ДОДАТОК А

Графічний матеріал кваліфікаційної роботи

Харківський національний університет радіоелектроніки
Кафедра ЕОМ

Модель нейромережевої системи для категоризації текстових документів

Кваліфікаційна робота
Другий (магістерський) рівень

Виконав: студент кафедри ЕОМ ХНУРЕ
Рибалов О.О. СПМ-22-5

Науковий керівник:
док. тех. наук:
професор кафедри ЕОМ ХНУРЕ
Фесенко Т.Г

Харків 2024

Об'єкт, предмет та мета дослідження

Об’єктом роботи є процеси автоматизованої категоризації текстових документів.

Предметом роботи є використання нейромережевих моделей, зокрема DistilBERT, для ефективної категоризації текстових документів у юридичній та інших сферах.

Метою роботи є розробка нейромережевої системи для категоризації текстових документів за рахунок застосування машинного навчання.

Завдання дослідження

1) **Проаналізувати методи категоризації текстових документів.** Дослідити сучасні підходи з використанням моделей BERT та DistilBERT.

2) **Запропонувати модель нейромережевої системи** на основі DistilBERT для автоматизації категоризації текстів.

3) **Сформувати та підготувати дані з різних джерел** для тренування та валідації запропонованої моделі.

4) **Навчити та налаштувати модель.** Вибір оптимальних параметрів моделі для досягнення високої точності та ефективності.

5) **Експериментально перевірити модель.** Проведення експерименту оцінки продуктивності та ефективності моделі на прикладі текстових документів.

6) **Оцінити результати** експериментальних досліджень та підтвердити чи спростувати дієвість моделі для категоризації текстових документів.

3

Методи розпізнавання текстової інформації

1. **Частотний аналіз (FA) та тематичне моделювання (LDA)**	Частотний аналіз використовує підрахунок частоти слів або фраз у тексті для виявлення основних тем або тенденцій. Тематичне моделювання, зокрема метод латентного розміщення Діріхле (LDA), дозволяє визначати приховані теми у великому корпусі текстів.
2. Аналіз сентиментів за допомогою ШНМ (ANN)	Аналіз сентиментів за допомогою штучних нейронних мереж (ШНМ) включає визначення емоційного тонусу тексту (позитивний, негативний, нейтральний) шляхом використання моделей машинного навчання.
3. Сетевий аналіз для вивчення зв'язків (NA)	Сетевий аналіз дозволяє досліджувати взаємозв'язки між різними елементами тексту (наприклад, між персонажами або темами) за допомогою графів та мереж.
4. Машинне навчання для класифікації текстів (ML)	Використання алгоритмів машинного навчання для автоматичної класифікації текстів на категорії, такі як спам/не спам, новини, огляди тощо.
5. Видобування інформації (IE)	Процес автоматичного витягання структурованої інформації з неструктурованого тексту, наприклад, імен, дат, місць, подій.

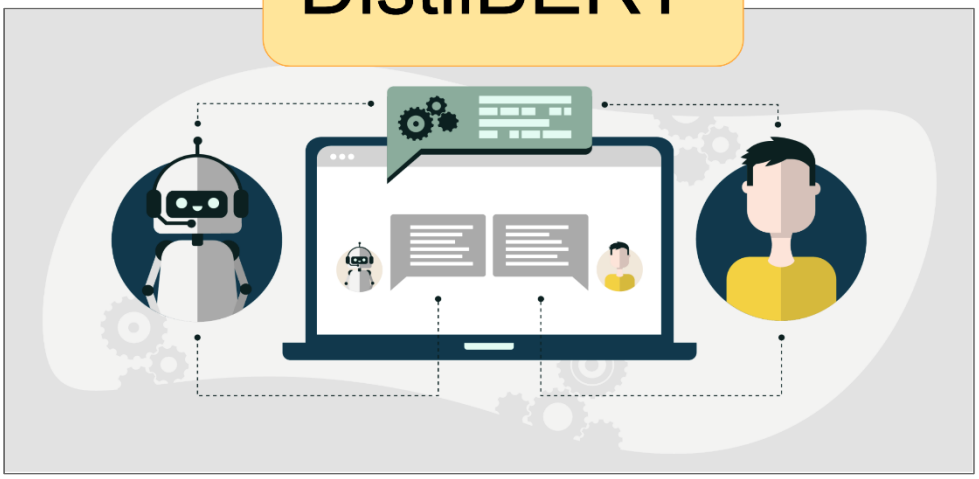
4

Методи машинного навчання

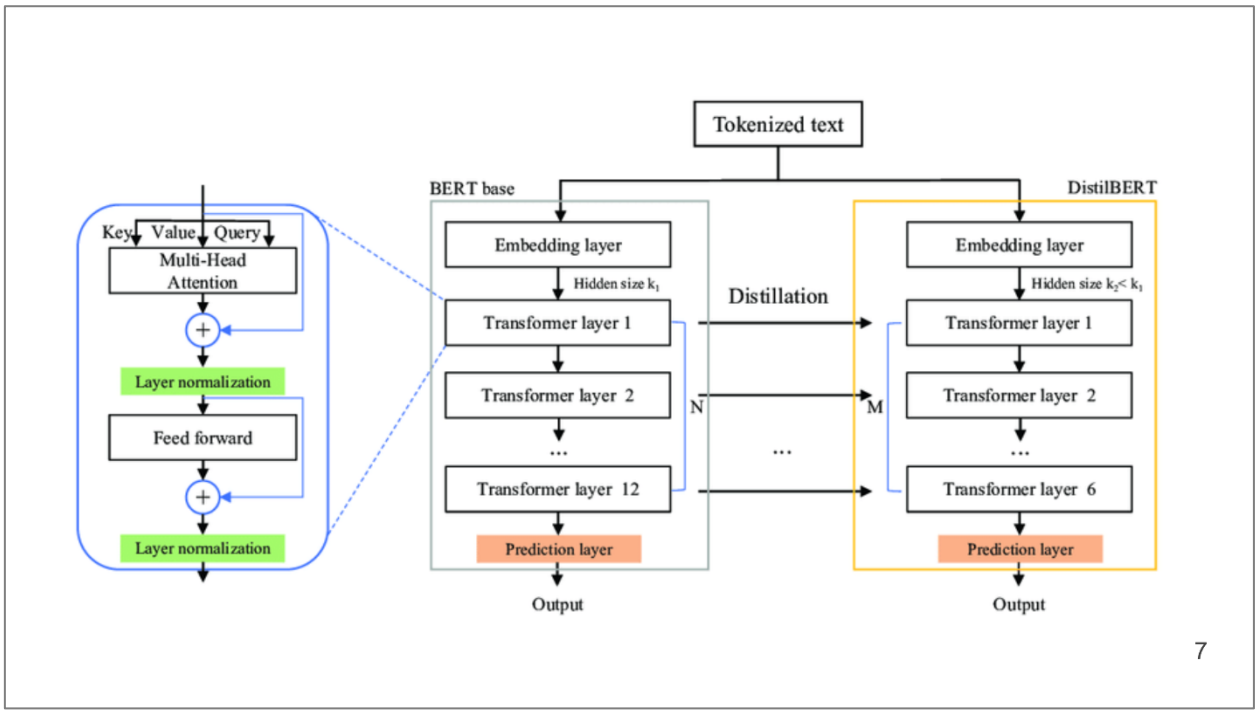
В даній роботі використовується такі методи машинного навчання, як трансформерна архітектура, дистиляція знань та оптимізація з використанням алгоритму Adam, для досягнення відповідних цілей в обробці природної мови (NLP).

5

DistilBERT

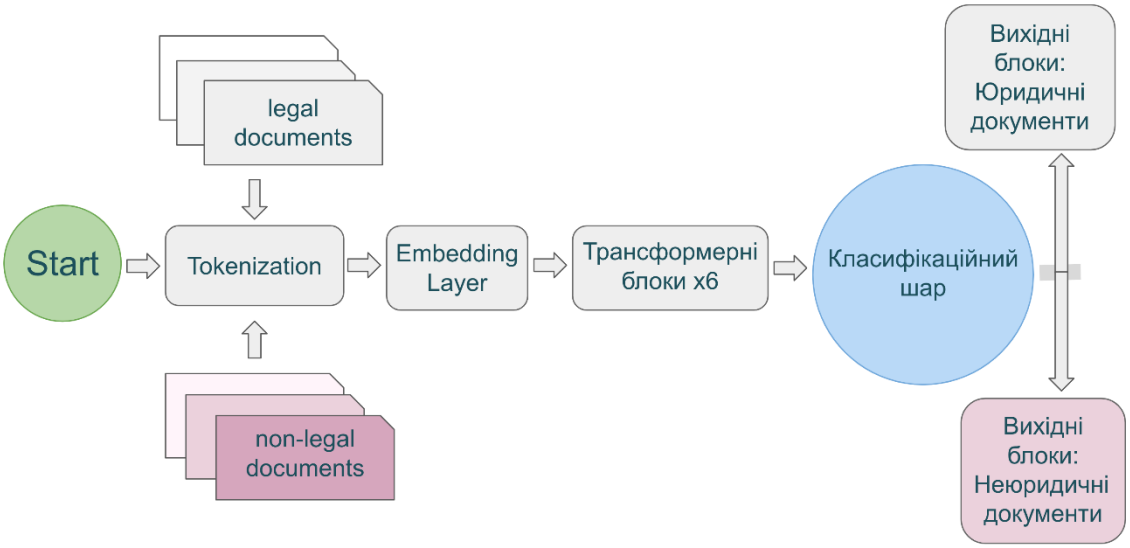


6



7

Схема моделі нейромережевої системи для класифікації текстових документів



8

Математичний опис моделі

$$X = [x_1, x_2, ..., x_n]$$

X — послідовність токенів тексту *T*,
xi *- індекс токена в словнику моделі,*
n *— довжина послідовності*

$$E_i = Embedding(x_i) + PositionEmbedding(i)$$

Ei — вектор ембедингу для *i-го токена,*
Embedding(xi) — ембединг токена xi,
PositionEmbedding(*i*) — позиційний ембедингу для *i-го токена,*

$$H^{(l)} = FFN(SelfAttention(H^{(l-1)}))$$

H(l) — вихід l-го трансформерного блоку,
FFN — підшар прямого поширення,
SelfAttention(H(l-1)) — механізм самоуваги,

9

Математичний опис моделі

$$P(class|X) = softmax(Z)$$

***P* (class|X)** — ймовірність того, що вхідна послідовність X належить до певного класу,
Z — вектор логітів,
softmax — функція, яка перетворює вектор логітів Z у розподіл ймовірностей по класах,

$$L(y, \hat{y}) = -\Sigma_{i=1}^{C} y_i \log(\hat{y}_i)$$

L(y, y) — функція втрат (loss function), яка вимірює,
y — істинна мітка класу в форматі цілого числа ,
y — прогнозований розподіл ймовірностей для всіх класів,
C — кількість класів,
yi — індикатор,
log(yi) — натуральний логарифм ймовірності,

$$w_{i+1} = w_t - \frac{\eta}{\sqrt{\hat{w}_t + \epsilon}} \hat{m}_t$$

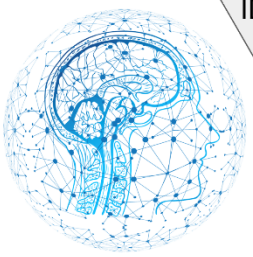
w_{i+1} — оновленим значенням ваги моделі
w_t — ваги на кроку t,
η — навчальний крок (learning rate),
$\hat{m}_t$ — оцінка першого моменту,
$\hat{v}_t$ — оцінка другого моменту,
ϵ — дуже мале число

10

Використані технології

мова програмування - Python 3

IDE - PyCharm 2023.3 (Community Edition)



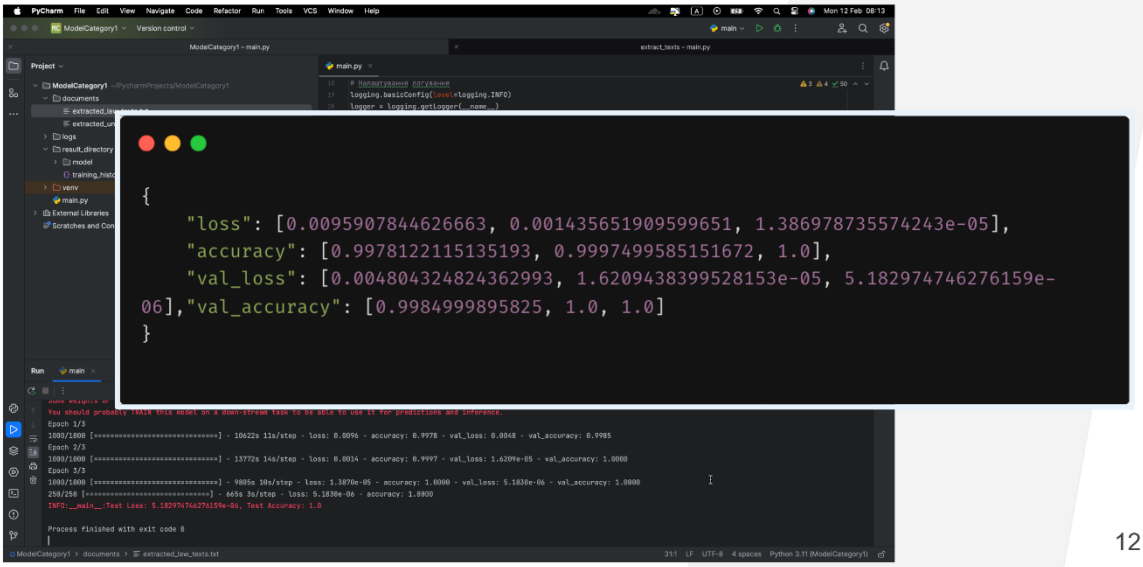
tensorflow - бібліотека машинного навчання та глибокого навчання

scikit-learn - бібліотека машинного навчання

transformers - бібліотека яка призначена для роботи з моделями обробки природної мови (NLP)

11

Процес тренування



12

Результати тестування моделі

Параметри ефективності моделі

Metric	Value
Eval loss	0.004677
Eval accuracy	0.998749
F1 Score	0.997504
Precision	0.995020
Recall	1.000000

13

Висновки

Досліджено сучасні методи та підходи до категоризації текстових документів із використанням штучних нейронних мереж, зокрема моделей BERT та DistilBERT. Визначено переваги та недоліки цих методів у контексті їх застосування в сфері категоризації текстових документів.

Проаналізовано існуючі дослідження в галузі категоризації текстових документів, включаючи роботи з використанням моделей BERT та Legal-BERT, та зроблено висновки щодо їх ефективності у вирішенні завдань автоматизованого розпізнавання та класифікації текстової інформації.

Запропоновано модель нейромережевої системи для категоризації текстових документів на основі моделі DistilBERT, яка поєднує високу точність та ефективність з низькими вимогами до обчислювальних ресурсів. На основі проведеного дослідження можна зробити висновок про виключну ефективність моделі DistilBERT у задачах категоризації текстових документів.

14

Висновки

Розроблено
програмну реалізацію системи, яка включає етапи токенізації, підготовки даних, навчання моделі та її оцінки. Програмна реалізація дозволяє автоматизувати процес категоризації документів та забезпечує високу точність результатів.

Апробовано
розроблену систему на тестових наборах даних, що включають юридичні та неюридичні тексти. Експериментальні дослідження підтвердили високу ефективність та точність моделі, зокрема досягнуто значення F1-оцінки 0.997504, що свідчить про високий рівень класифікації текстових документів.

15

Результати апробації

Рибалов О. О., Фесенко Т. Г. 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, С. 330–331

Рибалов О. О. Оцінка ефективності нейромережевої системи для категоризації текстових документів. 28-й Міжнародний молодіжний форум «Радіоелектроніка та молодь XXI століття», 16-18 квітня 2024 року: зб. матеріалів форуму. Т.5., конференція «Проблеми комп'ютерної інженерії та захисту інформації». Харків: ХНУРЕ, 2004. С. 56-57.

Рибалов О. О., Фесенко Т. Г. 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, С. 330–331

16

ДОДАТОК Б

Посилання на код моделі та дамп даних, які використовувались для тренування моделі

Посилання на код моделі та тестів – https://drive.google.com/drive/folders/1D0ULJoMQ97RmC0XX_Z3NYHhUf8_EjvVf

Lang-uk projects. UberText 2.0.: корпус сучасних українських текстів, призначених для задоволення різноманітних потреб НЛП). doi: https://lang.org.ua/en/ubertext/ (дата звернення: 09.03.2024)