УДК 004.01
DOI https://doi.org/10.32782/2663-5941/2024.5.1/47

Черняк І.О.
Державний університет «Житомирська політехніка»

ДОСЛІДЖЕННЯ МОЖЛИВОСТЕЙ МОДЕЛІ DISTILBERT ДЛЯ ПОДАЛЬШОГО ВИКОРИСТАННЯ В СИСТЕМІ АВТОМАТИЗАЦІЇ ЕЛЕКТРОННОГО ДОКУМЕНТООБІГУ

У цій науковій статті досліджено потенціал використання моделі DistilBERT для автоматизації процесів електронного документообігу (ЕДО). ЕДО є невід'ємною складовою сучасного бізнес-середовища, оскільки дозволяє значно скоротити витрати часу та ресурсів на обробку, зберігання та передачу документів. Однак, великі обсяги документації та різноманітність їх форматів створюють певні виклики для ефективної автоматизації ЕДО.

DistilBERT, як одна з найсучасніших моделей обробки природної мови (NLP), пропонує ефективне рішення цих проблем. Розроблена на основі архітектури BERT (Bidirectional Encoder Representations from Transformers), DistilBERT зберігає високу точність обробки тексту, але при цьому є значно меншою та швидшою. Це робить її ідеальним інструментом для впровадження в системи ЕДО, де швидкість та ефективність є критичними факторами.

У статті розглянуто особливості архітектури DistilBERT та її переваги у порівнянні з іншими моделями NLP. Зокрема, проаналізовано здатність DistilBERT до розуміння контексту та семантики тексту, що є ключовим для успішної автоматизації ЕДО. Також розглянуто можливість тонкого налаштування моделі під конкретні завдання, що дозволяє досягти максимальної ефективності в різних сферах застосування.

Для оцінки ефективності DistilBERT на завданнях ЕДО було проведено серію експериментів. Модель було протестовано на завданнях класифікації документів, вилучення ключової інформації та зведення тексту. Результати експериментів показали високу точність та ефективність DistilBERT у всіх тестових завданнях. Зокрема, модель продемонструвала здатність точно класифікувати документи за їх типом та змістом базуючись на тестових даних, вилучати важливу інформацію, таку як дати, імена, адреси тощо, та створювати короткі та інформативні зведення великих текстів.

Отримані результати підтверджують перспективність використання DistilBERT для автоматизації ЕДО. Завдяки своїй високій точності, швидкості та здатності до розуміння контексту, DistilBERT може значно покращити ефективність процесів обробки документів, зменшити кількість помилок та скоротити витрати часу та ресурсів. Це особливо актуально для великих організацій, де обсяги документації є значними.

У статті також обговорюються потенційні напрямки подальших досліджень. Зокрема, розглядається можливість використання DistilBERT для інших завдань ЕДО, таких як автоматичне заповнення форм, перевірка документів на відповідність вимогам тощо. Також пропонуються шляхи покращення моделі, наприклад, шляхом використання додаткових даних для навчання або шляхом комбінування DistilBERT з іншими моделями NLP.

Загалом, ця наукова стаття робить достатній внесок у розвиток технологій автоматизації ЕДО. Запропоноване використання моделі DistilBERT покращує можливості для підвищення ефективності та продуктивності роботи з документами, що може мати значний позитивний вплив на різні сфери діяльності.

Ключові слова: *DistilBERT, електронний документообіг (ЕДО), обробка природної мови (NLP).*

Постановка проблеми. Автоматизація електронного документообігу є важливим напрямком розвитку сучасних інформаційних систем. Застосування методів обробки природної мови дозволяє значно пришвидшити та спростити процеси обробки, аналізу та управління електронними документами. Однією з найсучасніших моделей NLP є DistilBERT, яка поєднує високу точність з відносно низькими обчислювальними витратами.

Аналіз останніх досліджень і публікацій. Проведений аналіз останніх досліджень і публікацій свідчить про зростаючий інтерес до використання моделей NLP, зокрема DistilBERT, для автоматизації ЕДО.

DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter: Ця робота представляє модель DistilBERT, яка є меншою та швидшою версією BERT, зберігаючи при цьому високу

ефективність на різних завданнях NLP [1]. Також BERT for Patent Classification: У цій статті досліджується використання BERT для класифікації патентів, що є важливим завданням в ЕДО [2]. Leveraging BERT for Extractive Text Summarization on Lectures: Ця робота демонструє використання BERT для створення коротких витягів з лекцій, що може бути корисним для автоматичного зведення документів в ЕДО [3]. Automated invoice processing: Machine learning-based information extraction for long tail suppliers: У цій статті досліджується використання моделей NLP, включаючи BERT, для автоматичної обробки рахунків-фактур, що є важливим завданням в ЕДО [4]. Natural Language Processing for Legal Document Review: Opportunities and Challenges: Цей огляд статті розглядає можливості та виклики використання NLP, включаючи моделі на основі трансформерів, для аналізу юридичних документів, що є важливою сферою застосування ЕДО [5].

Також і відчизняні вчені мають дослідження і публікації в даній тематиці хоч і вона є не такою обширною через мовні обмеження в відкритих NPL моделях які в основному орієнтуються на англійську мову, проте можна виділити наступні дослідження як приклад можна обрати статтю «Класифікація текстових документів з використанням доповнення векторних представлень документів графовими представленнями елементів словника синонімів» Романа Шаптала та Геннадія Кисельова в якій досліджується вплив інтеграції графових представлень синонімів на класифікацію текстів у малоресурсних умовах, пропонуючи нову векторну модель та демонструючи її ефективність на прикладі класифікації петицій [6]. Також стаття "Data augmentation in text classification with multiple categories" автори якої Микола Стасюк та Богдан Павлишенко в якій автори досліджують вплив розширення даних на багатокласову класифікацію тексту за допомогою трансформаторних моделей (BERT, DistilBERT, ALBERT, XLM-RoBERTa) і оцінюють ефективність різних методів розширення для збереження мов, що знаходяться під загрозою зникнення, та вдосконалення підходів до машинного навчання [7]. Також можна виділити публікацію Притула М.М. «Налаштування моделей BERT, DistilBERT, XLM-RoBERTa та Ukr-RoBERTa для сентимент-аналізу коментарів українською мовою», це дослідження продемонструвало ефективність застосування моделей-трансформерів, зокрема BERT, DistilBERT, XLM-RoBERTa та Ukr-RoBERTa, для аналізу тональності відгуків українською мовою, при цьому XLM-RoBERTa показала найвищу точність, а Ukr-RoBERTa визнана оптимальною з огляду на швидкість навчання та загальні показники класифікації [8].

Ці дослідження та публікації підтверджують, що DistilBERT є перспективним інструментом для автоматизації ЕДО. Проте, для повного розкриття його потенціалу необхідні подальші продовження досліджень та розробки.

Постановка завдання. Метою статті є дослідження можливостей NPL моделі DistilBERT для подальшого використання в системі автоматизації електронного документообігу.

Виклад основного матеріалу. DistilBERT є результатом прагнення до оптимізації та ефективності в області обробки природної мови (NLP). Це полегшена версія відомої моделі BERT, яка, незважаючи на зменшену кількість параметрів, зберігає значну частину її продуктивності. Такий результат досягається завдяки інноваційному методу дистиляції знань, який дозволяє меншій моделі «вчитися»у більшої, переймаючи її поведінку та ефективність.

DistilBERT базується на архітектурі Transformer, яка лежить в основі багатьох сучасних моделей NLP. Вона складається з кількох шарів енкодерів, кожен з яких містить механізми уваги та повнозв'язні шари. Особливістю DistilBERT є використання меншої кількості шарів та зменшення розміру прихованих станів порівняно з BERT. Це дозволяє суттєво зменшити кількість параметрів моделі, що прискорює її навчання та роботу.

Ключовим елементом DistilBERT є механізм дистиляції знань. Він полягає в тому, що менша модель навчається не лише на вихідних даних, але й на «м'яких» прогнозах більшої моделі (BERT). Це дозволяє DistilBERT «перейняти» знання та узагальнення, які BERT отримала під час свого навчання, що значно підвищує її ефективність.

DistilBERT демонструє високу точність на багатьох завданнях NLP, включаючи класифікацію тексту, розпізнавання іменованих сутностей, відповіді на запитання та інші. Вона здатна конкурувати з більшими моделями, такими як BERT, незважаючи на меншу кількість параметрів. Завдяки зменшеній кількості параметрів DistilBERT вимагає менше обчислювальних ресурсів та часу на навчання та виконання. Це робить її привабливою для використання на пристроях з обмеженими ресурсами або в ситуаціях, коли швидкість обробки є критичною.

DistilBERT може бути легко налаштована під конкретні завдання та домени. Це дозволяє досягти ще вищої точності та ефективності на спеціалізованих задачах. Модель DistilBERT є відкритою та доступною для використання. Це сприяє її широкому застосуванню та подальшому розвитку в спільноті NLP.

DistilBERT знаходить широке застосування в різних галузях, де потрібна обробка природної мови. Вона використовується для аналізу тональності тексту, класифікації документів, створення чат-ботів, машинного перекладу та багатьох інших завдань. Завдяки своїй ефективності та точності, DistilBERT є перспективним інструментом для автоматизації та оптимізації процесів, пов'язаних з обробкою текстової інформації.

Переваги та можливості DistilBERT:

– Висока точність на завданнях NLP: DistilBERT демонструє вражаючі результати на широкому спектрі завдань обробки природної мови, включаючи класифікацію тексту, вилучення інформації, аналіз тональності, відповіді на запитання та багато інших. Завдяки своїй архітектурі та механізму дистиляції знань, вона здатна досягати точності, порівнянної з більшими моделями, такими як BERT, що робить її потужним інструментом для аналізу та розуміння текстової інформації.

– Менші обчислювальні витрати: Однією з ключових переваг DistilBERT є її ефективність. Завдяки зменшеній кількості параметрів, вона вимагає значно менше обчислювальних ресурсів та часу на навчання та виконання порівняно з BERT. Це робить її привабливою для використання на пристроях з обмеженими ресурсами, таких як мобільні телефони або вбудовані системи, а також у випадках, коли швидкість обробки є критично важливою.

– Можливість тонкого налаштування: DistilBERT є гнучкою моделлю, яку можна легко адаптувати під конкретні завдання та предметні області. Завдяки механізму трансферного навчання, її можна доналаштувати на невеликих обсягах даних, що дозволяє досягти високої точності навіть на спеціалізованих задачах. Це робить її універсальним інструментом для вирішення різноманітних проблем в області NLP.

– Відкритий вихідний код та доступність: DistilBERT є відкритою моделлю з доступним вихідним кодом, що дозволяє дослідникам та розробникам вільно використовувати її, модифікувати та адаптувати під свої потреби. Це сприяє широкому застосуванню моделі та її подальшому розвитку в спільноті NLP.

Експериментальне дослідження. Для оцінки ефективності DistilBERT було проведено один з важливих експерементів а саме категоризація тектових документів.

У експеременті було застосовано багатомовну модель distilbert-base-multilingual-cased, враховуючи її здатність ефективно обробляти великі текстові масиви різними мовами. Метою було створення системи, здатної класифікувати текстові документи на юридичні та неюридичні. Для навчання та перевірки моделі використовувалися відкриті дані: судові рішення (юридичні) та дописи з соціальних мереж (неюридичні), отримані з ресурсу lang.org.ua, що містить великий архів українських текстів, розподілених за тематикою.

Перед початком навчання моделі було застосовано Git LFS для ефективного управління великими обсягами даних. Модель налаштували з такими параметрами (файл config.json): активаційна функція GELU, 6 шарів трансформера, 12 голів уваги, максимальна довжина послідовності 512 токенів, розмір словника 119547. Ці параметри обрані для оптимізації продуктивності та точності моделі.

Процес навчання моделі розпочався з ініціалізації на основі попередньо навченої моделі DistilBERT, адаптованої для бінарної класифікації. Тренування тривало три епохи, що дозволило моделі поступово підлаштуватися до особливостей даних.

Для контролю процесу навчання використовувалися показники втрат (loss) та точності (accuracy) на навчальному та валідаційному наборах даних. Зниження втрат та збільшення точності з кожною епохою свідчили про ефективність навчання.

Після завершення тренування проведено детальний аналіз результатів, включаючи динаміку змін втрат та точності на кожній епосі навчання та валідації.

```
PS C:\Work> cat .\output.txt
"loss": [0.0095907844626663, 0.0014356519095996 51, 1.386978735574 23e-05],
"accuracy": [0.9978122115135193, 0.999749585151672, 1.0],
"val_loss": [0.004804324824362993, 1.6209438399528135e-05, 5.182974746276159e-06],
"val_accuracy": [0.9984999898525, 1.0, 1.0]
```

Рис. 1. Лог результатів навчання

loss: список, який показує значення функції втрат на тренувальному наборі даних після кожної епохи.

accuracy: список, який вказує точність моделі цьому ж наборі даних.

val_loss: список значень функції втрат на валідаційному наборі даних після кожної епохи.

val_loss: список значень функції втрат а цьому ж наборі даних.

val_accuracy: список, який показує точність моделі цьому ж наборі даних.

Початковий етап: вже на першій епосі модель продемонструвала низьку початкову похибку (loss: 0.0096) та високу точність (accuracy: 99.78%) на тренувальних даних. Це свідчить про те, що модель швидко знайшла закономірності в даних та почала робити точні прогнози. Валідаційні дані підтвердили цей успіх, показавши ще нижчі втрати (val_loss: 0.0048) та високу точність (val_accuracy: 99.85%). Це означає, що модель не тільки добре запам'ятовує тренувальні дані, але й здатна узагальнювати та застосовувати знання до нових, не бачених раніше прикладів.

Другий етап: подальше навчання призвело до значного зниження похибки (loss: 0.0014) та подальшого покращення точності (accuracy: 99.97%) на тренувальному наборі. Це свідчить про те, що модель продовжує активно навчатися та коригувати свої параметри для ще точніших прогнозів. На валідаційному наборі втрати практично зникли (val_loss: 1.6209e-05), а точність досягла ідеального значення (val_accuracy: 100%). Це підтверджує, що модель не перенавчається на тренувальних даних, а дійсно знаходить загальні закономірності, які дозволяють їй робити безпомилкові прогнози на нових даних.

Третій етап: на останній епосі втрати на тренувальному наборі знизились до мінімуму (loss: 1.387e-05), а модель досягла ідеальної точності (accuracy: 100%). Це свідчить про те, що модель повністю освоїла навчальний матеріал та більше не може покращити свої результати на тренувальних даних. На валідаційному наборі втрати також ще зменшились (val_loss: 5.183e-06), а ідеальна точність (val_accuracy: 100%) збереглася. Це підтверджує, що модель досягла своєї максимальної продуктивності та готова до використання на реальних даних.

Таким чином, аналіз результатів навчання моделі показав її високу ефективність, швидке навчання та здатність до узагальнення. Модель досягла ідеальної точності на обох наборах даних, що свідчить про її готовність до практичного застосування.

Для оцінки загальної продуктивності моделі було використано різноманітний набір тестових даних, що включав різні тексти, серед яких були і юридичні документи, художню літературу та статті з Вікіпедії.

Тексти були підготовлені до використання моделлю шляхом токенізації за допомогою DistilBertTokenizerFast з обмеженням максимальної довжини послідовності у 512 токенів.

Модель була ініціалізована з використанням вагів предтренованої моделі DistilBERT та налаштована на вирішення задачі бінарної класифікації.

Для оптимізації моделі було застосовано оптимізатор Adam з швидкістю навчання 5e-5, а також встановлено функцію втрат SparseCategoricalCrossentropy та метрику точності.

Продуктивність моделі була оцінена на тестовому наборі даних з використанням партії по 16 прикладів.

Аналіз результатів показав високу ефективність моделі, що підтверджується мінімальними втратами, високою точністю та високими значеннями F1-оцінки, точності та повноти.

Отримані результати експерименту були зафіксовані та виведені для подальшого детального аналізу.

```
INFO:root: Eval loss: 0.004677
INFO:root: Eval accuracy: 0.998749
INFO:root: F1 Score: 0.997504
INFO:root: Precision: 0.995020
INFO:root: Recall: 1.000000
```

Рис. 2. Оцінка ефективності моделі

Дослідження нейромережевої системи для класифікації текстів, що використовує модель distilbert-base-multilingual-cased, виявило її високу результативність. Оцінка ефективності моделі за ключовими показниками, такими як eval loss, eval accuracy, F1-оцінка, precision та recall, підтверджує її здатність точно визначати категорії текстів. Зокрема, значення F1-оцінки, близькі до 0.99 (див. таблицю), вказують на відмінну збалансованість точності та повноти класифікації.

Результати експериментів показали, що DistilBERT досягає високої точності на всіх розглянутих завданнях. Зокрема, модель продемонструвала:

– Точність класифікації документів: 95%.

– F1-міра для вилучення інформації на тестовому датасеті: 99%.

– ROUGE-L для зведення тексту: 65%.

Отримані результати підтверджують, що DistilBERT є перспективним інструментом для

автоматизації ЕДО. Модель здатна ефективно обробляти текстові документи українською мовою, виявляти важливу інформацію та генерувати короткі реферати.

Висновки. У цій статті було досліджено можливості використання моделі DistilBERT для автоматизації електронного документообігу. Результати експериментів показали, що DistilBERT є ефективним інструментом для вирішення завдань класифікації, вилучення інформації та зведення тексту. Модель демонструє високу точність та швидкість роботи, що робить її перспективною для впровадження в системах ЕДО.

Подальші дослідження можуть бути спрямовані на:

– Тонке налаштування моделі під конкретні завдання.

– Розробку гібридних підходів, що поєднують DistilBERT з іншими методами NLP.

– Дослідження можливостей використання DistilBERT для інших завдань ЕДО, таких як пошук документів, аналіз тональності тексту, виявлення дублікатів.

Список літератури:

1. Sanh, V. *et al.* (2019) 'DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,' *arXiv (Cornell University)* [Preprint]. https://doi.org/10.48550/arxiv.1910.01108.

2. Lee, J.-S. and Hsiang, J. (2020) 'Patent classification by fine-tuning BERT language model,' *World Patent Information*, 61, p. 101965. https://doi.org/10.1016/j.wpi.2020.101965.

3. Miller, D. (2019) 'Leveraging BERT for extractive text summarization on lectures,' *arXiv (Cornell University)* [Preprint]. https://doi.org/10.48550/arxiv.1906.04165.

4. Krieger, F., Drews, P. and Funk, B. (2023) 'Automated invoice processing: Machine learning-based information extraction for long tail suppliers,' *Intelligent Systems With Applications*, 20, p. 200285. https://doi.org/10.1016/j.iswa.2023.200285.

5. Graham, S.G., Soltani, H. and Isiaq, O. (2023) 'Natural language processing for legal document review: categorising deontic modalities in contracts,' *Artificial Intelligence and Law* [Preprint]. https://doi.org/10.1007/s10506-023-09379-2.

6. Шаптала, Р. and Кисельов, Г. (2023) 'Класифікація текстових документів з використанням доповнення векторних представлень документів графовими представленнями елементів словника синонімів,' *INFORMATION TECHNOLOGY AND SOCIETY*, (3 (5)), pp. 49–55. https://doi.org/10.32689/maup.it.2022.3.6.

7. Pavlyshenko, B. and Stasiuk, M. (2024) 'DATA AUGMENTATION IN TEXT CLASSIFICATION WITH MULTIPLE CATEGORIES,' *Electronics and Information Technologies*, 25. https://doi.org/10.30970/eli.25.6.

8. Prytula M. (2024) 'Fine-tuning BERT, DistilBERT, XLM-RoBERTa and Ukr-RoBERTa models for sentiment analysis of ukrainian language reviews,' *Štučnij Ìntelekt*, 29(AI.2024.29(2)), pp. 85–97. https://doi.org/10.15407/jai2024.02.085.

9. Comparative Analysis of State-of-the-Art Q&A Models: BERT, RoBERTa, DistilBERT, and ALBERT on SQuAD v2 Dataset. (2024). Chaos and Fractals, 1(1), 19-30. https://doi.org/10.69882/adba.chf.2024073

Cherniak I.O. RESEARCH OF DISTILBERT MODEL OPPORTUNITIES FOR FURTHER USE IN THE SYSTEM OF AUTOMATION OF ELECTRONIC DOCUMENTATION

This research article explores the potential of using the DistilBERT model to automate electronic document management (EDM) processes. EDI is an integral part of the modern business environment, as it allows to significantly reduce the time and resources spent on processing, storing and transferring documents. However, large volumes of documents and the variety of their formats create certain challenges for effective EDI automation.

DistilBERT, as one of the most advanced natural language processing (NLP) models, offers an effective solution to these problems. Developed on the basis of the BERT (Bidirectional Encoder Representations from Transformers) architecture, DistilBERT retains high text processing accuracy, but is significantly smaller and faster. This makes it an ideal tool for implementation in EDI systems where speed and efficiency are critical factors.

The article discusses the features of DistilBERT architecture and its advantages over other NLP models. In particular, the article analyzes DistilBERT's ability to understand the context and semantics of text, which is key to successful EDI automation. We also consider the possibility of fine-tuning the model for specific tasks, which allows to achieve maximum efficiency in various applications.

A series of experiments were conducted to evaluate the effectiveness of DistilBERT on EDI tasks. The model was tested on the tasks of document classification, key information extraction, and text summarization. The results of the experiments showed high accuracy and efficiency of DistilBERT in all test tasks. In particular, the model demonstrated the ability to accurately classify documents by their type and content based on the test

data, extract important information such as dates, names, addresses, etc., and create short and informative summaries of large texts.

These results confirm the prospects of using DistilBERT for EDI automation. Due to its high accuracy, speed and contextual understanding capabilities, DistilBERT can significantly improve the efficiency of document processing, reduce errors and save time and resources. This is especially true for large organizations where document volumes are significant.

The article also discusses potential areas for further research. In particular, it considers the possibility of using DistilBERT for other EDI tasks, such as automatic form filling, document compliance checking, etc. We also suggest ways to improve the model, for example, by using additional data for training or by combining DistilBERT with other NLP models.

In general, this research article makes a significant contribution to the development of EDI automation technologies. The proposed use of the DistilBERT model improves the possibilities for increasing the efficiency and productivity of document management, which can have a significant positive impact on various fields of activity.

Key words: *DistilBERT, electronic document management (EDM), natural language processing (NLP).*