ОБРОБКА ПРИРОДНОЇ МОВИ УКРАЇНСЬКОЮ: ВИКЛИКИ ТА ПЕРСПЕКТИВИ ВИКОРИСТАННЯ ШТУЧНОГО ІНТЕЛЕКТУ В ОСВІТІ

NATURAL LANGUAGE PROCESSING IN UKRAINIAN: CHALLENGES AND PROSPECTS FOR THE USE OF ARTIFICIAL INTELLIGENCE IN EDUCATION

Стаття присвячена дослідженню проблем, пов'язаних із використанням технологій обробки природної мови (NLP) для аналізу та генерації навчальних матеріалів українською мовою. Автори акцентують увагу на труднощах, які виникають через обмежені ресурси української мови, зокрема недостатню кількість корпусів текстів для тренування моделей штучного інтелекту. У статті розглядаються основні причини низької якості результатів, отриманих від NLP-моделей, зокрема нерелевантні навчальні дані, неправильна токенізація, відсутність аналізу контексту та логічних зв'язків у тексті. Дослідження включає порівняння роботи мовних моделей OpenAI та BERT, зокрема їх точність, контекстуальність та адаптивність до української мови. Автори пропонують використання двонаправленого аналізу контексту, який застосовується в моделі BERT, для покращення розуміння тексту та генерації тестів. Експериментальна частина роботи демонструє, що налаштування токенізації, фільтрація стоп-слів та використання алгоритмів self-attention значно підвищують якість роботи моделей. Стаття підкреслює необхідність розробки спеціалізованих моделей, адаптованих до особливостей української мови, а також збільшення обсягів навчальних даних для професійних сфер. Висновки дослідження вказують на перспективність використання NLP у освіті, але за умови подальшого вдосконалення технологій та їх адаптації до мовних реалій. Дане дослідження може бути використано для подальшої адаптації мовних моделей для розробки тестових завдань.

Ключові слова: *обробка природньої мови, self-attention, векторні представлення, токенізація, BERT, штучний інтелект.*

УДК 004.662 (20.23.17)

DOI: https://doi.org/10.32782/dees.16-26

Пацай Б.Д.
к.ф.-м.н., доцент,
доцент кафедри статистики, інформаційно-аналітичних систем і демографії,
Київський національний університет імені Тараса Шевченка
Нечипорук І.М.
Ірпінський ліцей інноваційних технологій
Ковтун А.О.
Ірпінський ліцей інноваційних технологій

Patsai Bohdan
Taras Shevchenko National University of Kyiv
Nechyporuk Ivan
Irpin Lyceum of Innovative Technologies
Kovtun Anna
Irpin Lyceum of Innovative Technologies

TThe article is devoted to the study of problems associated with the use of natural language processing (NLP) technologies for analyzing and generating educational materials in the Ukrainian language. The purpose of the study is to analyze the results of test generation based on the proposed content and to identify possible causes of incorrect behavior in NLP models that process educational materials in the Ukrainian language. The study employs token filtering methods using self-attention algorithms. The BLEU score was used to evaluate the results obtained with BERT. The authors focus on the challenges arising from the limited resources available for the Ukrainian language, particularly the insufficient number of text corpora for training artificial intelligence models. The article examines the main reasons for the low quality of results produced by NLP models, including irrelevant training data, incorrect tokenization, a lack of contextual analysis, and weak logical connections in the text. The study includes a comparison of the performance of the OpenAI and BERT language models, focusing on their accuracy, contextual understanding, and adaptability to the Ukrainian language. The authors propose using bidirectional context analysis, as implemented in the BERT model, to improve text comprehension and test generation. The experimental part of the study demonstrates that adjusting tokenization settings, applying stop-word filtering, and using self-attention algorithms significantly improve model quality. The article emphasizes the need to develop specialized models adapted to the peculiarities of the Ukrainian language and to increase the volume of training data for professional domains. Based on the analysis of different token filtering methods, the study concludes that tokenization processes should be configured individually for each task, as this significantly affects model performance. The conclusions highlight the potential of NLP in education, provided there is further technological improvement and adaptation to linguistic realities. This study may serve as a foundation for the further adaptation of language models for developing test tasks.

Key words: *natural language processing, self-attention, vector representations, tokenization, BERT, artificial intelligence.*

Постановка проблеми. Стрімкий розвиток цифрової економіки вимагає від кожного максимального ефективного використання всіх наявних засобів для аналізу та прийняттю управлінських рішень.

Особливо це стосується використання штучного інтелекту. Одним із найпоширеніших напрямів використання штучного інтелекту в сучасному суспільстві – це обробка природної мови (NLP). Завдяки цій технології люди можуть щодня взаємодіяти зі штучним інтелектом, давати йому завдання та отримувати відповіді у текстовому вигляді. Хоча ця сфера, безумовно, є найбільш широко використовуваною завдяки своїй доступності у вигляді чат-ботів, які отримують запити людською мовою, якість одержаних результатів інколи є невисокою. В першу чергу, це стосується української мови, яка відноситься до мов з обмеженими ресурсами (low-resource language) [1, с. 28] для яких характерною є наявність мінімальних наборів даних для різних підзадач обробки природної мови, інструментів та людських ресурсів.

У порівнянні з англійською, китайською чи іспанською, українська має набагато менше корпусів текстів, які доступні для тренування моделей штучного інтелекту. Але, навіть для англійської

мови завдання пошуку змісту у наукових та технічних документах буде малоресурсним через обмежену кількість розмічених даних. Крім цього, складність розвитку NLP в українському сегменті полягає в тому, що значна частина даних походить із англійської мови і вони не адаптовані під особливості мови. Навіть найефективніші алгоритми працюють дуже погано, якщо їх не належним чином навчено або якщо змінюються контексти та домени [3, с. 9]. Такі алгоритми обмежені тим, що можуть обробляти лише ту інформацію, яку вони «бачать». Отже, однією з найбільших проблем, з якими стикаються мовні моделі є, в першу чергу, нестача даних для навчання.

Окремої уваги потребує складність самої мови. Наявність префіксів, суфіксів, наголосів та поширеність складнопідрядних речень. Але даний сегмент розвивається. Зокрема у 2024 році було презентовано Eval-UA-tion 1.0 набір нових україномовних наборів даних, спрямованих на оцінку ефективності мовних моделей спеціалізованих орієнтирів (benchmark) [2, с. 110]. Але останнім часом зростає інтерес до розвитку NLP моделей для обробки контексту українською мовою [4, с. 2].

Після зростання попиту до дистанційного навчання технічна галузь почала розвиватися, створивши у освітян нові потреби та надавши нові можливості. Освітній процес стикається з викликами, що пов'язані з великим обсягом інформації, яка потребує якісного попереднього аналізу. Для виконання цих завдань можуть бути застосовані сучасні технології та можливості штучного інтелекту, що оптимізують навчання.

Вже зараз за допомогою штучного інтелекту викладачі створюють тести для перевірки та оцінювання знань, надають додатковий матеріал для вивчення, формують тематичні плани тощо. Забезпечення правильної та якісної роботи NLP є актуальним та критично важливим завданням. Це пов'язано з тим, що коректність тестів безпосередньо впливає на якість навчання здобувачів освіти і справедливість їх оцінювання. У сферах діяльності де навчальні матеріали мають широке відображення у англомовному середовищі якість таких даних суттєво вища, в першу чергу це стосується інформаційних технологій.

Аналіз останніх досліджень і публікацій. Людська мова є різноманітною та непостійною внаслідок природного шляху формування. Нашій мові властиві символічність, та найголовніше, полісемантичність, тобто залежність значення слів від контексту. Вони сприймаються людиною завдяки глибокому розумінню мови [5, с. 75]. Проте, це є проблематичним для розуміння природної української мови комп'ютером, тому для її інтерпретації використовуються відповідні алгоритми [6, с. 320].

Формат, що буде зрозумілим для комп'ютерних систем передбачає представлення будь-якої інформації у вигляді числових значень. У роботі канадського науковця Йошуа Бенжіо, представлено feedforward neural network (нейронні мережі прямого зв'язку), які могли представляти слова як розподілені представлення, тобто як числові значення у багатовимірному просторі, які відображають інформацію про слово та його відношення до інших слів [7, с. 250]. Із розвитком вищеописаних нейронних мереж Томаш Міколов запропонував підхід до використання Word Embedding, що називався Word2Vec [8, с. 3]. Word embedding – метод обробки людської мови, де слова представляються як вектори із певними координатами у скінченному багатовимірному просторі. Слова зі схожими значеннями розташовані ближче одне до одного, а напрям векторів вказує на ступінь подібності.

Цей метод обробки базується на декількох концептах, але основним є Distributional Hypothesis (гіпотеза про розподіл значень). Суть гіпотези полягає в тому, що слова зі схожими значеннями зазвичай зустрічаються у схожих контекстах.

Важливо розрізняти, що Word2Vec – один із підходів до використання техніки Word Embedding. Серед інших є підходи викладені у роботі Педро Родріеса GloVe (Global Vectors for Word Representation), CBOW (Continuous Bag of Words) та інші, які можуть суттєво відрізнятись у підходах до навчання та визначення контексту [9, с. 105].

Очевидно, що роль NLP у всіх сферах життєдіяльності людини зростає. Ефективність діяльності будь якої компанії все частіше можна оцінити наскільки вона інтегрувала у власний бізнес процеси можливості штучного інтелекту. Дані процеси, в тому числі, стосується і освіти. Все частіше виникають думки про можливість інтеграції ШІ у навчальний процес, а в подальшому і побудова навчального процесу на основі технологій штучного інтелекту.

Постановка завдання. Метою дослідження є аналіз результатів генерації тестів на основі запропонованого контенту та виявлення можливих причин некоректної поведінки NLP-моделей, що опрацьовують навчальні матеріали українською мовою.

Виклад основного матеріалу дослідження. Для цього вирішення даної задачі було розглянуто загальні принципи роботи моделей NLP. Здійснено порівняння між мовними моделями OpenAI та BERT, а саме їх точність, контекстуальність та адаптивність до української мови.

Для представлення слів як числових значень дані мають пройти декілька етапів обробки, які зазвичай називають pipeline: збір даних, токенізація, стемінг/лематизація, алгоритми word embedding, алгоритми self-attention, аналіз даних відповідно до завдання [10, с. 3].

Токенізація зазвичай проходить у два етапи: сегментація тексту на речення та поділ речень

на слова. Сегментами (токенами), на які поділено текст є не лише змістові елементи. Токенами стають і розділові знаки, і сполучники, і прийменники, тощо.

З-поміж отриманих токенів відбираються виключно ті, що мають семантичну цінність для встановлення контексту (essential words). Такими токенами зазвичай *є іменники, дієслова, прикметники, прислівники*. Натомість токени, що не несуть цінності для встановлення контексту (non-essential tokens) здебільшого не враховуються при аналізі. До таких токенів часто належать *пунктуація, прийменники, сполучники, вигуки,* тощо. Випадком коли ці токени беруться до уваги є стилістичний аналіз тексту. Проте, у випадку із генерацією тестів наведені приклади non-essential words не враховуються. Важливо зазначити, що ці токени (також називаються stop-words) мають визначатись окремо, враховуючи вимоги завдання. Бібліотеки, за допомогою яких проводиться токенізація мають власний набір стоп-слів, проте для специфічних завдань вони мають визначатись власноруч [11, с. 2].

Протягом стемінгу із слів прибираються суфікси, префікси, постфікси та інші словотворчі частини, лишаючи лише початкову форму слова.

При використанні лематизації кінцевий продукт (лема) завжди є справжнім словом, але початковою його формою. Таким чином, лематизація є більш витратним, але більш точним процесом обробки.

З метою дослідження якості роботи штучного інтелекту (зокрема мовних моделей) було створено тестові завдання за темами з параграфів підручників. Для цього було використано інтернет ресурс https://www.quizrise.com/. Для надання посилань на джерела із додатковими матеріалами було використано https://docs.tavily.com/. Тестування було створене за предметами: фізика, біологія, історія України, зарубіжна література. Аби уникнути упередженості в проведенні опитування спеціальні форми з оцінкою роботи ШІ була сформована експертна група до якої входили вчителі-предметники з різних навчальних закладів. Наявні 13 питань були умовно розділені на дві групи: питання, що оцінюють якість обробки вхідного тексту, і питання, що оцінюють якість матеріалів, які надаються штучним інтелектом.

Оцінка результатів, що була надана освітянами виявилась незадовільною. Виходячи з цього, можна стверджувати, що мовні моделі (особливо ті, що працюють з українськими даними) потребують подальшого дослідження та виявлення причин неточностей, які виникають під час використання.

У результаті опитування середньою оцінкою першої групи запитань була 6.1 на шкалі від 1 до 10. Середня оцінка другої групи запитань – 5.1 на шкалі від 1 до 10 (рис. 1). Відповідаючи на запитання типу «так» чи «ні» 67 % вчителів вказували, що параграф, наданий для генерації тесту не був повністю опитаним та/або серед питань не було критично важливих, які відображали б знання матеріалу параграфа. У частині питань із відкритою відповіддю викладачі були незадоволені змістом тесту, зокрема, відсутністю перевірки розуміння учнями причинно-наслідкових зв'язків, недостатньою увагою до важливих речей і занадто великою кількістю питань про менш важливі речі. Таким чином, існування проблем у використанні штучного інтелекту в сфері освіти підтверджується результатами опитування.

До функцій, які виконує NLP належать синтаксичний аналіз та аналіз настроїв тексту,

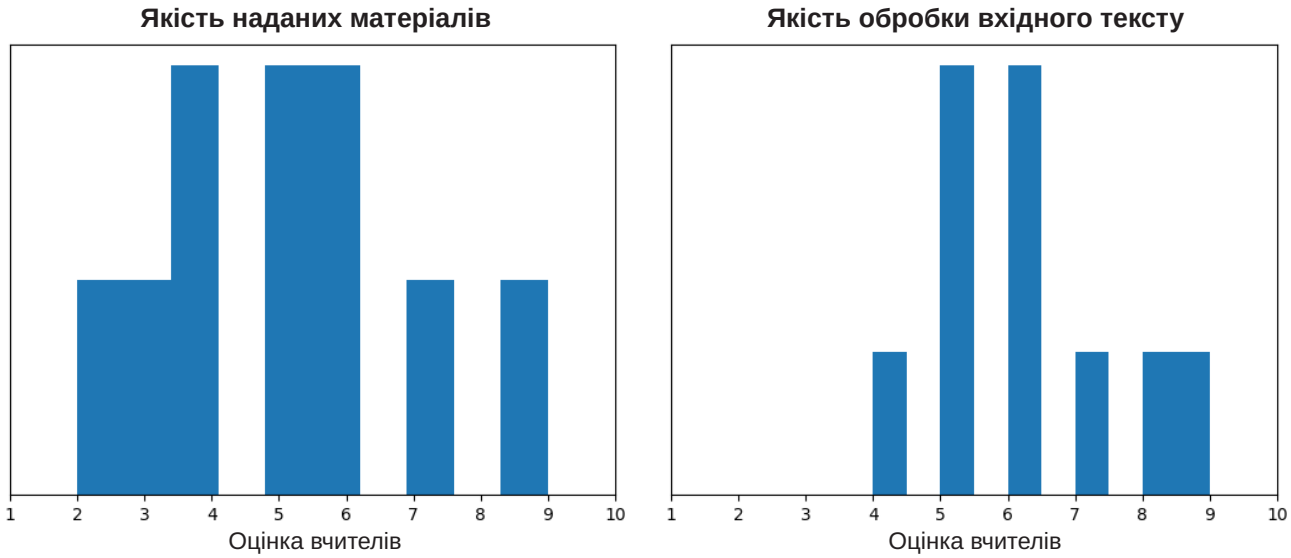

Рис. 1. Результати опитування викладачів

Джерело: сформовано автором

семантичний аналіз (вивчення зв'язків слів та контексту), NER (Name Entity Recognition, вилучення ключових елементів, таких як імена, адреси, номери телефонів) та багато інших.

Напрями, які охоплює NLP можна умовно розділити на дві групи: завдання обробки і завдання генерації людської мови. Метою NLP у завданні, що пов'язане із обробкою та розумінням мови є трансформація природної мови у формат, що буде зрозумілим для комп'ютерних систем. Ці моделі формують передбачення, проводять семантичний аналіз текстів або шукають інформацію, відповідно до визначеної моделлю теми. У даній роботі, модель в першу чергу проводить тематичне моделювання (topic modeling) тобто аналізує тему тексту, який вилучено із параграфа. Після цього, модель генерує тест, беручи за основу отриману тему, проте це вже є завданням генеративного NLP. У дослідженні основний акцент зроблений на аналізі мови, оскільки саме на цьому етапі, ймовірно, виникають неточності.

На рис. 2 зображено приклад роботи алгоритмів токенізації та стемінгу. Вони застосовуються до речення «Маленька дівчинка принесла мені величезний подарунок, який здивував усіх». У процесі токенізації речення розділяється на слова та розділові знаки. Після цього отримані токени проходять стемінг. Наприклад, у слові «принесла» виділено корінь слова (основну частину) «-несл-« та префікс «при-». У даному випадку префікс несе значення наближення дії. Тому він виділяється як окрема змістовна частина, оскільки модель має розуміти його значення.

Протягом опитування експертами було зазначено, що однією з основних причин непридатності використання ШІ для генерації тестів є неналежна увага логічним зв'язкам та висновкам. Інтерпретація штучним інтелектом таких зв'язків зумовлене їх аналізом контексту. Для виконання цієї функції є декілька підходів, робота яких буде розглянута на прикладі мовних моделей BERT (Bidirectional Encoder Representations from Transformers) та GPT4 (Generative Pre-trained Transformer 4).

Розуміння людьми логічних послідовностей та взаємозв'язків між словами в NLP реалізовується завдяки техніці передбачень. Аналізуючи контекст, навчена мовна модель передбачає тему тексту, наступне слово, фразу, висновки, тощо. Отримавши вектори як вхідні дані, за допомогою яких було представлено токени, штучний інтелект передбачає наступне слово.

Під час генерації тестів, найважливішими етапами є обробка вхідного тексту (параграфа) та генерація запитань. При аналізі даних, що потрапили у модель критично необхідним є виявлення логічних взаємозв'язків між словами, що і забезпечує розуміння штучним інтелектом основної думки тексту та відповідної генерації запитань.

У мовних моделях, які виконують завдання, що потребують глибокого аналізу тексту окрім вищеописаних етапів додається ще один: *self-attention*. *Self-attention* – технологія обробки тексту, що передбачає роботу із контекстом. Після надання токенам векторних значень мовні моделі, які потребують розуміння основної думки тексту застосовують цей алгоритм. Токенам привласнюються ще три значення: *query, key, value*.

Query (Q) – компонент, значення якого описує запит, який здійснює токен до інших токенів. *Key (K)* – вказує на характеристики токена за якими визначається його відповідність іншим токенам. *Value (V)* – визначає фактичну інформацію, яка може бути отримана із токена відповідно до контексту.

Після отримання значень *Query* та *Key* обчислюється ваговий коефіцієнт токена. Це скалярний добуток величин *Query* та *Key*, що показує наскільки токен впливає на контекст всього тексту.

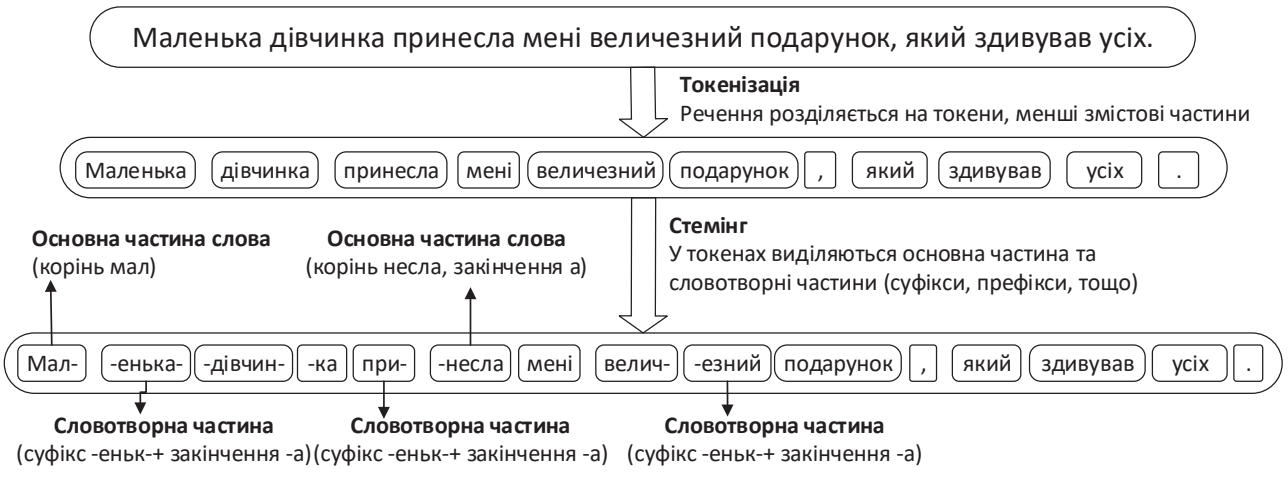

Рис. 2. Спрощена схема токенізації та стемінгу

Джерело: сформовано автором

Визначений ваговий коефіцієнт застосовується до компонента *Value*. У результаті кожен токен має нове представлення, що відповідає за його контекстуальний зв'язок. Таким чином, за допомогою алгоритму self-attention визначаються логічні послідовності тексту [12, с. 4].

$$Attention(Q,K,V) = \frac{softmax(QK^T)}{\sqrt{d_k}}V$$

На рис. 3 зображено приклад того, як може виглядати матриця ваг для частини речення «Маленька дівчинка принесла мені величезний подарунок». Зображено токени, які були виділені внаслідок роботи алгоритмів стемінгу. Кожному з них привласнено близько 100–300 числових значень, перетворюючи токени у вектори. Пізніше, кожен вектор множиться на матрицю ваг, яка виділяє значення *Key* (горизонтально) та *Value* (вертикально). Синіми кругами зображений скалярний добудок *Key* та *Value*. Чим яскравіший колір, тим більший добуток та більший зв'язок між токенами. На схемі можна побачити тенденцію, що словотворчі частини (суфікси та префікси) здебільшого мають менший зв'язок із іншими токенами. Так відбувається тому, що вони мають значний вплив лише на ті слова, до яких належать.

Підсумовуючи проблеми, які були описані та технології, що відповідають за їх виконання можна припустити, що причиною низької якості одержаних результатів є:

1. Нерелевантні для освітніх цілей навчальні дані.
2. Неправильно налаштовані для аналізу освітніх матеріалів стоп-слова, внаслідок чого можуть пропускатись важливі речі.
3. Використання менш дієвого способу попередньої обробки токенів (стемінг або лематизація).
4. Недостатня увага аналізу контексту та логічних взаємозв'язків.

Для вирішення вищеописаних проблем доречним буде пошук мовних моделей, що задовольнятимуть умови, які необхідні для виконання освітніх завдань. До того ж, обрані моделі мають проходити процес донавчання та тонкого налаштування для конкретних задач. Для порівняння із моделлю GPT4, на основі якої працює програма зі створення тестів обрана мовна модель BERT (Bidirectional Encoder Representations from Transformers).

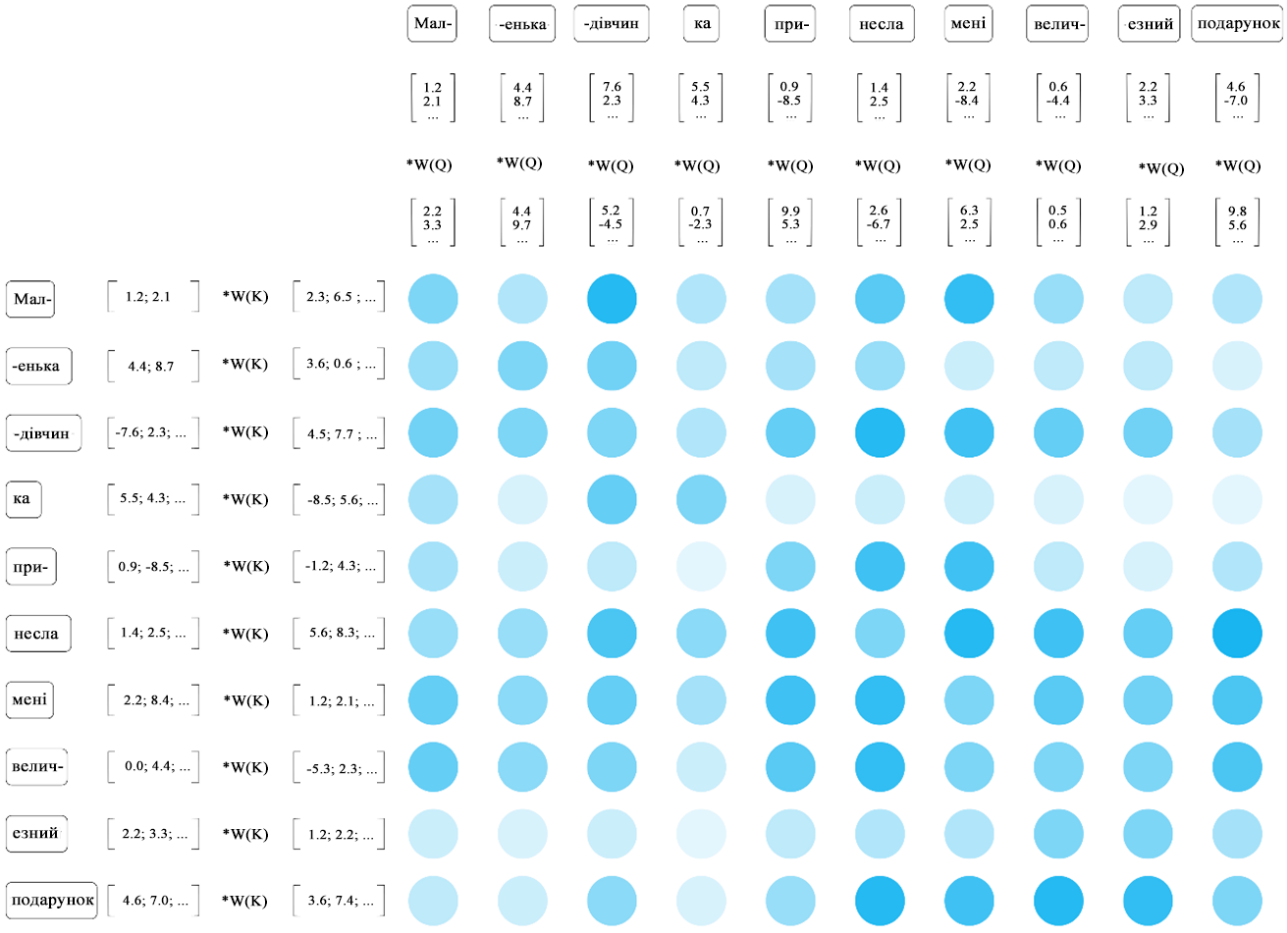

Рис. 3. Матриця ваг, спрощена схема

Джерело: сформовано автором

Основна задача, яку виконує модель GPT4 полягає у генерації зв'язного тексту. Це дозволяє їй уникнути технік та обчислень які потребують великих витрат ресурсів. Модель BERT представлена компанією Google та лягла в основу чат-боту Google Bard (нині – Gemini). BERT орієнтована на більш глибоке розуміння контекстів, що робить її придатною для використання у завданнях, де точність та надійність є важливими. Проте, витрати ресурсів та часу при такому підході зростатимуть [12, с. 4].

У моделей GPT мовні репрезентації, що були задані при навчанні моделі піддаються коригуванню та додатковому навчанню. Проте параметри, що коригуються покладаються на значення токена з урахуванням *лише попередньої частини тексту, що містить зміст*. Тобто, донавчаючи модель беруться до уваги лише ті токени, що знаходяться по один бік від об'єкту аналізу. Пропозиція BERT полягає у застосуванні підходу тонкого налаштування із двонаправленим *аналізом контексту*. Тобто враховується контекст по обидва боки від токена. Це дозволяє більш точно аналізувати логічні послідовності тексту завдяки урахуванню більшої кількості зв'язків, що мають токени один із одним [12, с. 4].

Моделям, які використовують бідирекційний підхід властивий недолік через те, що модель «бачить» усі слова, вона не передбачає значення наступного, а вже його знає. Це створює проблему під час навчання моделі. Розробники BERT знайшли вирішення проблеми. Протягом того, як модель навчається передбачати слово за контекстом навколо нього у реченні маскується 15 % токенів. Завдання моделі протягом навчання – передбачити, враховуючи контекст праворуч та ліворуч від слова, який токен був замаскований. Таким чином, під час навчання модель вчиться визначати та передбачити важливість токена у контексті та його взаємозв'язок з іншими. Моделі, що працюють за таким принципом були названі Masked Language Model (MLM) [13, с. 50].

У ході роботи для адаптації моделі BERT було написано код, основною метою якого була виділення найбільш важливих слів у тексті, використовуючи алгоритм self-attention, що був згаданий раніше. До того ж, було проведено експеримент: двом однаковим алгоритмам були дані BERT-checkpoints та безпосередньо модель BERT Multilingual для порівняння їх роботи у цьому конкретному завданні. Моделі отримали два тексти про погоду, що містили багато ключових слів із різних частин мови. Тексти були дані англійською та українською для порівняння якості роботи цих моделей із різними мовами. Результат оцінювався за допомогою BLEU score (bilingual evaluation understudy), що порівнював очікуваний результат із результатом програми. (0 – очікуваний результат жодним чином не співпадає із фактичним, 1 – очікуваний результат повністю ідентичний із фактичним).

Модель BERT Checkpoint (dbmdz/bert-large-cased-finetuned-conll03-english), яка класифікує токени за їх важливістю у тексті змогла передбачити найголовніші слова у тексті англійською мовою із правильністю у 33 % (0.3278), що говорить про незадовільний результат роботи моделі. Причиною такої поведінки є відсутність адаптації моделі під цю конкретну задачу та брак навчальних даних. Однак коли було прийнято рішення власноруч відфільтрувати стоп-токени якість моделі підвищилась на 14 %. Список відфільтрованих токенів не мав у собі займенників, прийменників, прислівників та артиклів. Відповідність цього списку до списку очікуваних токенів становить 47 % (0.4685).

Проте, оцінка роботи даної моделі на тексті українською мовою не підлягає оцінюванню, оскільки результати BLEU Score виявились близько 0 %. Причиною цього є те, що моделі не навчені на українських даних та, відповідно, не можуть виділяти токени та оцінювати їх значимість. Моделей BERT, які б пройшли процес тонкого налаштування для задачі класифікації токенів за їх важливістю або сумаризації тексту та були натреновані на україномовних даних на цей момент обмаль.

Результат цього експерименту свідчить про правильність зазначених раніше припущень щодо причин некоректної роботи моделі, а саме:

1. Неправильний підхід до налаштування токенізації та виділення стоп-слів.

2. Відсутність належного навчання на релевантних даних.

Для демонстрації того, як працюють алгоритми self-attention та як вони можуть бути застосовані у повсякденному житті було написано код на Python. Алгоритму дається задача з фізики у якій він має визначити основні слова. Потім код перевіряє, чи належать слова, які він виділив як основні, до завчасно визначених словників. Ці словники містять розділ фізики та слова, які найчастіше зустрічаються у задачах цієї теми. Знайшовши як мінімум 3 ключових слова у тексті, які збігаються із певним словником, він визначає розділ фізики, до якого належить ця задача.

Зміст алгоритму у заміні значення avg_attention так, щоб у нього не враховувались перший та останні токени [CLS] та [SEP]. Це зроблено оскільки вони мають великий зв'язок з іншими токенами моделі, що заважає сприйняттю матриці ваги. Однак, виділення цих токенів як дуже важивих є нормальною поведінкою моделі.

На рис. 4 можна побачити, що значну вагу мають перший токен та останній. Оскільки було налаштовано фільтрацію токенів таких як [CLS] та [SEP], увага змістилась на перший та останній токен, що обробляються моделлю. Першим токеном

у тексті є [CLS], а слідом за ним йде артикль «A», що також визначається як стоп-токен. Тому вага токена [CLS] змістилась на перший токен, який обробляється моделлю, а саме токен «double».

Обробка токенів з першого по передостанній, забезпечило більш точний результат роботи моделі який представлено на рис. 5.

У програмі було використано списки токенів, що відповідають певній темі у фізиці. Також здійснювалась перевірка скільки токенів збігаються із токенами у завчасно визначених словниках. У результаті була побудована теплова мапа для зображення матриці ваг.

Висновки. Підсумовуючи можна зробити висновки про можливість адаптації та налаштувань мовних моделей під конкретне завдання та двонапрямленого аналізу контексту робить такі моделі придатними до глибокого розуміння

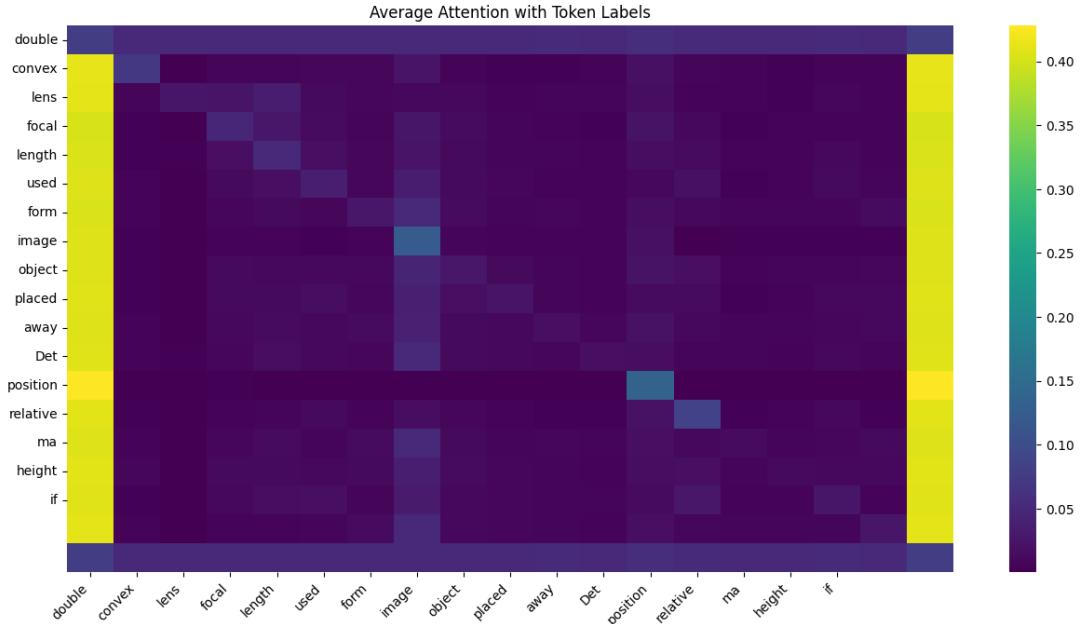

Рис. 4. Матриця уваги із урахуванням токенів CLS та SEP

Джерело: сформовано автором

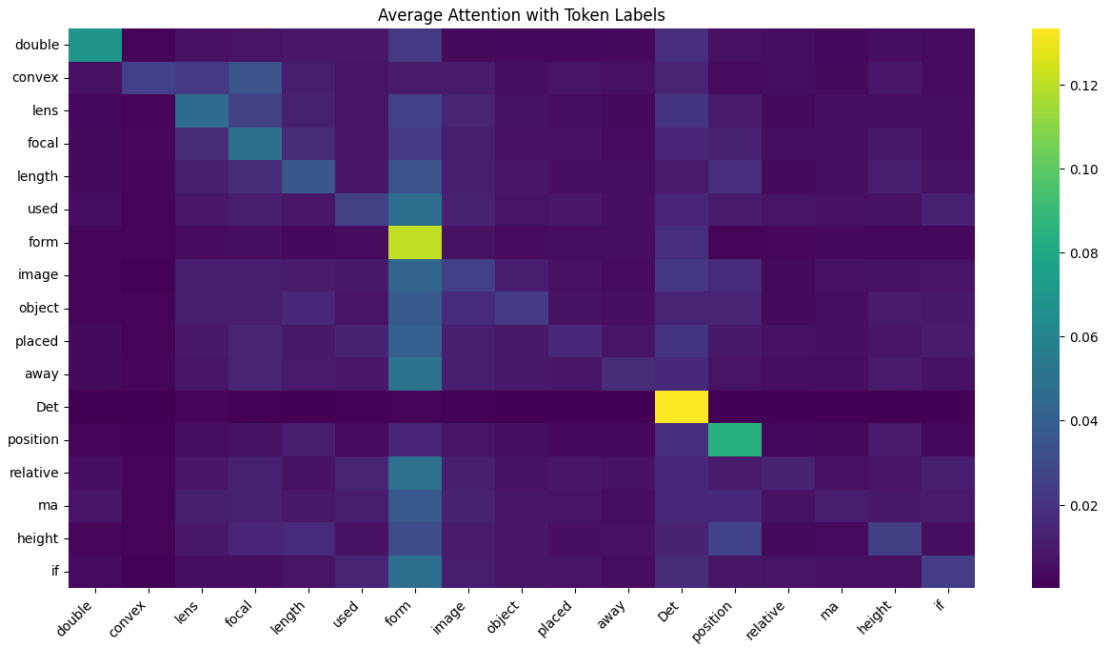

Рис. 5. Матриця уваги без урахування токенів CLS та SEP

Джерело: сформовано автором

людської мови. Базові налаштування бідирекційних моделей-енкодерів на кшталт BERT дозволяють проводити класифікацію токенів за їх важливістю, використовуючи алгоритми self-attention.

У ході аналізу різних способів фільрації токенів можна зробити висновок, що процеси токенізації мають бути налаштовані окремо для кожного завдання, адже це сильно впливає на якість роботи моделі.

Також, варто зазначити причини низької якості одержаних результатів для української мови:

1. Тренування на нерелевантних даних.
2. Неправильність або відсутність окремого налаштування роботи моделі під конкретні завдання.
3. Використання менш дієвого способу попередньої обробки токенів (стемінг або лематизація).
4. Відсутність правильного аналізу контексту та зв'язків у тексті.

Загалом для успішного використання NLP для аналізу контенту українською мовою необхідно збільшення даних для навчання, в першу чергу це стосується професійної сфери, формування тематичних словників та розробка спеціальних моделей які б враховували особливості української мови.

БІБЛІОГРАФІЧНИЙ СПИСОК:

1. Berment V. Méthodes pour informatiser les langues et les groupes de langues «peu dotées» : Doctoral dissertation, Université Joseph-Fourier-Grenoble I / Université Joseph-Fourier-Grenoble I, 2004.
2. Hamotskyi S., Levbarg A. I., Hänig C. Eval-UAtion 1.0: Benchmark for Evaluating Ukrainian (Large) Language Models: Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP)@ LREC-COLING 2024. 2024. May. P. 109–119.
3. Cambria E., White B. Jumping NLP Curves: A Review of Natural Language Processing Research. *IEEE Computational Intelligence Magazine*. 2014. Vol. 9, No. 2. P. 48–57. DOI: https://doi.org/10.1109/MCI.2014.2307227.
4. Vysotska V. Computer linguistic system modelling for Ukrainian language processing. *CEUR Workshop Proceedings*. 2024. Vol. 3722. P. 288–342.
5. Vysotska V., Pukach P., Lytvyn V., Uhryn D., Ushenko Y., Hu Z. Intelligent analysis of Ukrainian-language tweets for public opinion research based on NLP methods and machine learning technology. *International Journal of Modern Education and Computer Science (IJMECS)*. 2023. Vol. 15, No. 3. P. 70–93. DOI: https://doi.org/10.5815/ijmecs.2023.03.06.
6. Mashtalir S. V., Nikolenko O. V. Data preprocessing and tokenization techniques for technical Ukrainian texts. *Applied Aspects of Information Technology*. 2023. T. 6, № 3. C. 318–326. DOI: https://doi.org/10.15276/aait.06.2023.22.
7. Glorot X., Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. 2010. March. P. 249–256.
8. Norouzi M., Mikolov T., Bengio S., Singer Y., Shlens J., Frome A., Dean J. Zero-shot learning by convex combination of semantic embeddings. arXiv preprint arXiv:1312.5650. 2013. DOI: https://doi.org/10.48550/arXiv.1312.5650.
9. Rodriguez P. L., Spirling A. Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics*. 2022. Vol. 84, No. 1. P. 101–115. DOI: https://doi.org/10.1086/715162.
10. Tenney I. BERT rediscovers the classical NLP pipeline. arXiv preprint arXiv:1905.05950. 2019. DOI: https://doi.org/10.48550/arXiv.1905.05950.
11. Elov B. B., Khamroeva S. M., Xusainova Z. Y. The pipeline processing in NLP: E3S Web of Conferences. 2023. DOI: https://doi.org/10.1051/e3sconf/202341303011.
12. Im J., Cho S. Distance-based self-attention network for natural language inference. arXiv preprint arXiv:1712.02047. 2017. DOI: https://doi.org/10.48550/arXiv.1712.02047.
13. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Polosukhin I. Attention is all you need: NIPS. 2017. December.