ISSN 2707-1898 (print)
ISSN 2786-9172 (online)

Correspondence author
@ V. A. Vysotska
victoria.a.vysotska@lpnu.ua

Український журнал інформаційних технологій
Ukrainian Journal of Information Technology
http://science.lpnu.ua/uk/ujit
https://doi.org/10.23939/ujit2025.02.120
Article received 29.04.2025 p.
Article accepted 30.10.2025 p.
UDC 004.8:9

Р. Б. Федчук, В. А. Висоцька
Національний університет "Львівська політехніка", м. Львів, Україна

МАТЕМАТИЧНА МОДЕЛЬ ІДЕНТИФІКАЦІЇ ПОМИЛОК В ТЕКСТАХ УКРАЇНОМОВНОГО КОНТЕНТУ

Проблема автоматизованого виявлення помилок у текстах українською мовою набуває особливої актуальності в умовах зростання обсягів цифрового контенту. Розроблено математичну модель системи підтримки прийняття рішень для виявлення помилок в україномовних текстах. Досліджено процес ідентифікації помилок як задачу багатокласової класифікації на рівні токенів з урахуванням контексту тексту. Запропоновано використання ймовірнісних моделей для визначення типу помилки залежно від оточення токенів у тексті. Виявлено доцільність формування навчальних вибірок, що містять як реальні, так і штучно внесені помилки, для забезпечення збалансованості навчального процесу. Встановлено ефективність підходів до векторизації текстів із урахуванням морфологічної та синтаксичної структури української мови, що підвищує точність роботи моделі. З'ясовано, що інтеграція контекстуальної інформації істотно покращує результати ідентифікації помилок. Побудовано детальні DFD-діаграми, які формалізують процеси функціонування системи та взаємодію її компонентів. Здійснено експериментальне навчання моделі ukr-roberta-base на корпусі UA-GEC для завдання ідентифікації помилок в українських текстах. Отримано такі результати якості моделі: F1 – 0,736, accuracy – 0,76, precision – 0,85, recall – 0,65. Надано приклади роботи моделі на тестових даних. Виявлено, що модель уже навчилась виявляти пунктуаційні та базові орфографічні помилки, що свідчить про її ефективність і перспективи подальшого розвитку. Перспективи подальших досліджень – масштабування розробленої моделі та її адаптація для розширення покриття складніших типів мовних помилок.

Ключові слова: опрацювання природної мови, ідентифікація помилок, класифікація, машинне навчання, українська мова, ймовірнісна модель.

Вступ / Introduction

У сучасних реаліях зростання обсягів текстових даних проблема автоматизованої перевірки та корекції помилок у текстах набуває особливої актуальності. Хоча для англійської мови розроблено велику кількість ефективних методів вирішення завдання автоматичного виправлення граматичних помилок (GEC)[1], для української мови такі підходи залишаються малодослідженими. Це зумовлено як браком великих анотованих корпусів, так і специфічними лінгвістичними особливостями української мови – морфологічною складністю, синтаксичним різноманіттям, наявністю діалектів і високою контекстною залежністю. Виявлення та корекція орфографічних, граматичних, пунктуаційних і семантичних помилок у текстах українською мовою – важливі завдання для забезпечення якості інформації, її достовірного сприйняття та подальшого опрацювання в інформаційних системах.

Сьогодні вирішення цього завдання для україномовних текстів ускладнюють кілька причин. Передусім, більшість традиційних систем автоматичної перевірки текстів розраховані на менш складні з граматичного погляду мови, такі як англійська. Застосування цих систем до української мови виявляється малоефективним, оскільки вони не враховують специфіку її граматики та синтаксису. Окрім цього, в українській мові є значні лексико-семантичні залежності, які залежать від контексту.

Традиційні системи ідентифікації помилок, як правило, побудовані на основі набору лінгвістичних правил [2], які сформулювали експерти, і словникових баз. Такі системи перевіряють орфографічну правильність, узгодження чисел, родів чи відмінків між словами і навіть базові пунктуаційні конструкції. Однак складність і багатогранність граматичних залежностей української мови спричиняють численні обмеження для таких

систем. Вони не враховують контексти, в яких інтерпретація правил часто змінюється. Крім того, створення та підтримка обширного набору правил для таких систем потребують значних людських і часових ресурсів. Як наслідок, традиційні системи демонструють низьку продуктивність у реальних умовах і майже не адаптувалися до нових даних та стилів тексту.

Складніші системи, побудовані на основі статистичного машинного перекладу (SMT), стали наступним етапом розвитку і запропонували гнучкіший підхід до виявлення помилок. В основі таких моделей – аналіз великих корпусів текстів. Для кожного речення була відома правильна і помилкова версія. Хоча цей підхід певною мірою покращений порівняно із попереднім, він все ще залишався обмеженим через залежність від якості навчального корпусу. Крім того, SMT-моделі недостатньо враховують глобальний контекст тексту.

Сучасна епоха вирішення завдання виявлення помилок охоплює використання машинного навчання, яке має вагомі переваги порівняно з традиційними підходами. Моделі машинного навчання, такі як трансформерні архітектури, здатні враховувати не лише локальні, але і глобальні залежності між словами в тексті. Для ідентифікації помилок використовують підхід класифікації; кожен токен тексту аналізують й позначають відповідною міткою. В українській мові цей етап стикається зі значними труднощами, пов'язаними з морфологічною складністю, адже більшість слів може набувати різних форм залежно від відмінка, числа чи роду. Також граматичні залежності часто визначаються взаємодією кількох токенів, розташованих на значній відстані один від одного в тексті.

Використання методів машинного навчання має істотні переваги. Моделі, побудовані на таких архітектурах, адаптивні у роботі з реальними текстами й здатні якісно враховувати контекст. Вони позбавлені жорсткої обмеженості, притаманної лінгвістичним правилам, і можуть навчатися на великих корпусах даних, поступово покращуючи свої прогнози.

Об'єкт дослідження – процеси автоматичної ідентифікації та корекції помилок в українських текстах з урахуванням їх морфологічної та синтаксичної складності.

Предмет дослідження – математична модель, що дає змогу вирішувати завдання ідентифікації помилок.

Мета дослідження – розроблення математичної моделі системи підтримки прийняття рішень для ідентифікації помилок в україномовних текстах із використанням методів машинного навчання. Визначено такі *завдання дослідження*:

- формалізація процесів аналізу текстів українською мовою;
- побудова математичної моделі для ідентифікації помилок із урахуванням морфологічної, синтаксичної та контекстної специфіки мови;
- створення детальних DFD-діаграм для описання інформаційних потоків і функціонування системи;
- опис алгоритму для задач ідентифікації помилок;
- навчання моделі на корпусі українських текстів із помилками експериментальній перевірці, оцінюванню якості роботи та аналізу результатів.

Аналіз останніх досліджень та публікацій. Сучасні засоби автоматичного виправлення текстів стали важливим інструментом для подолання різноманітних видів орфографічних, граматичних, пунктуаційних і стилістичних помилок. У цьому контексті найбільш відомими й ефективними рішеннями є програмні застосунки та платформи, інтегровані в текстові процесори, онлайн-інструменти та системи машинного навчання.

Grammarly є одним із найпопулярніших інструментів для автоматичної перевірки текстів [3]. На жаль, поки що Grammarly не підтримує перевірки граматики, орфографії чи стилістики української мови. Основна функціональність сервісу зосереджена на англійській мові. Проте компанія активно працює над розвитком української мови в галузі комп'ютерної лінгвістики. Зокрема, Grammarly створила та опублікувала у відкритому доступі анотований GEC-корпус української мови, який містить майже 34 000 речень [4–6]. Цей ресурс призначений для наукового та практичного вивчення мови, а також для тренування та оцінювання програм виявлення та виправлення граматичних помилок.

Серед спеціалізованих українських інструментів виділимо OnlineCorrector [7], орієнтований на перевірку правопису, пунктуації та стилістики в текстах. Цей онлайн-засіб дає змогу завантажувати тексти для перевірки та отримувати результат із виправленими помилками. Особливу увагу він приділяє роботі із текстами різного формату, враховуючи специфіку української мови.

Сучасні моделі машинного навчання, зокрема GECToR [8], відкривають нові можливості для автоматизації процесів коригування текстів. Ця модель побудована на трансформерній архітектурі та доступна для адаптації до будь-якої мови, ураховуючи українську. GECToR використовує механізм уваги для врахування контексту у виправленні текстів, що дає змогу досягати високої точності та ефективності.

Сьогодні для розробників також доступні платформи, такі як Hugging Face [9], які пропонують попередньо натреновані моделі, зокрема багатомовні варіанти BERT, GPT чи T5. Ці платформи дають змогу створювати кастомізовані рішення для завдань виявлення та виправлення помилок та інших задач опрацювання природної мови, адаптувавши моделі до специфіки мови та стилю тексту.

Чудовим прикладом інтерактивного багатомовного інструменту є сучасні GPT-моделі, зокрема ChatGPT, які можуть не лише перевіряти і виправляти тексти, але й адаптуватися до складних стилістичних особливостей, ураховуючи контекст тексту. Ці моделі здатні працювати із будь-яким типом тексту, забезпечуючи як базову перевірку, так і контекстно обґрунтовані виправлення.

У [10] дослідженні вивчено ефективність застосування GPT-3.5 для виправлення граматичних помилок (GEC) у багатомовному контексті в різних сценаріях, ураховуючи нульове налаштування (zero-shot learning), точне налаштування (fine-tuning) та використання моделі для ранжування гіпотез виправлень, створених іншими моделями. Автори оцінюють продуктивність GPT-3.5 через методи автоматичного оцінювання, такі як оцінювання граматичності за допомогою мовних моделей (LM), тест Скрібенді та аналіз семантичних вкладень. Результати демонструють, що для англійської мови GPT-3.5 забезпечує високий рівень узгодженості виправлень, зберігаючи семантичну цілісність тексту, завдяки чому модель добре виправляє помилки та генерує плавні речення. Проте в інших мовах, разом з українською, чеською, німецькою, російською та іспанською, GPT-3.5 часто істотно змінює вихідні речення, що іноді призводить до зміни їхньої семантики та створює складнощі для оцінювання. Основною особливістю цієї роботи є виявлення тенденції GPT-3.5 до надмірної корекції, а також аналіз її обмежень у виправленні помилок пунктуації, узгодженні часових форм, синтаксичних залежностей та лексичної сумісності. Незважаючи на потужні можливості моделі, автори наголошують на проблемах її адаптації для задач GEC у багатомовному середовищі, особливо для мов із багатою граматичною структурою, таких як українська.

У роботі [11] автори запропонували новий підхід до багатомовної ідентифікації і корекції граматичних помилок (GEC), який ґрунтується на використанні попередньо натренованих моделей машинного перекладу. Особливістю їх підходу є інтеграція методів нейронного машинного перекладу у завдання, що дало змогу ефективно адаптувати моделі для роботи з текстами різними мовами, зокрема низькоресурсними. Автори досягли виходу за межі обмежень традиційних GEC-моделей завдяки оптимізації попередньо натренованих трансформерів, що забезпечило високу точність коригування, урахування контексту та природність виправленого тексту. Основна відмінність їхньої роботи полягає у тому, що вони адаптували моделі перекладу, зокрема трансформери, для багатомовної корекції, зберігши їх універсальність і здатність масштабуватися на нові мовні набори. Це дає змогу не лише виявляти помилки з високою точністю, але й ефективно працювати з мовами, для яких доступні обмежені навчальні ресурси.

Результати дослідження та їх обговорення / Research results and their discussion

Система підтримки прийняття рішень для ідентифікації помилок в україномовних текстах ґрунтується на використанні методів обробки природної мови (NLP) та машинного навчання. Важливим аспектом є можливість інтеграції додаткових компонентів, таких як лексичні бази даних, семантичні мережі та алгоритми підвищення точності аналізу тексту. Модуль ідентифікації помилок виконує аналіз тексту для виявлення некоректних слів, граматичних, орфографічних і стилістичних помилок. Його основні завдання: розподіл вхідного тексту на окремі речення; видалення стоп-слів, які не мають змістового навантаження і лише сповільнюватимуть ідентифікацію помилок; токенізація тексту; лематизація чи стемінг; класифікація слів на правильні та потенційно помилкові й контекстний аналіз через оцінювання ймовірності помилки на основі навколишніх слів.

Система отримує на вхід текст у різних формах, який може містити орфографічні, граматичні, пунктуаційні та стилістичні помилки. Допоміжні параметри: формат тексту (plain text, HTML, XML, JSON); контекстні метадані (жанр тексту (науковий, публіцистичний, розмовний), стиль (формальний / неформальний)).

Опрацювання тексту передбачає кілька етапів аналізу. Основні проміжні параметри: лінгвістичні особливості, зокрема POS-теги $P(\text{xi})$, морфологічні ознаки $M(xi)$, синтаксичні залежності $S(xi, xj)$; ймовірнісні характеристики; кандидати на виправлення, мітки помилок. До вихідних параметрів належать перелік знайдених помилок, їх тип, а також метрики оцінювання.

Розробляючи системи підтримки прийняття рішень для ідентифікації помилок в україномовних текстах, важливо враховувати низку обмежень, що впливають на її функціональність, а також чітко визначити критерії ефективності, які дають змогу оцінити якість її роботи. Ці аспекти є визначальними для побудови надійної та точної математичної моделі.

Одним із ключових обмежень є лінгвістична специфіка текстів, з якими працює система. Українська мова має багату морфологічну структуру, велику кількість граматичних правил, а також численні винятки, що ускладнює процес автоматизованого аналізу. Важливим викликом стає обробка полісемії – випадків, коли одне слово має кілька значень залежно від контексту. Окрім того, система може стикатися із труднощами під час опрацювання неологізмів, професійної лексики або діалектних особливостей, які не завжди входять до навчальних корпусів.

Технологічні обмеження також відіграють важливу роль у роботі системи. Одним із критичних аспектів є продуктивність алгоритмів, що використовують для аналізу тексту. Деякі сучасні моделі опрацювання природної мови, такі як нейромережеві трансформери, потребують значних обчислювальних ресурсів, що може створювати проблеми у разі їх застосування у реальному часі або на пристроях з обмеженою обчислювальною потужністю. Крім того, довжина аналізованого тексту може накладати обмеження на ефективність роботи системи, оскільки великі фрагменти доводиться ділити на менші частини, що може знижувати точність контекстного аналізу.

Визначення критеріїв ефективності є важливим етапом оцінювання якості системи. Основним показником є точність виявлення та корекції помилок, що оцінюють за допомогою таких метрик, як precision (точність), recall (повнота) та F1-міра. Висока точність означає, що система виявляє саме ті помилки, які справді містяться в тексті, а

висока повнота вказує на здатність розпізнавати якомога більше помилок без пропусків. Оптимальним варіантом є баланс між цими показниками, оскільки надмірна кількість виявлених помилок без достатньої точності може призводити до помилкових виправлень.

$$A = \frac{TP + TN}{TP + TN + FP + FN}, \quad (1)$$

$$R = \frac{TP}{TP + FN}, \quad (2)$$

$$P = \frac{TP}{TP + FP}, \quad (3)$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (4)$$

де A – точність; R – повнота; P – точність; TP – кількість правильно визначених помилок; TN – кількість правильних виправлень; FP – кількість помилково ідентифікованих помилок; FN – кількість пропущених помилок.

Важливим аспектом є ресурсозатратність, яку оцінюють за використанням оперативної пам'яті та процесорного часу. Ще один важливий аспект – якість виправленого тексту. Навіть якщо система правильно ідентифікує помилки, її виправлення повинні бути релевантними та відповідати стилю й семантиці початкового тексту.

Завдання ідентифікації полягає у виявленні фрагментів тексту, які можуть містити орфографічні, граматичні, лексико-семантичні чи синтаксичні помилки, а також у визначенні типу цих помилок. Це є ключовим етапом у системах автоматичної корекції текстів, де ефективність коригування безпосередньо залежить від точності та якості ідентифікації.

Найважливіші дані модуля формалізуються як:

$$X = \{x_1, x_2, \ldots, x_n\}, \quad (5)$$

$$Y = \{y_1, y_2, \ldots, y_N\}, \quad (6)$$

$$C = \{c_1, c_2, \ldots, c_K\}, \quad (7)$$

де X – вхідний текстовий масив, що складається з n токенів (слів, символів, речень); x_i – окремий токен або слово у тексті, y_i – мітки помилок, C – множина помилок, де c_0 позначає "без помилки", а $(c_1, \ldots, c_K)$ – конкретні типи помилок.

$$P(y_i = c_k \mid x_i, context) = f_{id}(x_i, context) \quad (8)$$

де f_{id} – це функція розпізнавання (ідентифікації), що враховує як токен x_i, так і його контекст *context*. Основне завдання моделі ідентифікації – навчитися точно оцінювати функцію f_{id}, яка зможе правильно визначати тип помилки або її відсутність.

Українська мова має специфічні риси, які істотно ускладнюють ефективну ідентифікацію помилок. По-перше, це морфологічна складність, яка полягає у багатстві форм словозміни. Наприклад, прикметники, дієслова чи іменники можуть змінюватися за родами, числами та відмінками, що ускладнює виявлення як орфографічних, так і граматичних помилок. По-друге, контекстна багатозначність українських слів, тобто слова з однаковим написанням можуть мати різні значення чи функції залежно від контексту. Наприклад, слово "пішли" може належати до дієслівної форми минулого часу або слугувати розмовною формою для опису руху. Саме тому система ідентифікації повинна враховувати не лише локальний аналіз токена, а й синтаксичні та семантичні зв'язки у межах усього речення.

Крім того, ускладнює ідентифікацію варіативність регіональних діалектів, запозичень, а також використання розмовних та неформальних конструкцій, характерних для української мови. Наслідком такого лінгвістичного розмаїття є необхідність адаптації моделей класифікації до датасетів, які містять приклади як стандартної мови, так і розмовних форм. Окрім цього, важливо зазначити, що ідентифікація помилок у текстах тісно пов'язана із опрацюванням різних типів помилок. Орфографічні помилки можна легко помітити за допомогою лексичних словників (наприклад, перевірка непізнаних слів), тоді як граматичні або пунктуаційні помилки залежать від складної структурної залежності між словами у реченні.

Отже, завдання ідентифікації помилок в україномовному тексті – багаторівнева проблема, яка потребує використання складних математичних моделей, здатних ураховувати як контекстні залежності між словами, так і особливості морфології та синтаксису. Ефективність таких моделей великою мірою залежить від якості машинного навчання, яке спирається на великі корпуси текстів із чітко анотованими помилками. Оскільки моделі, застосовувані для вирішення цієї проблеми, здебільшого ґрунтуються на алгоритмах машинного навчання та глибоких нейронних мережах, текст необхідно переводити у форму, якою можуть скористатися алгоритми для обчислення. З цією метою текст подають через математично формалізовані дані, які враховують як самі токени, так і їхній контекст.

Для роботи із сучасними методами глибокого навчання текст необхідно подавати у вигляді багатовимірних векторів, які є числовими представленнями тексту. Формат потоку вхідних даних подано як:

$$W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix}, \quad (9)$$

де (w_i) – це векторне представлення токена (x_i), (x_i), $W \in R^{\{N \times d\}}$ – матриця таких представлень. Векторизація тексту передбачає перетворення слів чи токенів на багатовимірні числові вектори – ембедінги (embeddings), що зберігають семантичну чи контекстну інформацію про слова. Використовують два основні підходи до векторизації: статичні та динамічн. У підході

статичних векторів уявлення кожне слово з тексту має постійний вектор, який не залежить від його контексту. Методи, як-от Word2Vec або FastText, представляють слова на основі їхньої співзустрічності в тексті. Наприклад, у методі FastText враховуються підслова (n-грамні структури) для моделювання винятків, таких як помилки чи незнайомі токени.

Динамічні контекстуальні векторні уявлення – сучасніший підхід, вони враховують контекст, у якому використовується слово. Популярні моделі, як-от BART, XLM-RoBERTa, або mBERT, генерують такі динамічні вектори, де для кожного токена (x_i) у тексті формується: $w_i = f(x_i, context)$, де (f) – трансформерна модель, яка враховує контекст у межах усього речення (або навіть тексту). У цьому підході вектор (w_i) не лише передає зміст слова, але й ураховує залежності між іншими токенами. Це критично важливо для завдання ідентифікації помилок, оскільки деякі помилки можна ідентифікувати лише у межах ширшого контексту.

Для ідентифікації помилок недостатньо лише аналізувати індивідуальні слова або токени. Потрібно враховувати їхню позицію у реченні та зв'язок із сусідніми елементами. Контекст моделюється такими способами: послідовність токенів обробляється рекурентними або трансформерними нейронними мережами, які зберігають інформацію про попередні та наступні токени; використання спеціальних механізмів, таких як механізм уваги (attention), що дають змогу виділити значущі частини тексту для конкретного токена.

Математично контекст можна подати як комбінацію токенів у вікні довжиною (m) навколо аналізованого токена (x_i).

$$context(x_i) = \{x_{i-m}, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{i+m}\}, \qquad (10)$$

де m – довжина вікна контексту. У трансформерних архітектурах довжина контексту може охоплювати все речення або навіть кілька речень. Описуючи задачу ідентифікації у формалізованому вигляді, можна структурувати вхідні дані у вигляді двох взаємопов'язаних компонентів: матриці векторів слів та міток помилок.

Формалізація вхідних даних має враховувати специфіку української мови, зокрема відмінювання, дієвідміни, узгодження за родами і числами. Для українських текстів характерні скорочення ("тр", "д/з") та регіоналізми, що також важливо враховувати. Також слід забезпечувати адаптивність до різних варіацій української мови, які можна виявити в розмовних і художніх текстах.

Основне завдання під час побудови системи ідентифікації помилок в україномовних текстах – розроблення математичної моделі класифікації, здатної визначати, чи містить текстовий фрагмент (токен або речення) помилку, а якщо так – класифікувати її за типом. Оскільки класифікація помилок є проблемою багатокласової класифікації, математична модель повинна визначати ймовірність належності кожного фрагмента до певного класу (ck).

Ідентифікація помилок у тексті – складний процес, який передбачає кілька етапів опрацювання. Ключове завдання – пошук тих токенів або сегментів тексту, які потенційно можуть містити помилки, на основі формальних критеріїв, мовних правил і статистичних особливостей. Алгоритм роботи системи ідентифікації складається із кількох обов'язкових етапів:

1. Попереднє опрацювання тексту: нормалізація тексту, токенізація, лематизації або стемінгу. Цей етап передбачає видалення будь-якого шуму, який може заважати правильному навчанню моделі.

2. Аналіз синтаксичної та морфологічної структури тексту. На цьому етапі текст обробляють для виокремлення його граматичних і синтаксичних залежностей. Для цього виконують POS-тагування (розпізнавання частин мови), побудову синтаксичного дерева. Для речення створюють синтаксичну структуру, яка моделює взаємозалежності між словами.

3. Створення мовної моделі для аналізу контексту. Після синтаксичного аналізу текст сегментується у вікна контексту, які дають змогу враховувати сусідні токени під час ідентифікації помилок. Контекст критично важливий, оскільки багато помилок неможливо виявити без урахування граматичних і смислових залежностей між словами.

4. Виділення токенів-кандидатів для корекції. На цьому етапі система ідентифікує токени, які потенційно можуть містити помилки. Цього досягають за допомогою сукупності правил, лексичних знань і статистичних особливостей. Токенами-кандидатами можуть бути слова, фрази або символи, які задовольняють один чи кілька критеріїв, що сигналізують про можливу помилку. Визначення таких токенів ґрунтується на комбінованому підході, що враховує як лінгвістичні правила, так і результати статистичного аналізу даних, а саме лексичного аналізу, аналізу граматичних узгоджень, пунктуаційного аналізу, статистичної частоти та контекстної невідповідності.

5. Фільтрація кандидатів. Після початкового виділення потенційних помилок система виконує фільтрацію токенів-кандидатів для вилучення хибно позитивних прогнозів. Це забезпечує зниження навантаження на наступний етап – модуль корекції, де опрацьовуються лише ймовірні помилки.

У результаті роботи цього алгоритму визначають підмножину токенів, які є кандидатами для коригування. Їх передають до системи наступного рівня аналізу, яка може приймати конкретне рішення щодо виправлення.

Для візуалізації роботи алгоритму використано DFD діаграму, що являє собою схематичний інструмент для візуалізації потоків даних у системі. DFD дає змогу зрозуміти основні потоки даних у системі й допомогти чітко і структурно відобразити процес ідентифікації помилок.

На рівні DFD 0 система представлена як єдиний блок, що взаємодіє із зовнішніми сутностями. На ній видно, що користувач вводить текст у систему. Система

повертає виведення: список токенів із мітками помилок та типи помилок.

Далі деталізується потік даних, це і є рівнем DFD рівень 1: систему розділяють на основні функціональні блоки.

DFD рівень 2 – розширений алгоритм. На цьому рівні відображаються всі операції, що відбуваються всередині кожного блока.

Для навчання моделі необхідний корпус текстів, що містить позначені мовні помилки. Джерела таких корпусів: реальні тексти, зокрема студентські есе, наукові роботи з помилками, соціальні мережі, блоги, особисті повідомлення, онлайн-форуми та анотовані корпуси, як корпуси для української мови (UA GEC, GRAC [12] та інші), де тексти вже анотували фахівці для задач коригування.

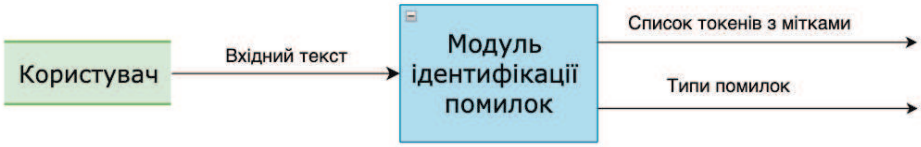

Рис. 1. Діаграма DFD рівня 0 модуля ідентифікації помилок / DFD level 0 of the error identification module

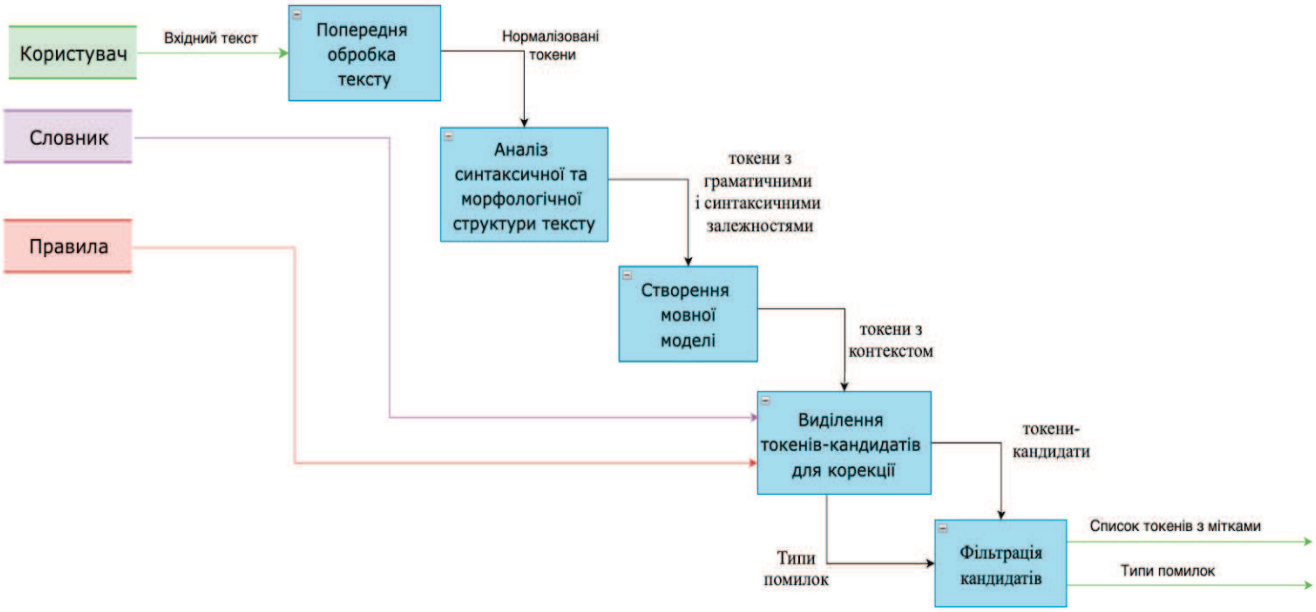

Рис. 2. DFD рівня 1 для модуля ідентифікації помилок / DFD level 1 for the error identification module

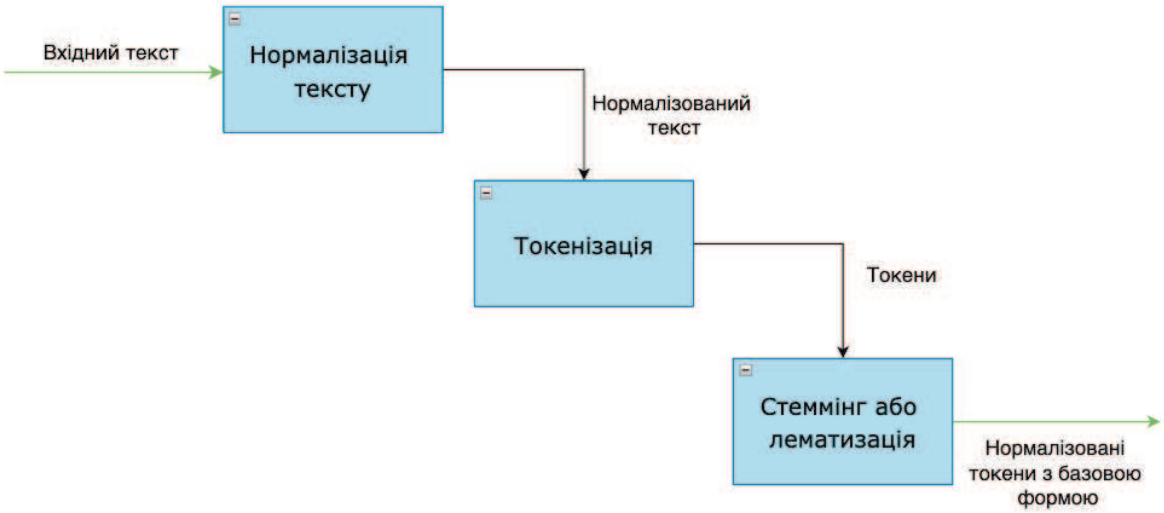

Рис. 3. DFD рівня 2 блока “Попередня обробка тексту” / DFD level 2 diagram of the “Text Preprocessing” block

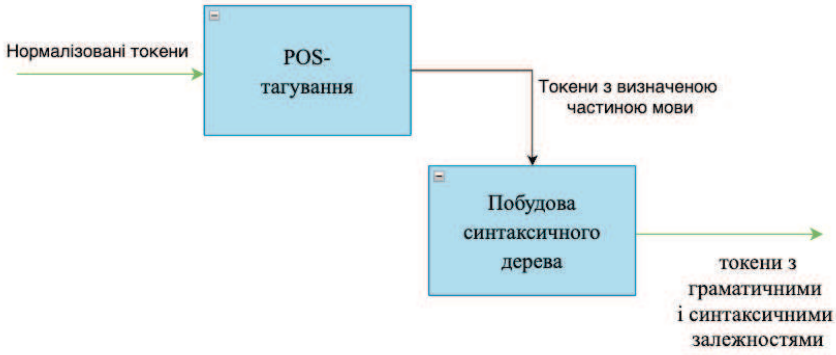

Рис. 4. DFD рівня 2 блока "Аналіз синтаксичної та морфологічної структури тексту" / DFD level 2 block "Analysis of syntactic and morphological structure of the text"

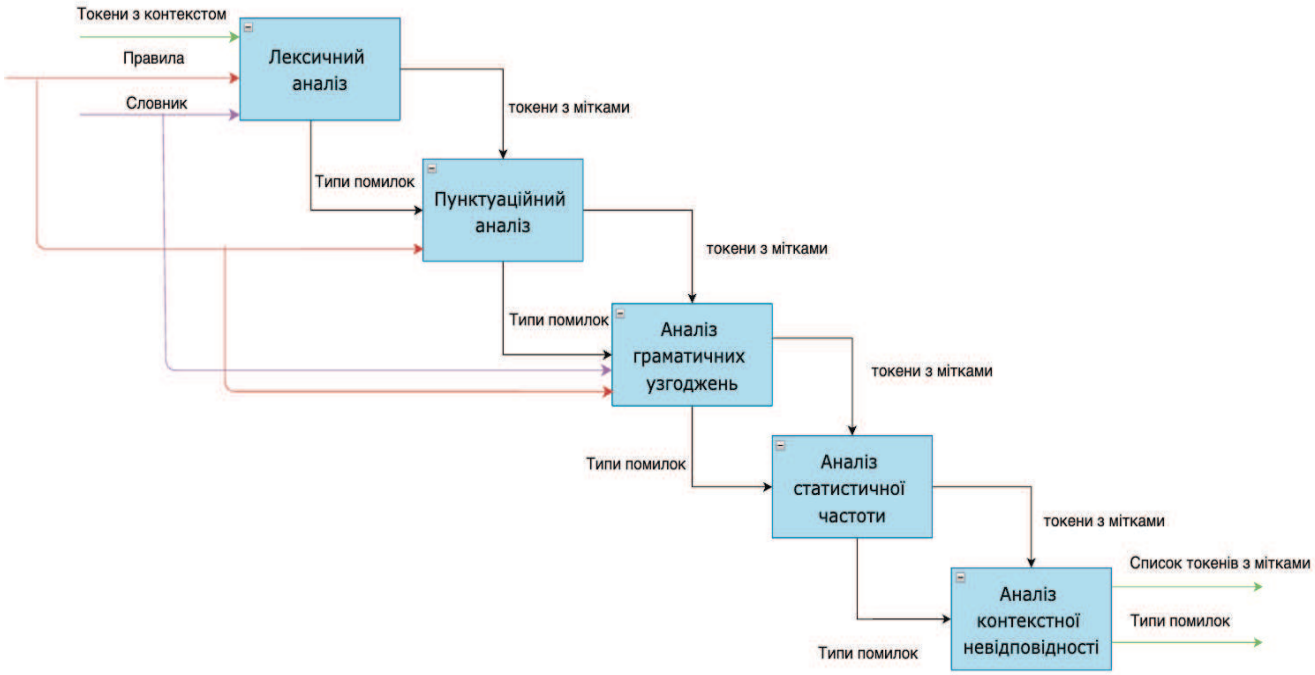

Рис. 5. DFD рівня 2 блока "Виділення токенів-кандидатів для корекції" / DFD level 2 of the block "Selecting candidate tokens for correction"

Тексти без помилок також важливі для побудови навчальної вибірки. Їх використовують, щоб ввести штучні помилки для моделювання реальних сценаріїв та для того, щоб навчити модель правильно "читати" контекст і визначати непроблемні ділянки тексту. Тут джерелами можуть бути книги, формальні документи, новини; Вікіпедія та інші відкриті бази даних текстів; художні твори українською мовою.

У навчальній вибірці повинні міститися різноманітні типи помилок. До основних категорій належать орфографічні (випадково додані букви, заміна однієї букви на іншу, пропуск літер або зайві літери), граматичні (неправильне узгодження слів, помилки відмінювання чи дієвідмінювання, неправильний порядок слів) пунктуаційні, лексичні та контекстуальні помилки. Часто постає питання про збільшення розміру навчальної вибірки. Для цього можна використати автоматичну генерацію помилок із використанням правил чи статичних алгоритмів або ж готові мовні моделі (GPT, T5) для генерації текстів із типовими помилками. Для автоматичного генерування помилок можна використовувати алгоритм випадкового внесення помилок, замінюючи літери на основі частотності реальних помилок, видалення або додавання символів у слова, пропускання ком або додавання зайвих розділових знаків. Також часто використовують заміну правильного відмінка на випадковий або випадкове додавання слова не з контексту.

Під час навчання моделі налаштовують такі гіперпараметри: розмір вікна контексту, що визначає, скільки сусідніх токенів враховувати під час аналізу; розмір ембедингів (64–512); кількість епох, яку зазвичай вибирають експериментально, залежно від швидкості досягнення збіжності (3–20), розмір батчу, що являє собою кількість текстів, оброблюваних за один раз (4–64); темп навчання, який знижуватиметься зі зростанням

кількості епох. Після навчання модель стає здатною точно класифікувати токени тексту за типом помилки.

У межах цього дослідження для задачі автоматичного виправлення помилок в україномовних текстах вибрано модель *youscan/ukr-roberta-base* [13], що являє собою трансформер на основі архітектури RoBERTa, попередньо навчений на великому корпусі україномовних текстів, зокрема новин, соцмереж, Вікіпедії та інших джерел. Її навчання було спрямоване на мовне моделювання, тобто передбачення замаскованих токенів у реченнях, що дало змогу моделі глибоко засвоїти граматичні, лексичні та синтаксичні закономірності української мови.

Навчання виконано в середовищі Kaggle [14] із доступним GPU та обмеженим обсягом оперативної пам'яті. Для навчання використано корпус UA-GEC, який містить приклади речень із типових помилок української мови. Попереднє оброблення даних здійснено за допомогою бібліотеки Stanza [15] у форматі пар: вхідне речення (source) та мітка (label), яка вказує, чи містить речення помилку. Мітка набуває значення 1, якщо в реченні є помилка, і 0, якщо помилок не виявлено.

Для навчання моделі вибрано такі основні гіперпараметри: кількість епох n_epochs = 3, розмір батчу batch_size = 16, довжина вхідної послідовності max_token_length = 256, а також початкове значення швидкості навчання learning_rate = 3e–5. Ці параметри типові для навчання трансформерних моделей на завданнях опрацювання природної мови, де збереження стабільності градієнтів і узгодженість оновлень ваг мають критичне значення. Вибір помірного значення learning_ rate забезпечив поступове зменшення функції втрат без надмірних коливань.

Упродовж трьох епох спостерігалося систематичне зменшення функції втрат на тренувальній вибірці: з 0,5490 на першій епосі до 0,3017 на третій. Це свідчить про те, що модель навчалася узгоджено та успішно узагальнювала закономірності в даних. На валідаційній вибірці втрати зменшилися з 0,5132 до 0,4845 між першою та другою епохами, однак на третій епосі спостерігалося зростання до 0,5876, що вказує на потенційне перенавчання. Метрики додатково підтверджують зазначену динаміку. Значення accuracy поступово зростало, досягаючи максимуму 0,763 на третій епосі. Аналогічно зростали і значення precision та f1-score. Однак показник recall залишався стабільним між другою і третьою епохами (≈0,649), що може свідчити про насичення навчання та зменшення здатності моделі виявляти нові типи помилок.

Табл. 1. Значення гіперпараметрів для навчання моделі протягом ітерацій / Hyperparameter values for training the model over iterations

Гіперпараметр	Ітерація 1	Ітерація 2
Кількість епох	2	2
Розмір батчу	16	16
Максимальна довжина токенів	256	256
Швидкість навчання	3e–5	1e–5

З урахуванням зростання валідаційних втрат на третій епосі та незначного приросту f1 було прийнято рішення використати чекпоінт, збережений після другої епохи. Саме на цьому етапі модель досягла найкращого балансу між точністю, повнотою та функцією втрат на валідації. Для подальшого навчання вибрали менше значення learning_rate, що дасть моделі змогу продовжити оптимізацію з меншими кроками, знижуючи ризик перенавчання і сприяючи кращому узагальненню на нових прикладах.

Потім було зроблено спробу донавчання моделі з меншою швидкістю навчання learning_rate = 2e-5. Проте, попри зниження тренувальних втрат, значення Validation Loss зросло до 0,6936, що свідчить про погіршення узагальнювальної здатності моделі.

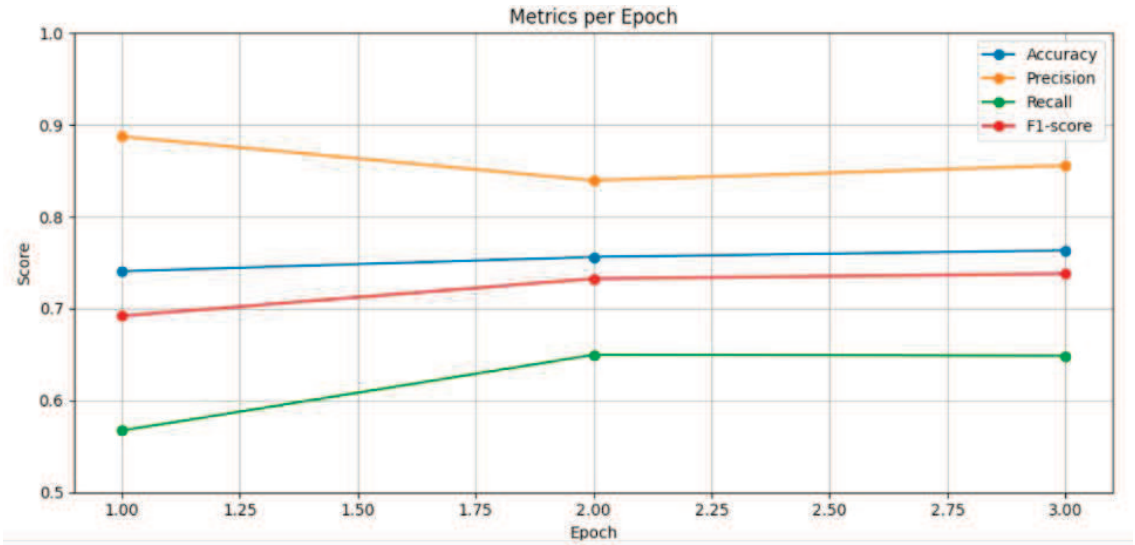

Рис. 6. Графік втрат і метрик навчання моделі на UA_GEC корпусі з learning_rate = 3e-5 / Graph of model training losses and metrics on the UA_GEC corpus with learning_rate = 3e-5

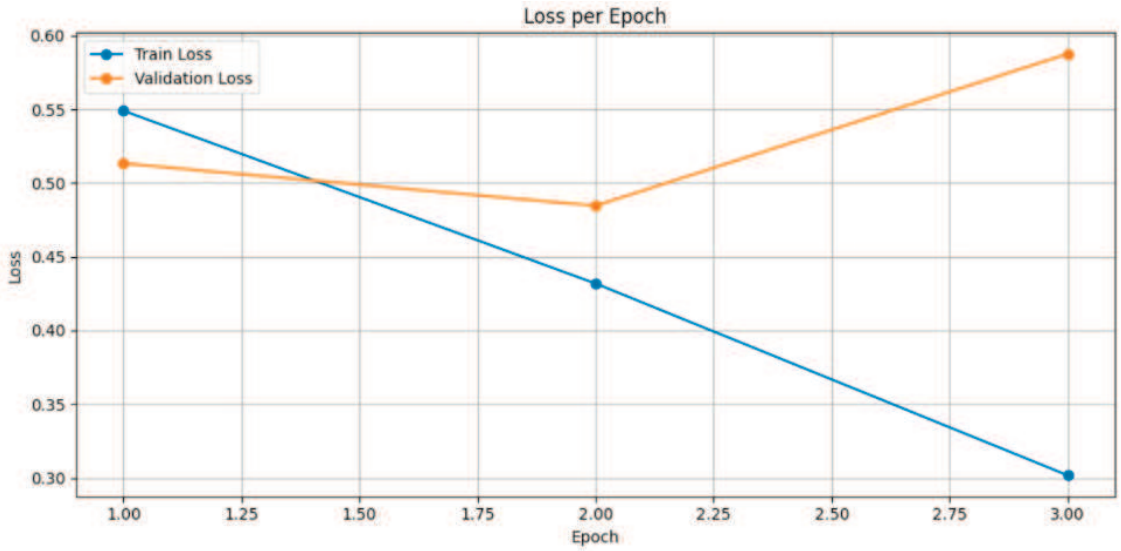

Рис. 7. Графік втрат і метрик навчання моделі на UA_GEC корпусі з learning_rate = 3e-5 / Graph of model training losses and metrics on the UA_GEC corpus with learning_rate = 3e-5

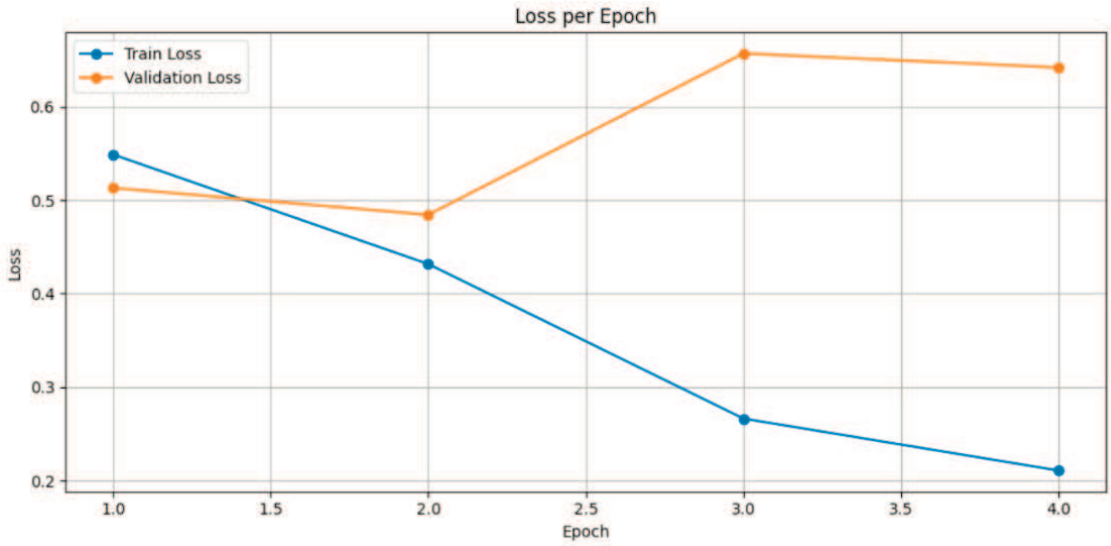

Рис. 8. Графік втрат навчання моделі на UA_GEC корпусі з learning_rate = 1e-5 / Graph of model training loss on the UA_GEC corpus with learning_rate = 1e-5

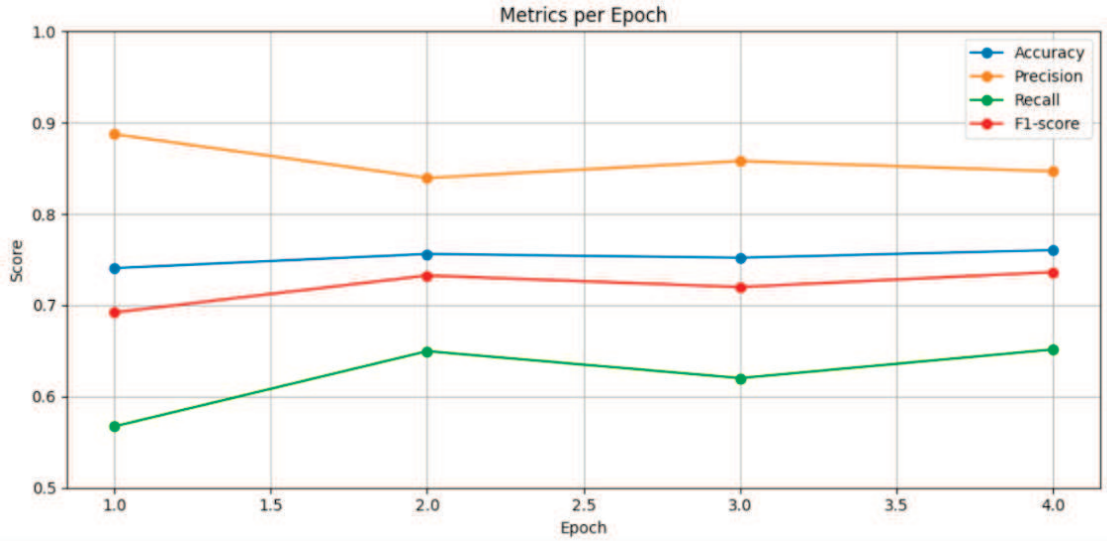

Рис. 9. Графік метрик навчання моделі на UA_GEC корпусі з learning_rate = 1e–5 / Graph of model training metrics on the UA_GEC corpus with learning_rate = 1e–5

Після невдалого зменшення швидкості навчання до 2e–5 ми вирішили спробувати ще нижче значення – 1e–5. У цьому режимі модель досягла кращих результатів на другій епосі, зменшивши Validation Loss до 0,6421 і показавши найбільше значення F1-метрики – 0,7364. Третя епоха призвела до погіршення як Validation Loss, так і метрик, що свідчить про початок перенавчання. Отже, найкращий результат отримано у разі комбінації двох епох із learning_rate = 3e–5, після чого досягнення оптимального рівня тривало ще упродовж двох епох із зменшеним learning_rate = 1e–5. Модель для виправлення граматичних помилок показала добрі результати під час аналізування простих українських речень, однак деякі неточності все ще залишаються.

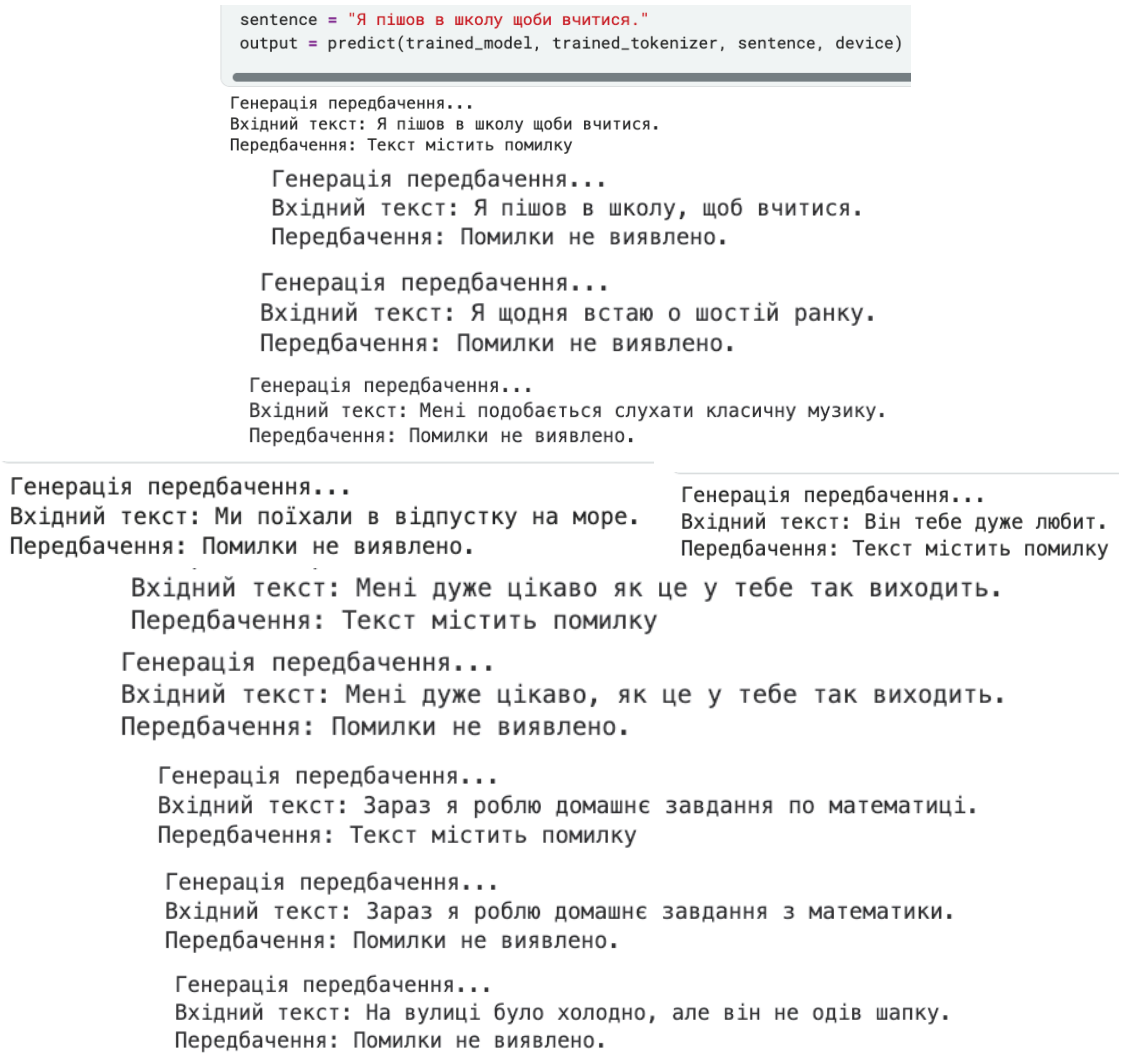

Рис. 10–20. Результати роботи моделі після навчання / Model performance results after training

У першому прикладі “Мені подобається слухати класичну музику” модель обгрунтовано не виявила помилок, оскільки це граматично правильне речення. Модель правильно обробляє стандартні речення без помилок, що свідчить про її здатність визначати граматичну правильність у простих конструкціях. В іншому прикладі, де вжито фразу “Я пішов в школу щоби вчитися”, модель правильно виявила помилку у використанні частки “щоби”. Правильний варіант у цьому випадку – “щоб”. Це свідчить про здатність моделі виявляти типові граматичні помилки, що стосуються використання часток у складних реченнях. Проте модель не виявила помилки у реченні “Ми поїхали в відпустку на море”. Це вказує на те, що модель може не завжди правильно визначати контексти, у яких допускають помилки, типові для розмовного стилю.

У прикладі “Він тебе дуже любит” модель правильно виявила орфографічну помилку. У реченні “На вулиці було холодно, але він не одів шапку” модель не виявила помилки, хоча в ньому орфографічна помилка також є. Це свідчить про те, що модель може пропускати деякі орфографічні помилки. Також модель правильно виявила помилку в реченні “Мені дуже цікаво як це у тебе так виходить”, де необхідно поставити кому перед “як”. Це свідчить про здатність моделі працювати з пунктуацією, зокрема коли коми потрібні для відокремлення частин складного речення. Приклад “Зараз я роблю домашнє завдання по математиці” модель опрацювала правильно, виявивши помилку у використанні прийменника. Правильний варіант – “з математики”, що вказує на здатність моделі знаходити важчі помилки у використанні прийменників.

Загалом модель показала добрі результати в розпізнаванні граматичних помилок, зокрема в контекстах, що стосуються орфографії, пунктуації та вживання часток. Однак деякі аспекти, як-от опрацювання розмовної мови або контексту специфічних виразів, залишаються для моделі складними. Перспективи удосконалення передбачають покращення здатності моделі розпізнавати складніші помилки, тобто у варіантах, які виникають у розмовному мовленні, або у регіональних варіантах. Подальші покращення можуть передбачати додаткове навчання на корпусах, що охоплюють різні стилі мовлення, та оптимізацію для роботи із помилками, характерними для неформальної української мови. І навіть більше, важливе навчання моделі не лише на реченнях, а й на словосполученнях і словах, щоб забезпечити краще навчання моделі.

Обговорення отриманих результатів. Розроблено математичну модель ідентифікації помилок у текстах українською мовою. Здійснено формалізацію потоку даних через побудову діаграм потоків даних (DFD), які відображають взаємодію між основними компонентами системи. Здійснено донавчання трансформерної моделі youscan/ukr-roberta-base, адаптованої для виконання класифікації речень за наявністю мовних помилок. Найкращих показників якості було досягнуто за стратегії поетапного донавчання із поступовим зменшенням learning_rate, що дало змогу уникнути перенавчання та покращити узагальнювальну здатність моделі. Зокрема, за результатами останньої ітерації навчання досягнуто F1-міри 0,736, що є обнадійливим результатом для задачі бінарної класифікації мовленнєвих помилок у текстах.

Наукова новизна отриманих результатів – розроблено математичну модель для задачі ідентифікації помилок, що ґрунтується на ймовірнісному підході до класифікації речень, де кожне вхідне речення подається у вигляді векторного простору, а мета моделі – оцінити ймовірність наявності помилки на основі контексту; запропоновано представлення даних у вигляді пар "речення – 0/1" (відсутність або наявність помилки), що дає змогу застосовувати сучасні моделі трансформерного типу без токенізованої розмітки.

Практична значущість отриманих результатів полягає у створенні ефективного інструменту, який може бути інтегрований у системи попередньої перевірки текстів перед подальшим глибоким обробленням. Запропонована модель може слугувати першим фільтром для систем автоматичного редагування, зменшуючи кількість речень, які потребують складнішої синтаксичної чи семантичної перевірки, тим самим оптимізуючи роботу всієї системи.

Висновки / Conclusions

Результатом роботи є математична модель ідентифікації помилок в україномовних текстах, орієнтована на підходи машинного навчання. У підсумку здійснених досліджень сформульовано математичне підґрунтя для вирішення поставлених завдань, ураховуючи формалізацію потоку даних, розміщення компонентів системи та подання текстів у вигляді, придатному для машинного оброблення.

Побудовано математичну модель для задачі ідентифікації помилок, що спирається на ймовірнісний підхід. Основний акцент у моделюванні зроблено на збереженні контексту та врахуванні залежностей між токенами для підвищення точності ідентифікації. Побудовано математичну модель, яка передбачає обчислення умовної ймовірності вибору найкращого кандидата для виправлення помилкового токена в межах певного контексту тексту. Розглянуто підхід до підготовки даних через векторизацію текстів, формування навчальної вибірки та організацію процесу навчання моделей на основі великих корпусів текстів. Наголошено на важливості побудови репрезентативного корпусу навчальних даних, що міститиме тексти із реальними помилками, а також штучно змодельовані приклади помилок із урахуванням їхнього типологічного розподілу.

Запропоновано архітектуру, орієнтовану на машинне навчання із використанням трансформерної моделі youscan/ukr-roberta-base. Модель донавчалась на корпусі UA-GEC, де кожен приклад був представлений парою "речення – клас (0 або 1)". У ході експериментів випробовували різні значення learning_rate, і найкращих результатів досягли за стратегії: дві епохи з learning_rate = 3e–5 та ще дві епохи з learning_rate = 1e–5, використовуючи збереження другої епохи як стартову точку. Отримано F1 = 0,736, accuracy = 0,76, що демонструє стабільну здатність моделі до виявлення помилок у простих реченнях.

У підсумку розроблена математична модель є універсальним підходом до опрацювання українських текстів, що дає змогу вирішувати завдання ідентифікації помилок. Визначені формальні аспекти взаємодії між компонентами моделі створюють підґрунтя для її ефективного навчання й подальшого впровадження у практичні завдання.

Подяка

Статтю підготовано завдяки грантовій підтримці Національного фонду досліджень України, реєстраційний номер проєкту 33/0012 від 3/03/2025 (2023.04/0012) "Розроблення інформаційної системи автоматичного виявлення джерел дезінформації та неавтентичної поведінки користувачів чатів" за конкурсом "Наука для зміцнення обороноздатності України".

References

1. Bryant, C., Yuan, Z., Qorib, M. R., Cao, H., Ng, H. T., & Briscoe, T. (2023). Grammatical Error Correction: A Survey of the State of the Art. *Computational Linguistics*, 49(3), 643–701. https://doi.org/10.48550/arXiv.2211.05166, https://doi.org/10.1162/coli_a_00478
2. Smith, O. B., Ilori, J. O., & Onesirosan, P. (1984). The proximate composition and nutritive value of the winged bean Psophocarpus tetragonolobus (L.) DC for broilers. *Animal Feed Science and Technology*, 11(1), 231–237. https://doi.org/10.1016/0377-8401(84)90066-X

3. Grammarly Inc. Free Grammar Checker. Retrieved from: https://www.grammarly.com/grammar-check
4. Meet UA-GEC – a grammar correction dataset for the Ukrainian language. Retrieved from: https://dou.ua/forums/topic/33272/
5. Syvokon, O., Nahorna, O., Kuchmiichuk, P., & Osidach, N. (2023). UA-GEC: Grammatical Error Correction and Fluency Corpus for the Ukrainian Language. Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), 96–102. https://doi.org/10.18653/v1/2023.unlp-1.12
6. Syvokon, O., & Nahorna, O. (2021). UA-GEC: Grammatical Error Correction and Fluency Corpus for the Ukrainian Language. ArXiv. https://doi.org/10.48550/arXiv.2103.16997
7. Writing correctly is easy. OnlineCorrector. Retrieved from: https://onlinecorrector.com.ua/ [in Ukrainian]
8. Omelianchuk, K., Atrasevych, V., Chernodub, A. N., & Skurzhanskyi, O. (2020). GECToR – Grammatical Error Correction: Tag, Not Rewrite. *ArXiv*. https://doi.org/ 10.48550/arXiv.2005.12592, https://doi.org/10.18653/v1/2020.bea-1.16
9. HuggingFace. Transformers Documentation. Retrieved from: https://huggingface.co/docs/transformers/index
10. Katinskaia, A., & Yangarber, R. (2024). GPT-3.5 for Grammatical Error Correction. *ArXiv*. https://doi.org/10.48550/arXiv.2405.08469
11. Luhtaru, A., Korotkova, E., & Fishel, M. (2024). No Error Left Behind: Multilingual Grammatical Error Correction with Pre-trained Translation Models. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Vol. 1: Long Papers), 1209–1222. Association for *Computational Linguistics*. https://doi.org/10.18653/v1/2024.eacl-long.73
12. Shvedova, M., et al. (2017–2022). General Regionally Annotated Corpus of Ukrainian Language (GRAC). Network for Ukrainian Studies Jena. *ArXiv*. https://doi.org/10.48550/arXiv.1911.02116
13. Ukrainian RoBERTa base model. Hugging Face. Retrieved from: https://huggingface.co/youscan/ukr-roberta-base
14. Kaggle. Learn Documentation. Retrieved from: https://www.kaggle.com/learn
15. Stanza – A Python NLP Package for Many Human Languages. Stanford NLP Group. Retrieved from https://stanfordnlp.github.io/stanza

R. B. Fedchuk, V. A. Vysotska
Lviv Polytechnic National University, Lviv, Ukraine

MATHEMATICAL MODEL OF ERRORS IDENTIFICATION IN TEXTS OF UKRAINIAN CONTENT

The problem of automated error detection in Ukrainian texts is becoming particularly relevant in the context of the growth of digital content. A mathematical model of a decision support system for detecting errors in Ukrainian-language texts has been developed. The process of error identification has been studied as a multi-class classification task at the token level, considering the context of the text. The use of probabilistic models has been proposed to determine the type of error depending on the environment of tokens in the text. The feasibility of forming training samples containing both real and artificially created errors has been identified to ensure a balanced learning process. The effectiveness of approaches to vectorizing texts has been established, considering the morphological and syntactic structure of the Ukrainian language, which increases the accuracy of the model. It has been found that the integration of contextual information significantly improves the results of error identification. Detailed DFD diagrams have been constructed that formalize the processes of the system's functioning and the interaction of its components. Experimental training of the ukr-roberta-base model on the UA-GEC corpus for the task of identifying errors in Ukrainian texts has been carried out. The following model quality results were obtained: F1 – 0.736, accuracy – 0.76, precision – 0.85, recall – 0.65. Examples of the model's performance on test data are provided. It was found that the model has already learned to detect punctuation and basic spelling errors, which indicates its effectiveness and prospects for further development. Prospects for further research include scaling the developed model and adapting it to expand the coverage of more complex types of language errors.

***Keywords**:* natural language processing, error identification, classification, machine learning, Ukrainian language, probabilistic model.

Інформація про авторів:

Висоцька Вікторія Анатоліївна, д-р техн. наук, професор, кафедра інформаційних систем та мереж.
Email: victoria.a.vysotska@lpnu.ua; https://orcid.org/0000-0001-6417-3689

Федчук Ростислав Богданович, аспірант, кафедра інформаційних систем та мереж.
Email: rostyslav.b.fedchuk@lpnu.ua; https://orcid.org/0009-0002-6669-0369

Цитування за ДСТУ: Федчук Р. Б., Висоцька В. А. Математична модель ідентифікації помилок в текстах україномовного контенту. *Український журнал інформаційних технологій*. 2025, т. 7, № 2. С. 120–131.

Citation APA: Fedchuk, R. B., & Vysotska, V. A. (2025). Mathematical model of errors identification in texts of Ukrainian content. *Ukrainian Journal of Information Technology, 7*(2), 120–131. https://doi.org/10.23939/ujit2025.02.120