

ВИКОРИСТАННЯ ШІ ДЛЯ ВИЗНАЧЕННЯ МЕТАТЕГІВ У НАВЧАЛЬНИХ І НАУКОВИХ ДОКУМЕНТАХ

Зростання обсягу даних, які генерують освітні установи та наукові організації, створює нові виклики для їхньої обробки, класифікації та управління. Однією з важливих складових управління великих наборів інформаційних документів є метатеги, що виконують функцію маркерів для структуризації даних. Їхня наявність дозволяє суттєво зменшити витрати часу на пошук потрібних матеріалів, забезпечити якісну навігацію в архівах, а також створити умови для аналітики даних. Проте, у багатьох освітніх та наукових архівах документи або зовсім не мають метатегів, або вони заповнені частково. У більшості випадків додавання метаданих виконується вручну, що займає багато часу та може бути джерелом помилок [1].

Штучний інтелект (ШІ) відкриває нові можливості для автоматизації цього процесу. В цих матеріалах розглядається використання генеративного моделі штучного інтелекту GPT-4, яка демонструє високу точність у текстових завданнях, зокрема у створенні метаданих для наукових і навчальних документів. Також оцінюється можливість використання цієї моделі для файлів розміщених на платформи Microsoft 365 Education, зокрема Teams та SharePoint Online. Ці хмарні сервіси широко використовуються для організації навчальних процесів в закладах освіти по всьому світу і зокрема в Україні, це набуло більшої актуальності в світлі останніх подій [2, с. 99].

Значна кількість освітніх та наукових організацій, зокрема університети, дослідницькі центри, бібліотеки та електронні архіви, стикаються з такими проблемами як незручність пошуку, ручне впорядкування, втрата контексту. Відсутність стандартизованих метатегів ускладнює пошук релевантних матеріалів. Процеси додавання метаданих часто виконуються вручну, що потребує значні людські ресурси. Неповнота або некоректність метаданих зменшує доступність важливих документів, ускладнюючи їхню подальшу обробку та використання.

У відповідь на ці виклики GPT-4 надає функціонал що дозволяє розробити автоматизовані рішення, які здатні значно спростити роботу з великими обсягами документів. Це особливо актуально

в умовах переходу на хмарні платформи, де потрібно оперативно впорядковувати документи.

Автоматизація визначення метатегів із використанням GPT-4 включає кілька ключових етапів:

- збір даних – формування вибірки документів із різними форматами: PDF, Word, презентації PowerPoint;
- попередня класифікація – визначення типів файлів, що обробляються: навчальні програми, дослідницькі статті, лекційні матеріали тощо;
- навчання моделі (fine-tune) – використання спеціально підготовлених даних для навчання GPT-4 на розпізнаванні типових структур документів та їхньої класифікації;
- інтеграція в інформаційні системи закладу – реалізація автоматизованого процесу додавання метатегів у хмарному середовищі, наприклад, SharePoint Online.

Підготовка fine-tune моделі GPT-4 є важливим етапом для забезпечення високої точності класифікації підкріпленої додатковим навчанням базової моделі на власних наборах даних [3].

Було проведено навчання на наборах даних, що містили:

- розмічені файли – документи з чітко визначеними типами, структурою та ключовими словами;
- підготовлені приклади класифікації – файли, які демонструють характерні ознаки певних категорій, наприклад, навчальні програми мають заголовки модулів, а наукові статті містять розділи «Анотація» та «Методика»;
- проблемні кейси – документи зі складними або нестандартними структурами для моделювання рідкісних сценаріїв.

Модель навчалася визначати ключові атрибути файлів, такі як структура тексту, наявність ключових розділів, характерні формулювання та ключові слова. Після навчання модель була протестована на контрольній вибірці документів, які не входили до навчального набору. Основні критерії оцінки включали: точність визначення метатегів, коректність класифікації типу документів, релевантність метаданих змісту документа. Тестування показало, що GPT-4 здатен забезпечити точність класифікації до 85–90 %, залежно від типу файлу та його складності.

Під час інтеграції у хмарну платформу SharePoint Online була реалізована така послідовність:

- аналіз поточного стану архівів – виявлення документів без метаданих або з неповними тегами;
- обробка масиву даних GPT-4 – автоматичне додавання метатегів для кожного документа;

- верифікація результатів – перевірка метаданих бібліотекарями чи експертами;
- автоматизація процесу – створення скриптів для регулярної обробки нових документів.

Реалізація рішення продемонструвала такі результати: значне скорочення часу на впорядкування архівів (на 70 %), зменшення людських помилок у метаданих, покращення доступності документів завдяки якісному пошуку. Досвід роботи з бібліотечною системою підтвердив доцільність використання GPT-4 як інструменту для оптимізації освітньої діяльності.

Попри значний прогрес, існує кілька викликів: труднощі з обробкою нестандартних документів, етичні аспекти, інтеграція в різні системи. Тому для частини файлів GPT-4 потребує додаткового навчання на тренувальних наборах більшого обсягу. Також використання ШІ викликає питання щодо прозорості, об'єктивності та контролю, які іноді важко аргументовано пояснити і може бути пов'язано із надмірним навчанням моделі. Різні університетські платформи мають різні технічні обмеження, що потребує додаткової адаптації рішень. Тому подальші дослідження можуть бути зосереджені на підвищенні гнучкості моделей, зокрема адаптації до різних мов та спеціалізованих форматів документів.

Інтеграція GPT-4 у бібліотечні системи відкриває нові можливості для автоматизації рутинних завдань, таких як визначення метатегів. Це сприяє оптимізації роботи освітніх установ, полегшує доступ до матеріалів для викладачів і студентів та стимулює впровадження інновацій у наукову діяльність. Подальші дослідження у цьому напрямі можуть забезпечити ще більшу ефективність моделей та їх інтеграцію з інструментами аналізу великих даних, що особливо актуально в умовах розвитку цифрових освітніх екосистем.

Література:

1. Herrera-Urtiaga, A. P., Torres-Cruz, F., Mendoza-Mollocondo, C. I., Paredes-Quispe, J. R., Chambi-Mamani, E. W. Automatic extraction of metadata based on natural language processing for research documents in institutional repositories. *Smart Innovation, Systems and Technologies*. 2023. Pp. 189–197. DOI: 10.1007/978-3-031-31007-2_19
2. Hlazunova, O. H., Korolchuk, V. I., Voloshyna, T. V. The Strategy of Digital Competence Formation Using a Hybrid Learning Environment Based on Microsoft 365 Services: collective monograph. *National University of Life and Environmental Sciences of Ukraine. Košice: Vysoká škola bezpečnostného manažérstva v Košiciach*. 2022. 203 p. DOI: 10.5281/zenodo.7633313
3. Liyanage, C. R., Gokani, R., Mago, V. GPT-4 as an X data annotator: Unraveling its performance on a stance classification task. *PLoS ONE*. 2024; 19(8): e0307741. DOI: 10.1371/journal.pone.0307741