УДК 004.89:004.93 DOI: https://doi.org/10.30837/IYF.PCEIP.2024.056

ОЦІНКА ЕФЕКТИВНОСТІ НЕЙРОМЕРЕЖЕВОЇ СИСТЕМИ ДЛЯ КАТЕГОРИЗАЦІЇ ТЕКСТОВИХ ДОКУМЕНТІВ

Рибалов О.О.
Науковий керівник – д.т.н. проф. Фесенко Т.Г.
Харківський національний університет радіоелектроніки, каф. ЕОМ,
м. Харків, Україна
тел +38(067) 451-10-97, oleksandr.rybalov@nure.ua

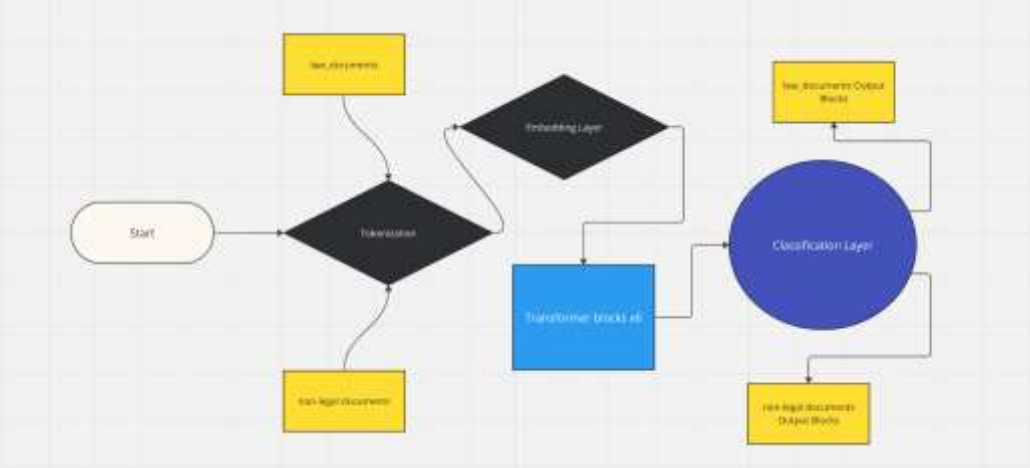
This paper presents an evaluation of the DistilBERT model's effectiveness for categorizing Ukrainian text documents. DistilBERT, a streamlined version of BERT, aims to retain the original's performance with reduced size and increased speed. This study focuses on the model's application for classifying texts into legal and non-legal categories using publicly available data, including court decisions and social media posts. The training encompassed several epochs, enhancing the model's adaptation to data peculiarities. The results, including high accuracy and precision metrics, affirm DistilBERT's efficacy in this context. This research highlights the potential of neural network systems for automating the processing and categorization of Ukrainian texts in various fields.

Відомо, що Natural Language Processing (NLP) – галузь комп'ютерних наук та штучного інтелекту, що займається розробкою методів та технологій для взаємодії між комп'ютерами та людьми через природну мову. Для обробки та категоризації природної мови використовується велика кількість різних моделей. Зокрема, командою вчених у Google запропонована модель Bidirectional Encoder Representations from Transformers (BERT). Для вирішення задач зменшення обсягу пам'яті та ресурсів розроблена скорочена версія моделі BERT – «DistilBERT» [2].

Застосування моделі «DistilBERT» дозволило розробити нейромережеву систему категоризації текстових документів українською мовою на «юридичні» та «неюридичні» (малюнок №1) [2]. Тренування і валідація системи відбувалось із використанням текстів судових рішень, розміщених на платформі Єдиного державного реєстру судових рішень (https://reyestr.court.gov.ua/).

Ефективність нейромережевої системи категоризації текстових документів оцінюється параметрами:

1) Eval loss – середня втрата (або помилка) моделі на тестових даних. Чим нижче цей показник, тим краще модель впоралася із завданням;
2) Eval accuracy – точність моделі, відсоток правильно класифікованих прикладів серед усіх тестових прикладів;
3) F1 Score – гармонічне середнє між точністю (precision) та відтворенням (recall). F1-оцінка, наближена до 0.99 вказує на високий рівень балансу між точністю та відтворенням;



Малюнок №1 – Загальна схема нейромережевої системи для категоризації текстових документів

4) Precision – відсоток правильних позитивних передбачень відносно усіх позитивних передбачень, які зробила модель.
5) Recall – відсоток правильних позитивних передбачень відносно усіх позитивних прикладів у тестовому наборі.

Аналіз продуктивності нейромережевої системи для категоризації на прикладі судових рішень дозволило отримати наступні оцінки параметрів: Eval loss – 0,004677; Eval accuracy – 0,998749; F1 Score – 0,997504; Precision – 0,995020; Recall – 1,000000. Отримані результати демонструють високий рівень точності класифікації документів та підтверджує доцільність застосування моделі DistilBERT. Подальші дослідження ефективності використання нейромережевих систем для категоризації текстів будуть пов’язані з роботою документів з іншої галузі (наприклад, з висновками про результати акредитаційної експертизи освітньої програми [3]).

Список використаних джерел:

1. Sanh V., Debut L., Chaumond J. & Wolf T. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *Cornell Universiti*. doi: https://doi.org/10.48550/arXiv.1910.01108.

2. Рибалов О.О. & Фесенко Т.Г. (2023). Дослідження засобів інтелектуального аналізу текстових документів. Збірник наукових праць XVI Міжнародної науково-практичної конференції «Академічна й університетська наука: результати та перспективи», 12–13 грудня 2023 року. Полтава: Полтавська політехніка, 330–331.

3. Fesenko T., Ruban I., Karpenko K., Fesenko G., Kovalenko A., Yakunin A. & Fesenko H. (2022). Improving of the decision-making model in the processes of external quality assurance of higher education. Eastern-European Journal of Enterprise Technologies. Vol.1(3(115)), 74–85. doi: https://doi.org/10.15587/1729-4061.2022.253351.