Шелюжак Ярослав Сергійович, здобувач СВО «бакалавр»

Турченко Ірина Василівна, к. т. н, доцент

Західноукраїнський національний університет, м. Тернопіль

**ПРОГРАМНИЙ МОДУЛЬ КЛАСИФІКАЦІЇ ЕЛЕМЕНТІВ АНОТАЦІЙ НАУКОВИХ СТАТЕЙ**

Написання наукових статей відіграє важливу роль у процесі розвитку науки і технологій. Це основний спосіб спілкування науковців зі спільнотою, що дозволяє обмінюватися ідеями, відкриттями та досвідом. Анотація статті надає короткий огляд всього дослідження і дозволяє швидко оцінити актуальність та значущість дослідження, не читаючи всю статтю. Враховуючи величезну кількість робіт, що публікуються щодня, можливість швидко переглянути анотації для виявлення релевантних досліджень є надзвичайно цінною, а тому розробка програмних засобів, що реалізовуватимуть цю можливість, є актуальною.

Сьогодні обробка природньої мови, галузь штучного інтелекту, має великий потенціал для автоматизації аналізу та класифікації тексту. Останні досягнення в цій сфері дозволяють комп’ютерам краще розуміти та інтерпретувати людську мову.

Різні методи інтерпретації людської мови для класифікації анотацій наукових статей досліджували автори в [1].

Зазвичай анотації включають в себе речення, які відносяться до однієї з наступних категорій: ОГЛЯД, МЕТА, МЕТОД, РЕЗУЛЬТАТИ, ВИСНОВКИ, кожна з яких пояснює зміст дослідження, цілі, методи, висновки, а також наслідки [2].

На рисунку 1 наведено алгоритм роботи програмного модуля класифікації елементів анотацій наукових статей на основі штучного інтелекту. Першим кроком є ввід анотації у програмний модуль за допомогою текстового файлу, який записується у змінну. Далі анотація поділяється на речення, кожне з яких проходить через ідентифікацію та заміну спеціальних символів: таких як числа, розділові знаки, елементи обчислення та формул та інші символи, які не є частиною мови. Алгоритм ініціює змінну, яка контролює кількість ітерацій у циклі. Змінна «і» представляє поточну ітерацію, «M» відображає кількість речень (елементів) у анотації. Під час кожної ітерації «і»-те речення проходить процес токенізації. Таке речення, представлене у вигляді тензора, повністю готове до подачі на вхід моделі.



Рисунок 1 – Схема алгоритму роботи програмного модуля

Після успішного прогнозування результат зберігається у змінну та програмний модуль виконує прогнозування для решти ітерацій. Після виконання прогнозування для всіх елементів, програмний модуль зберігає результат класифікації у вигляді текстового файлу.

У процесі реалізації програмного модуля на першому етапі було проведено навчання моделі обробки природної мови для виконання класифікації елементів анотацій наукових статей. Модель навчалася на наборі даних PubMed 200k RCT, який містить близько 200 000 анотацій із випадково обраних статей, при цьому загальна кількість речень в яких становить близько 2,3 мільйона. Кожне речення в цих анотаціях позначено відповідною роллю, яку вона виконує в анотації. Набір даних структурований так, щоб відображати типовий формат анотацій, таким чином зберігаючи послідовність, яка відображає логічний потік речень у реальній літературі [3].

Запропонована архітектура моделі обробки природної мови, яка застосована в розробленому програмному модулі, базується на архітектурі BERT, та доповнена рівнями нейронних мереж для інтеграції з контекстними функціями. Модель використовує складну структуру вхідних даних для покращення її розуміння та точності класифікації.

Для реалізації програмного модуля та навчання моделі було використано мову програмування Python, бібліотеку машинного навчання Tensorflow, середовище Visual Studio Code та середовище керування версіями бібліотек Anaconda.

Тестування програмного модуля продемонструвало його потенціал для автоматизації класифікації елементів анотацій наукових статей, що робить його цінним інструментом для науковців та дослідників.

**Література:**

1. Dernoncourt F., Lee J. Y. PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts. In Proceedings of the Eighth International Joint Conference on Natural Language Processing, 2017, Volume 2: Short Papers, pp. 308–313.
2. Plaxco K. W. The art of writing science. Protein Science, 2010. Pp. 2261-2266.
3. PubMed 200k RCT. URL: <https://www.kaggle.com/datasets/matthewjansen/pubmed-200k-rtc/data>