Снитюк Віталій Євгенович, д.т.н., проф.

Пономарьова Даріна Андріївна, студент

Київський національний університет імені Тараса Шевченка

м. Київ, Україна

0009-0001-7620-2050

**Розробка системи виявлення семантичних запозичень у наукових текстах з використанням трансформерів**

У сучасному академічному світі та професійному контексті важливість академічної доброчесності стає особливо актуальною. Легкість доступу до невпинно зростаючих обсягів інформаційних ресурсів через Інтернет призводить до збільшення випадків плагіату. Це підриває основи наукової етики та довіру до наукових досліджень.

Плагіат – це неетична практика привласнення письмової роботи іншого автора і представлення її як власного творіння [1]. Він може проявлятися в різних формах: більш простих, таких як пряме копіювання контенту з інших джерел з незначними змінами, або більш складних формах – семантичних, таких як зміна граматичної структури тексту, перефразування, переклад змісту роботи з іншого джерела тощо [2]. Остання категорія плагіату може включати в себе комбінацію перелічених стратегій для приховування своєї наявності.

Разом з тим, методи протидії плагіату також удосконалюються. Проте, виявлення семантичного плагіату досі залишається невирішеною проблемою, адже такий вид плагіату не завжди підлягає легкому розпізнаванню, оскільки вимагає значно глибшого аналізу тексту.

Актуальність даного дослідження випливає з необхідності розробки спеціалізованої системи виявлення семантичних запозичень у наукових текстах, що дозволить ефективно протидіяти такому виду академічної недоброчесності. Метою дослідження є підвищення ефективності процесів виявлення плагіату шляхом розробки та аналізу інтелектуальної технології розпізнавання текстових запозичень.

У науковій літературі запропоновано численні методи виявлення текстових запозичень. Основні підходи, які в них висвітлені, можна розділити на декілька основних категорій:

* методи, що базуються на вимірюванні текстової відстані,
* статистичні методи,
* методи машинного та глибинного навчання.

З аналізу сучасних досліджень у галузі виявлення семантичного плагіату було визначено, що найкращі результати можна одержати, використовуючи моделі нейромереж із глибинним навчанням, зокрема трансформери. Ці технології ефективно розв’язують задачі семантичного аналізу тексту, демонструючи високу точність у моделюванні схожості речень та виявленні плагіату.

Розглянемо сучасні моделі – трансформери для забезпечення глибокого і точного аналізу текстових даних: моделі E5-Multilingual та ColBERT.

Модель E5-Multilingual є частиною сімейства моделей E5, розроблених для глибокого розуміння тексту [3]. Дана модель базується на моделі Sentence Transformers, зокрема на `paraphrase-multilingual-mpnet-base-v2`, яка використовувалася для ефективного векторного представлення тексту різними мовами. E5-Multilingual – сучасніша і високоточна модель, що дозволяє найточніше обчислювати схожість тексту різними мовами.

Модель ColBERT (Contextualized Late Interaction over BERT) ефективно використовує контекстуальні можливості та методику пізньої взаємодії. З використанням даної методики модель розширює ідеї BERT і оптимізує обробку та ранжування текстів. Так, замість агрегування на ранньому етапі, ColBERT зберігає індивідуальні вектори для кожного вхідного токена запиту і документа аж до самого кінця процесу обробки [4]. Це означає, що модель здійснює детальне і динамічне порівняння між токенами, що дозволяє зберегти та використати більш тонку семантичну інформацію.

На основі дослідження та детального аналізу результатів виявлення плагіату окремими методами, було розроблено комбінацію технологій, спрямовану на підвищення точності та швидкості виявлення як прямого, так і семантичного плагіату. Ключова стратегія оптимальної комбінації полягає у використанні E5-Multilingual для первинної швидкої перевірки великих масивів текстів, тобто пошуку семантично схожих статей, за яким слідує застосування ColBERT для більш детального аналізу фрагментів, ідентифікованих як потенційно запозичені. Розроблена комбінація методів дозволила не лише підвищити точність виявлення семантичного плагіату до 93.4%, але й оптимізувати час обробки даних.

Тренування моделей було здійснено на штучно створеному наборі даних з симульованими випадками прямого та семантичного плагіату. Для формування бази знань системи, з якою відбувається порівняння вхідних статей на наявність плагіату, використовувались збірки матеріалів міжнародних конференцій «Інформаційні технології та впровадження» різних років, які містять наукові статті на різні теми, пов’язані з інформаційними технологіями.

Розроблено інформаційно-аналітичну систему виявлення семантичних запозичень у наукових текстах (рис. 1). Для оцінки її ефективності у виявленні семантичного плагіату підготовлено статтю зі штучно внесеними перефразуваннями тексту з інших джерел. Ці сегменти включали як загальні переформулювання, так і більш дрібні модифікації оригінальних текстів, що дозволяло оцінювати систему в широкому спектрі сценаріїв.



Рис. 1 Результат виявлення семантичного плагіату системою.

В даному випадку розроблена система виявила загальний відсоток плагіату – 18%, що збігається із попередньо закладеним відсотком штучно доданого семантичного плагіату. При цьому, системою було вірно розпізнано всі випадки семантичних запозичень, а також вірно вказано джерела плагіату.

Література

1. Закон України "Про авторське право і суміжні права" [Текст] : від 23 лютого 1994 р. № 3792-XII // Відомості Верховної Ради України. — 1994. — № 13. — Ст. 64.
2. Рижко О. Поняття, види, класифікації плагіату [Текст] / О. Рижко // Записки Львівської національної наукової бібліотеки України імені В. Стефаника. – 2016. – № 8. – С. 134-150. – Режим доступу до журналу: http://nbuv.gov.ua/UJRN/lnnbyivs\_2016\_8\_12
3. Wang L., Yang N., Huang X., Yang L., Majumder R., Wei F. Multilingual E5 Text Embeddings: A Technical Report [Електронний ресурс] // arXiv preprint arXiv:2004.02743. – 2020. – Режим доступу: https://arxiv.org/abs/2004.02743
4. Khattab O., Zaharia M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT [Електронний ресурс] // arXiv preprint arXiv:2004.12832. – 2020. – Режим доступу: <https://arxiv.org/abs/2004.12832>