Магістерська кваліфікаційна робота

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ КЛАСИФІКАЦІЇ БАНКІВСЬКИХ ТЕКСТІВ НА ОСНОВІ ЗГОРТКОВОЇ НЕЙРОННОЇ МЕРЕЖІ

Виконав: Переродов А. О.
Науковий керівник: Колесницький О.К.

Актуальність

класифікація текстових документів для їх подальшого перенаправлення у відповідні відділи в банківських установах є досить актуальною проблемою через велику кількість вхідної кореспонденції, яка призводить до перезавантаженості служб банківського моніторинг

* **Метою магістерської кваліфікаційної роботи** є підвищення достовірності класифікації банківських текстів програмними засобами за рахунок застосування згорткових нейронних мереж.
* Для досягнення мети розробки необхідно виконати такі **задачі:**
  * провести аналіз проблеми розв'язання задачі класифікації банківських текстів;
  * розглянути існуючі методи вирішення задачі класифікації банківських текстів та обрати й обґрунтувати вибір методу, який задовольняє мету даної магістерської кваліфікаційної роботи;
  * розробити математичну модель класифікації банківських текстів;
  * сформулювати стадії інформаційної технології, розробити структуру та алгоритм роботи програмного засобу;
  * виконати програмну реалізацію запропонованої інформаційної технології;
  * провести тестування програмного продукту та виконати аналіз отриманих результатів.

* **Об’єкт дослідження** – процес класифікації банківських текстів з використанням згорткових нейронних мереж.
* **Предмет дослідження** – інформаційна технологія та програмні засоби класифікації банківських текстів з використанням згорткових нейронних мереж та достовірність їх роботи.
* **Методи дослідження.** У роботі використані наступні методи наукових досліджень: системного аналізу, інтелектуального аналізу даних, теорії штучних нейронних мереж для реалізації інформаційної технології класифікації банківських текстів, методи математичної статистики для розробки процесу класифікації банківських текстів та обрахунків результатів експериментів із програмним засобом, об’єктно-орієнтованого програмування.

НАУКОВА НОВИЗНА ОДЕРЖАНИХ РЕЗУЛЬТАТІВ

•Набула подальшого розвитку інформаційна технологія класифікації банківських текстів, яка відрізняється використанням згорткової штучної нейронної мережі, що дозволило підвищити достовірність класифікації банківських текстів.

• Удосконалено метод попередньої обробки тексту для подальшого розпізнавання нейронною мережею, який відрізняється використанням процедури видалення стоп-слів перед подачею тексту на вхід нейронної мережі, що дозволило уникнути «зашумлення» ознак і тим самим підвищити достовірність класифікації банківських текстів.

ПРАКТИЧНЕ ЗНАЧЕННЯ ОДЕРЖАНИХ РЕЗУЛЬТАТІВ

• розроблено алгоритм роботи програмного забезпечення класифікації банківських текстів на основі згорткової нейронної мережі;

• розроблено програмні засоби для класифікації банківських текстів на основі згорткової нейронної мережі;

Аналіз предметної області класифікації банківських текстів

Основними методами класифікації текстів є:

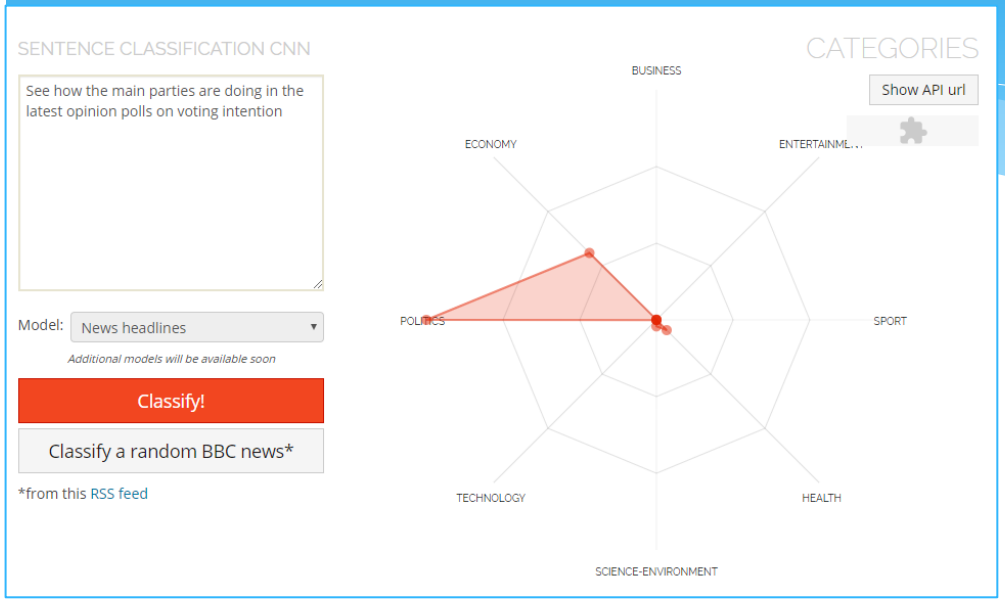
- наївний баєсівський класифікатор;
- метод k-найближчих сусідів;
- дерева рішень;
- метод опорних векторів;
- методи на основі штучних нейронних мереж.

Аналіз методів векторизації слів

Основні методи, які використовуються для перетворення слів у вектор:

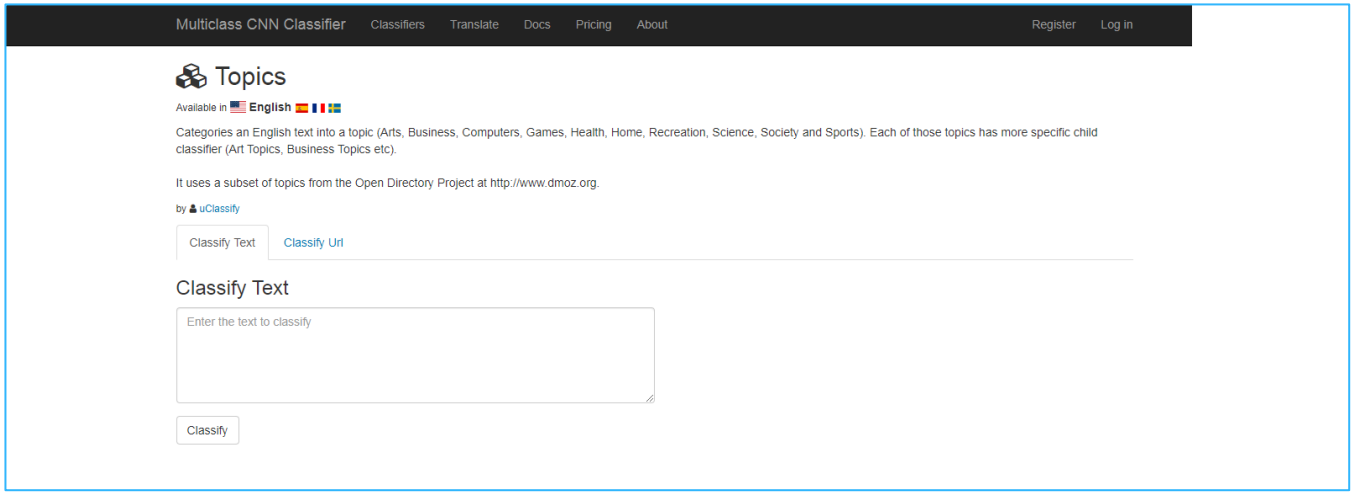
- one-hot encoding;
- word2vec;
- glove.

Вибір і обґрунтування аналогу



Програма «Sentence Classification CNN»

Програма «Multiclass CNN Classifier»



Постановка задачі

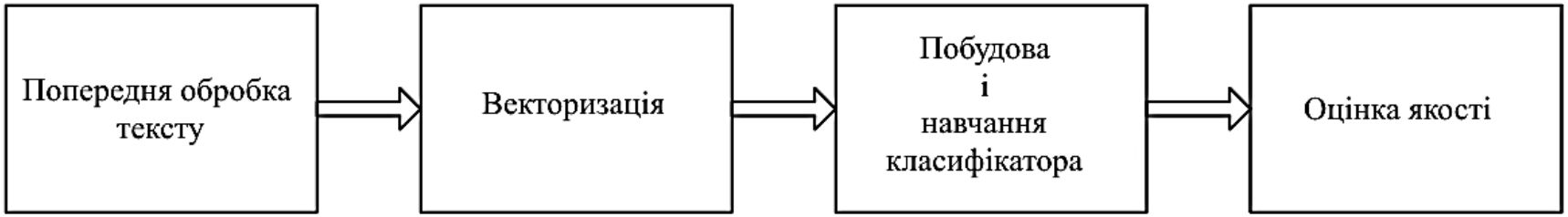
$D = \{ d_1, \dots, d_n \}$ – множина текстових документів. Кожний документ $d \in D$ представляє собою послідовність слів $W_d = \{w_1, \dots, w_{n_d}\}, n_d$ – довжина документа d.

$Y = \{y_1, \dots, y_n\}$ – кінцева множина класів.

$y^*: D \to Y$ – невідома цільова залежність, значення якої відомі тільки для об'єктів навчальної вибірки $D^m = \{(d_1, y_1), \dots, (d_m, y_m)\}$.

Потрібно розробити інтелектуальний модуль, в якому буде реалізовано алгоритм $a: D \to Y$.

Етапи процесу автоматичної класифікації текстів



Метод класифікації – згорткова нейронна мережа

Основні переваги:

- автоматичний вибір ознак класифікації в ході навчання нейронної мережі;
- можливість використання готових векторизаторів слів;
- зменшення розмірності вхідних даних (особливість згорткових нейронних мереж).

Векторизація вхідного тексту

Для векторизації вхідного тексту була застосована **word2vec** векторизація.

- Відносно невелика розмірність вектору;
- Враховує семантичну схожість слів;
- Нейромережевий алгоритм;
- Для успішного навчання моделі необхідний великий корпус текстів.

Попередня обробка тексту

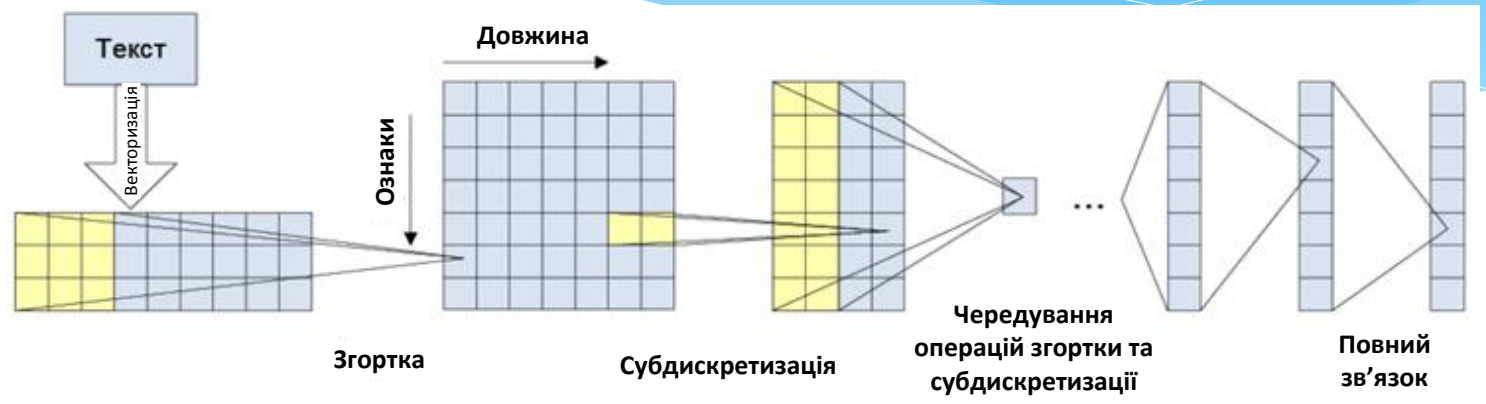
- Видалення стоп-символів

SC = { «"», «!», «#», «$», «%», «&», «'», «(», «)», «*», «+», «“», «”», «-», «,», «.», «/», «:», «;», «<», «=», «>», «?», «@», «\», «^», «_», «`», «{», «|», «}», «~», «\» }

- Видалення стоп-слів

SW = { «do», «does», «did», «a», «an», «the», «be», «been», «was», «were», «could», «will», «would», «shall» }

Архітектура згорткової нейронної мережі



Структурні елементи згорткової нейронної мережі для задачі класифікації текстів є:

- вхідний шар (400 * 100);
- згортковий шар з ядром згортки 3 x 100 (398 * 1 * 32) -> агрегувальний шар (32 * 1 * 1);
- згортковий шар з ядром згортки 4 x 100 (397 * 1 * 32) -> агрегувальний шар (32 * 1 * 1);
- згортковий шар з ядром згортки 5 x 100 (396 * 1 * 32) -> агрегувальний шар (32 * 1 * 1);
- повнозвязний вихідний шар (12).

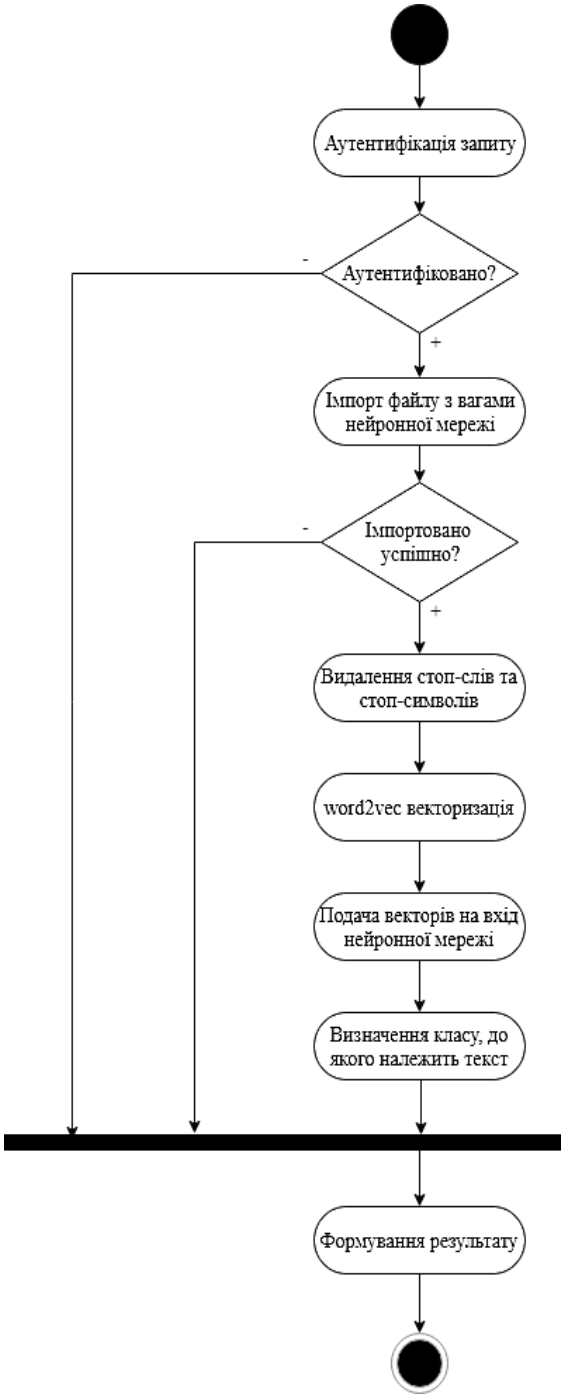
Структура інформаційної технології класифікації банківських текстів

Програмне забезпечення класифікації текстів

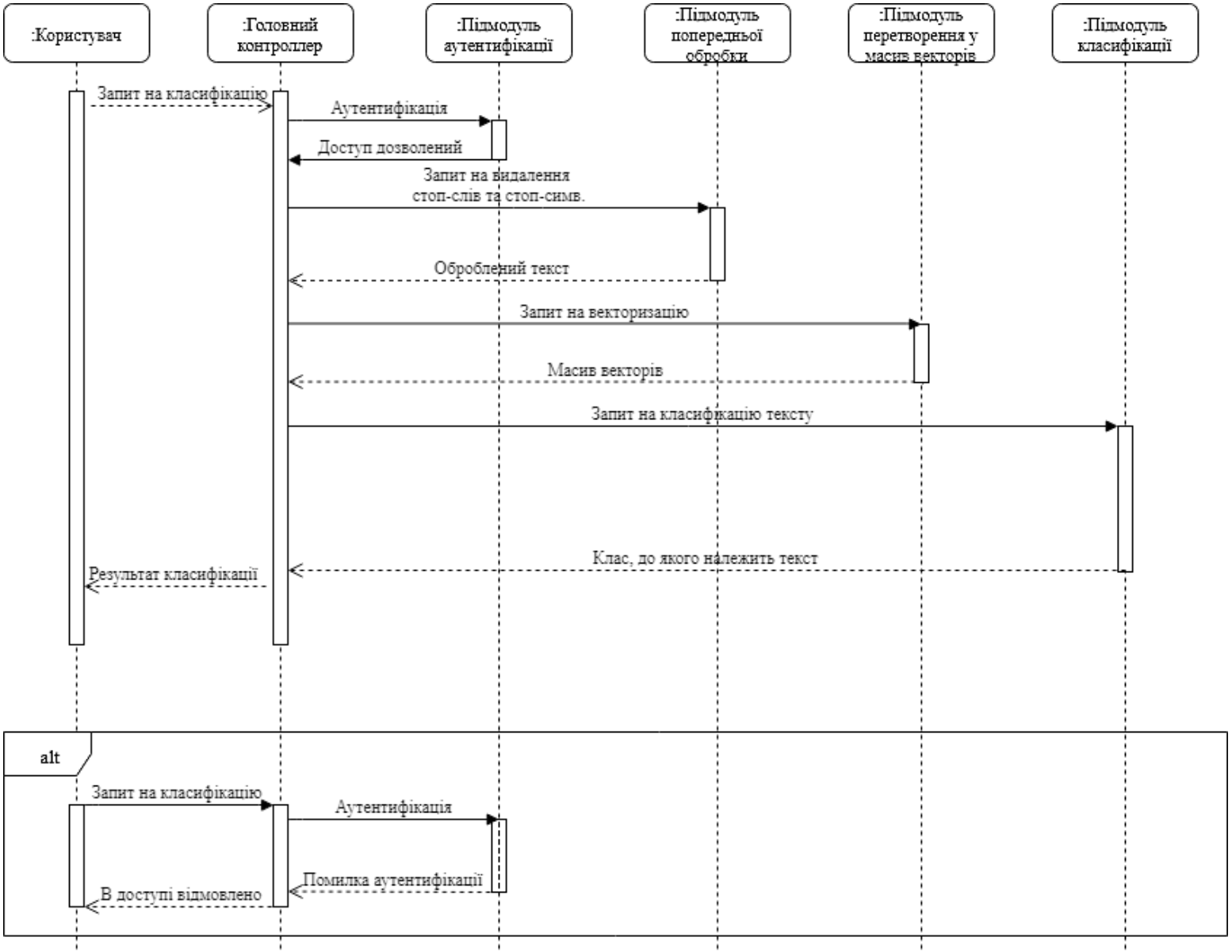
* модуль попередньої обробки слів (Python);
* модуль векторизації (Python);
* модуль класифікації (Python);
* модуль аутентифікації (C#).

Програмне забезпечення класифікації текстів працює в режимі веб-сервісу (WebAPI) з аутентифікацією запитів по токену.

Загальна UML-діаграма активності програми класифікації текстів

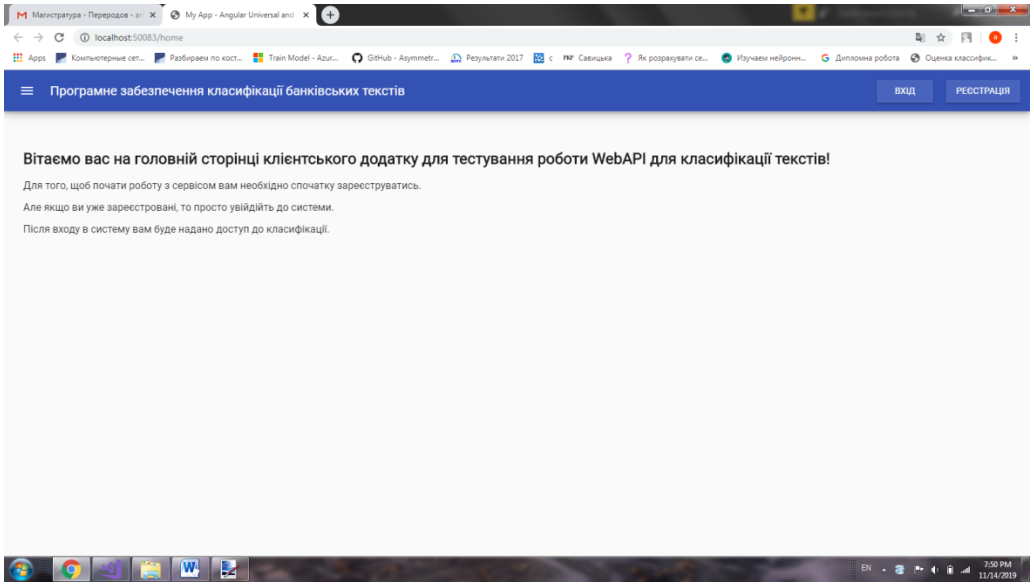


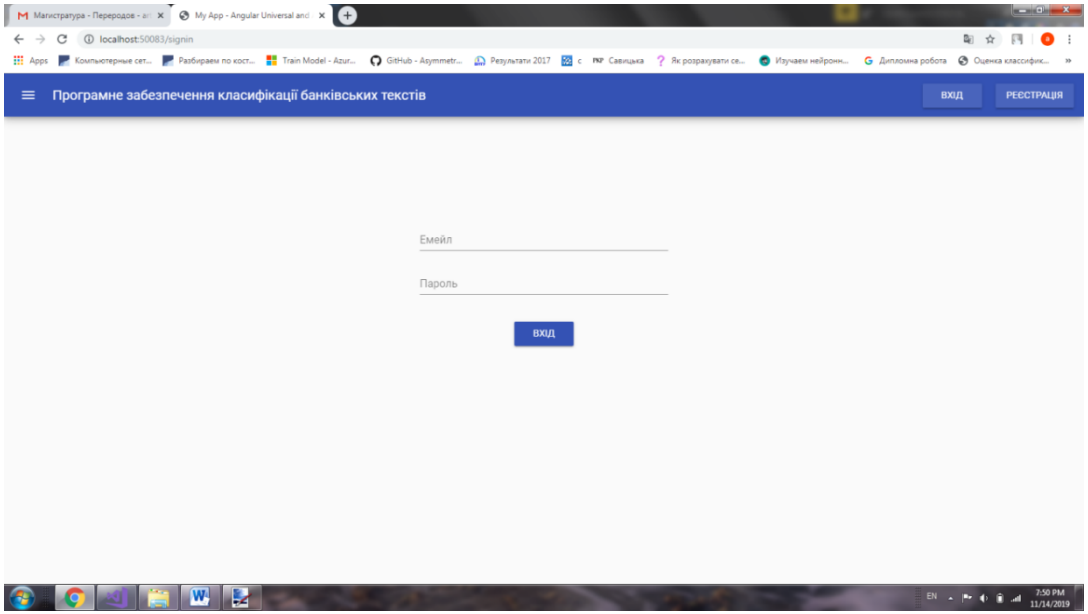
UML-діаграма послідовностей програмного забезпечення класифікації текстів



СТАРТОВІ ВІКНА ПРОГРАМИ КЛАСИФІКАЦІЇ БАНКІВСЬКИХ ТЕКСТІВ НА ОСНОВІ ЗГОРТКОВОЇ НЕЙРОННОЇ МЕРЕЖІ

Головна сторінка

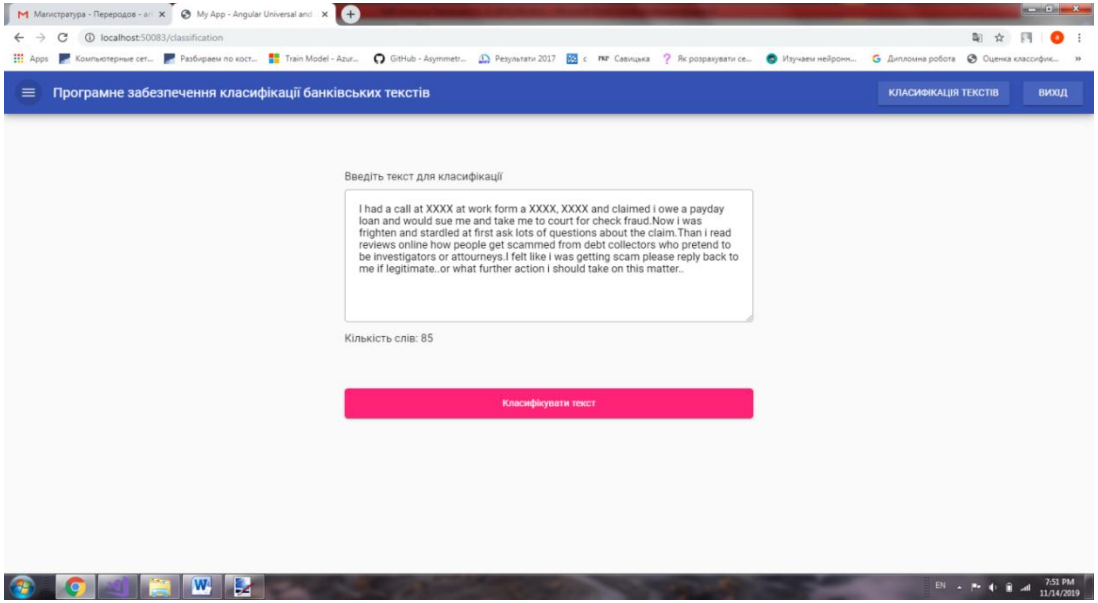


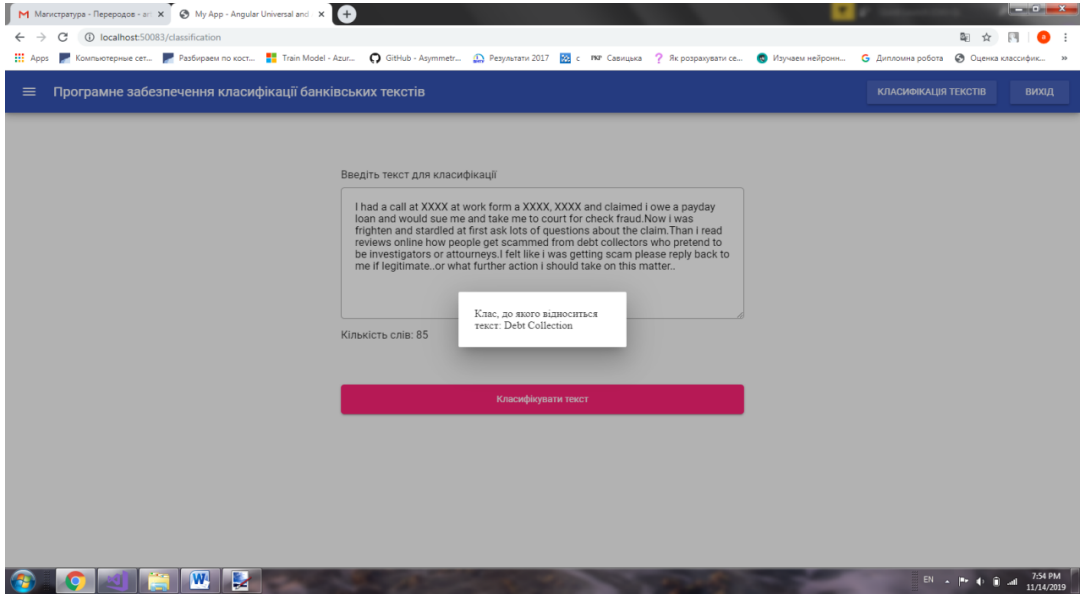


Сторінка входу до системи

РЕЗУЛЬТАТИ РОБОТИ ПРОГРАМИ КЛАСИФІКАЦІЇ БАНКІВСЬКИХ ТЕКСТІВ НА ОСНОВІ ЗГОРТКОВОЇ НЕЙРОННОЇ МЕРЕЖІ

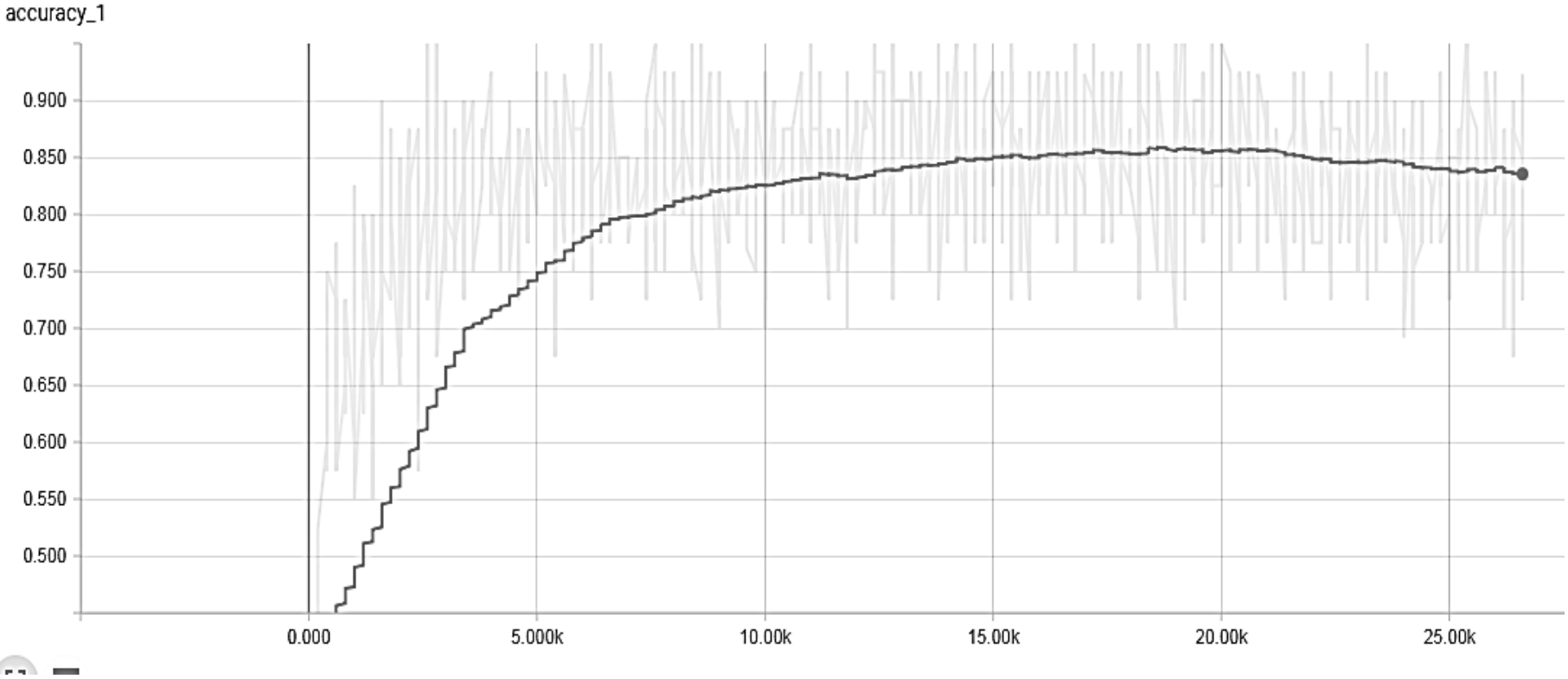
Сторінка класифікації текстів





Результат класифікації тексту

Результати навчання нейронної мережі



Порівняння результатів

Характеристика | Засіб	«Multiclass CNN Classifier»	«Sentence Classification CNN»	Розроблена програма класифікації текстів
Точність класифікації	78.1%	75.8%	85.3%
Видалення стоп-символів	+	+	+
Видалення стоп-слів	-	-	+
Використання готових векторизаторів, навчених на великій вибірці даних	-	-	+
Робота у режимі WebAPI	-	-	+
Аутентифікація користувача	-	-	+

Формула оцінки точності

$$accuracy = \frac{Nt}{Nall} * 100\%$$, де

Nt — кількість правильно класифікованих текстів тестової вибірки,

$Nall$ — загальна кількість текстів тестової вибірки,

Із табл. видно, що розроблена програма має вищу достовірність класифікації (85,3%), ніж краща з аналогічних програм (78,1%), а значить достовірність класифікації текстів покращена щонайменше на 7,2%, тобто мета роботи досягнута.

ЕКОНОМІЧНА ЧАСТИНА

Було проведено економічне обґрунтування доцільності розробки програми класифікації банківських текстів, нова розробка має високий рівень комерційного потенціалу- середньоарифметична сума балів становить 43,5. Загальна сума витрат на виконання означених робіт склала 32384,05 грн., абсолютна ефективність вкладених інвестицій становить 206775,47 грн. Відносна (щорічна) ефективність вкладених в наукову розробку інвестицій – 94 %, отже інвестор буде зацікавлений у фінансуванні даної наукової розробки. Термін окупності складає 1,06 року. В загальному можна зробити висновок, що фінансування розробки програми класифікації банківських текстів з використанням згорткової нейронної мережі є економічно доцільним проектом.

Апробація наукових досліджень

Результати даного дослідження доповідались на

1) конференції «ІОН-2018» («Методи перетворення слова у вектор фіксованої довжини для задачі класифікації текстів») та
2) опубліковані у матеріалах конференції «Молодь в науці: дослідження, проблеми, перспективи-2020», Вінниця, 2019 (Інформаційна технологія класифікації банківських текстів на основі згорткової нейронної мережі)

Публікації.

За результатами магістерської кваліфікаційної роботи опубліковано 2 тез доповідей на конференції

Висновки

- В результаті виконання МКР розроблено інформаційну технологію та програмне забезпечення для класифікації банківських текстів на основі згорткової нейронної мережі. Програмне забезпечення створено такими мовами програмування: модуль класифікації текстів – Python, модуль аутентифікації – C#, клієнтський веб-додаток тестування роботи WebAPI – Typescript та бібліотеки Tensorflow та Flask. Програма має вищу на 7,2% достовірність класифікації. Таким чином, мета роботи досягнута – достовірність класифікації банківських текстів підвищена.

ДЯКУЮ ЗА УВАГУ!