang X., Yan Y., Gao Y., Yan G. Exploring influencing mechanism of herd behavior in academic information use: The perspective of cognitive load. *The Journal of Academic Librarianship.* 2023. Vol. 49. № 3. DOI: 10.1016/j.acalib.2023.102705

3. Maharani A. A. An analysis of the readability level in Life Today textbook for twelfth grade of senior high school using Flesch Reading Ease formula by Rudolf Flesch. Diploma thesis. Bandar Lampung : UIN Raden Intan Lampung, 2025. 70 p. URL: https://repository.radenintan.ac.id/39235/
4. Chamovitz E., Abend O. Cognitive simplification operations improve text simplification. *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL).* Abu Dhabi, United Arab Emirates (Hybrid), 2022. P. 241–265. DOI: 10.18653/v1/2022.conll-1.17
5. Lipianina-Honcharenko K., Soia M., Yurkiv K. Evaluation of the effectiveness of machine learning methods for detecting disinformation in Ukrainian text data. *CEUR Workshop Proceedings.* 2024. Vol. 3702. P. 1–8. URL: https://ceur-ws.org/Vol-3702/paper9.pdf
6. Krak I., Zalutska O., Molchanov M., Mazurets O., Bahrii R. Abusive speech detection method for Ukrainian language using recurrent neural network. *CEUR Workshop Proceedings.* 2024. Vol. 3688. P. 1–9. URL: https://ceur-ws.org/Vol-3688/paper2.pdf
7. Lomovatskyi A., Basyuk T. Methods of machine learning and design of a system for determining the emotional coloring of Ukrainian-language content. *Scientific Bulletin of Lviv Polytechnic National University.* 2024. № 2. P. 78–90. DOI: 10.23939/sisn2024.15.074
8. Uhryn D., Vysotska V., Chyrun L., Chyrun S., Hu C. Intelligent application for textual content authorship identification based on machine learning and sentiment analysis. *International Journal of Intelligent Systems and Applications.* 2024. Vol. 17. № 2. P. 44–53, DOI: 10.5815/ijisa.2025.02.05
9. Lytvyn V., Pukach P., Vysotska V., Vovk M., Kholodna N. Identification and correction of grammatical errors in Ukrainian texts based on machine learning technology. *Mathematics.* 2023. Vol. 11. № 4. Art. 904. DOI: 10.3390/math11040904
10. ParlaMint II: advancing comparable parliamentary corpora across Europe / T. Erjavec, M. Kopp, N. Ljubešić et al. *Language Resources and Evaluation.* 2025. Vol. 59. P. 2071–2102. DOI: 10.1007/s10579-024-09798-w
11. Mondal H., Mondal S., Majumber R., De R. Conduct Common Statistical Tests Online. *Indian Dermatology Online Journal.* 2022. Vol. 13. № 4. P. 539–542. DOI: 10.4103/idoj.idoj_605_21

REFERENCES

1. Hartley, S. (2011). *World report on disability (WHO)*. 10.13140/RG.2.1.4993.8644
2. Zha, X., Wang, X., Yan, Y., Gao, Y., & Yan, G. (2023). Exploring influencing mechanism of herd behavior in academic information use: The perspective of cognitive load. *The Journal of Academic Librarianship, 49* (3), 102705. 10.1016/j.acalib.2023.102705
3. Maharani, A. A. (2025). *An analysis of the readability level in Life Today textbook for twelfth grade of senior high school using Flesch Reading Ease formula by Rudolf Flesch* [Diploma thesis, UIN Raden Intan Lampung]. UIN Raden Intan Repository. https://repository.radenintan.ac.id/39235/
4. Chamovitz, E., & Abend, O. (2022) *Cognitive simplification operations improve text simplification.* Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL). Abu Dhabi, United Arab Emirates (Hybrid). P. 241–265. DOI: 10.18653/v1/2022.conll-1.17
5. Lipianina-Honcharenko, K., Soia, M., & Yurkiv, K. (2024). Evaluation of the effectiveness of machine learning methods for detecting disinformation in Ukrainian text data. *CEUR Workshop Proceedings, 3702*, 1–8. https://ceur-ws.org/Vol-3702/paper9.pdf
6. Krak, I., Zalutska, O., Molchanov, M., Mazurets, O., & Bahrii, R. (2024). Abusive speech detection method for Ukrainian language using recurrent neural network. *CEUR Workshop Proceedings, 3688*, 1–9. https://ceur-ws.org/Vol-3688/paper2.pdf
7. Lomovatskyi, A., & Basyuk, T. (2024). Methods of machine learning and design of a system for determining the emotional coloring of Ukrainian-language content. *Scientific Bulletin of Lviv Polytechnic National University*, (2), 78–90. https://science.lpnu.ua/sites/default/files/journal-aper/2024/aug/35646/maket2402951-78-90_0.pdf
8. Uhryn, D., Vysotska, V., Chyrun, L., Chyrun, S., & Hu, C. (2024). Intelligent application for textual content authorship identification based on machine learning and sentiment analysis. *International Journal of Intelligent Systems and Applications, 17* (2), 44–53. https://www.mecs-press.org/ijisa/ijisa-v17-n2/IJISA-V17-N2-5.pdf
9. Lytvyn, V., Pukach, P., Vysotska, V., Vovk, M., & Kholodna, N. (2023). Identification and correction of grammatical errors in Ukrainian texts based on machine learning technology. *Mathematics, 11* (4), 904. DOI: 10.3390/math11040904
10. Erjavec, T., Kopp, M., Ljubešić, N., et al. ParlaMint II: advancing comparable parliamentary corpora across Europe. Lang Resources & Evaluation, 59, 2071–2102 (2025). DOI: 10.1007/s10579-024-09798-w
11. Mondal H., Mondal S., Majumder R., & De R. (2022). Conduct Common Statistical Tests Online. Indian Dermatology Online Journal, 13 (4), 539–542. DOI: 10.4103/idoj.idoj_605_21

Дата першого надходження рукопису до видання: 28.08.2025
Дата прийнятого до друку рукопису після рецензування: 25.09.2025
Дата публікації: 31.12.2025