Кваліфікаційна робота на тему «Автоматична класифікація текстів»

ВИКОНАВ СТУДЕНТ ІПЗ-4 ДУБОВИК АНДРІЙ

НАУКОВИЙ КЕРІВНИК: ВОЛИНЕЦЬ Є. А.

Мета роботи

Аналіз та порівняння різних методів класифікації текстів

Реалізація власного рішення для класифікації

Огляд відомих підходів

Naive Bayes classifier

Support Vector Machine

Recurrent Neural Network

BERT

LLM

Алгоритм розв’язку

Використання набору даних «AG News Classification Dataset»

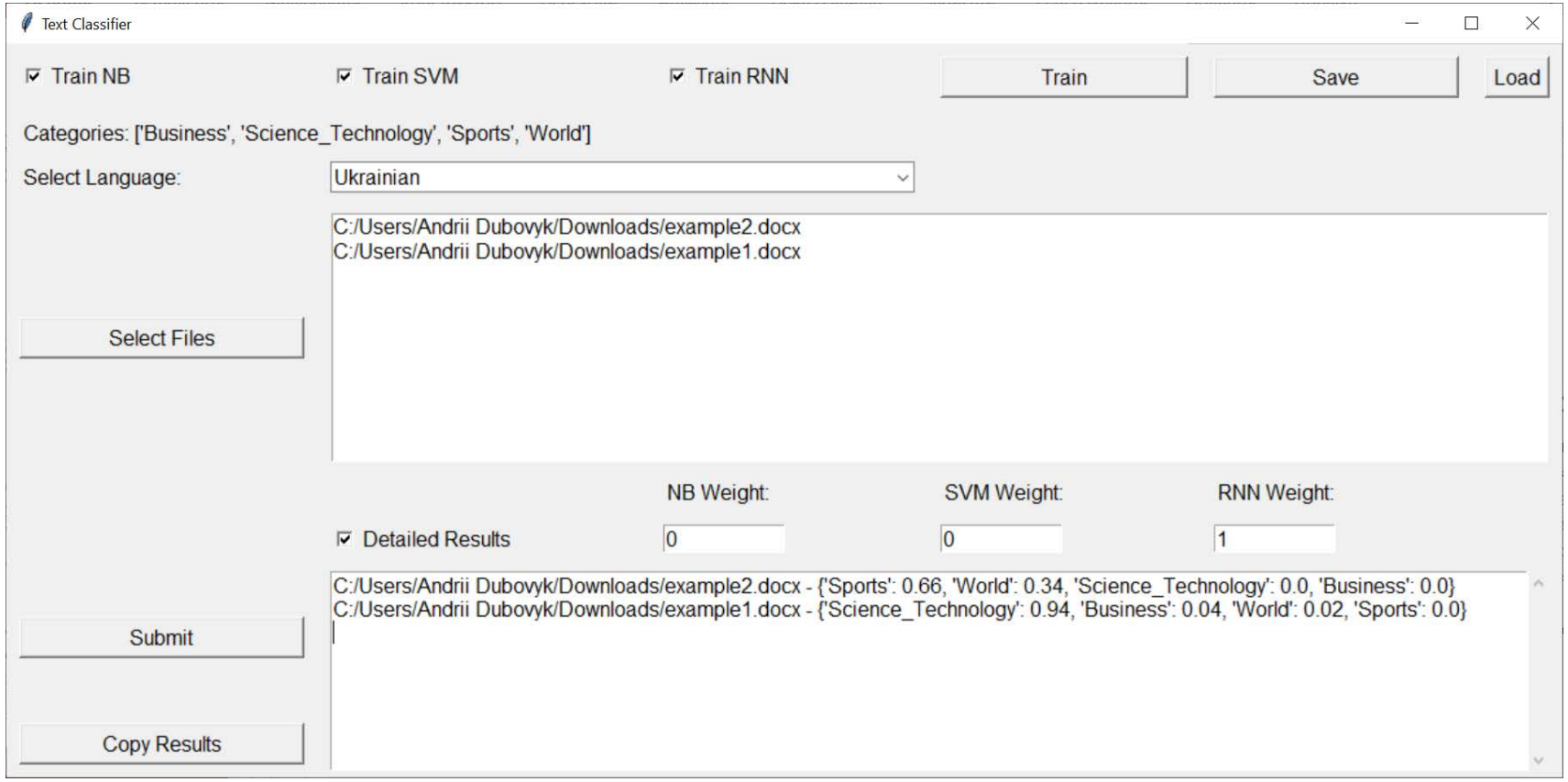
Обробка тексту

Створення й тренування трьох моделей:

- Naive Bayes
- SVM
- RNN

Розробка графічного інтерфейсу для використання моделей та їх тренування на даних користувача

Графічний інтерфейс



Графічний інтерфейс

example1.docx

Хоч найранішу модель машинного навчання й представили в 1950-х роках, коли Артур Семюель винайшов програму, що обчислювала шанси на перемогу в шашках для кожної зі сторін, історія машинного навчання сягає десятиліть людського бажання й зусиль досліджувати людські когнітивні процеси.[13] 1949 року канадський психолог Дональд Гебб опублікував книгу «Організація поведінки[en]», в якій він запропонував теоретичну нейронну структуру, утворювану певними взаємодіями нейронів.[14] Геббова модель взаємодії нейронів між собою заклала основу того, як працюють алгоритми ШІ та машинного навчання на рівні вузлів, або штучних нейронів, які комп'ютери використовують для передавання даних.[13] Інші дослідники, які досліджували людські когнітивні системи, також зробили свій внесок до сучасних технологій машинного навчання, серед них логік Волтер Піттс[en] та Воррен Маккалох, які запропонували ранні математичні моделі нейронних мереж для розробки алгоритмів, що імітують процеси людського мислення.[13]

example2.docx

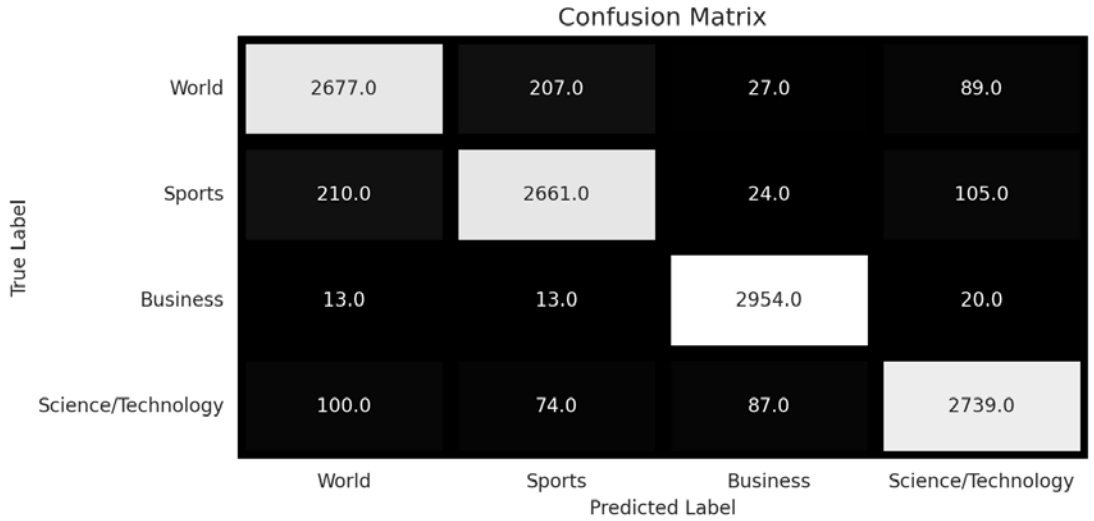
Офіційно першим футбольним матчем в Україні зараз вважається матч у Львові 14 липня 1894 року між командами Львова і Кракова. Матч відбувся у Стрийському парку в рамках Загальної виставки краєвої і тривав 7 хвилин до першого голу, оскільки після матчу мали відбуватися показові виступи спортсменів інших видів спорту. На 6-тій хвилині переможний гол забив спортсмен Львова Володимир Хомицький — учень другого року навчання учительської семінарії. У 2004 році, з нагоди відзначення 110-ї річниці першого матчу в Україні у Стрийському парку Львова відбулося урочисте відкриття пам'ятника українському футболу роботи львівського скульптора Ярослава Скакуна[4].

```
C:/Users/Andrii Dubovyk/Downloads/example2.docx - Sports
C:/Users/Andrii Dubovyk/Downloads/example1.docx - Science_Technology
```

```
C:/Users/Andrii Dubovyk/Downloads/example2.docx - {'Sports': 0.66, 'World': 0.34, 'Science_Technology': 0.0, 'Business': 0.0}
C:/Users/Andrii Dubovyk/Downloads/example1.docx - {'Science_Technology': 0.94, 'Business': 0.04, 'World': 0.02, 'Sports': 0.0}
```

Приклад застосування

	Precision	Recall	F1-Score
World	0.89	0.89	0.89
Sports	0.90	0.89	0.89
Business	0.96	0.98	0.97
Science/Technology	0.93	0.91	0.92



Порівняння результатів

Розроблена RNN модель

11031 / 12000

91.925 %

ChatGPT 3.5

176 / 200

88 %

Інші приклади застосування

	Naive Bayes	SVM	RNN	Комбінація
(10)Dataset Text Document Classification (1000 зразків)	97 %, 0.2 с	98 %, 0.59 с	91 %, 974 с	-
Spam Mails Dataset (5171 зразок)	92.07 %, 0.41 с	99.23 %, 0.54 с	98.07 %, 1153 с	-
Emotions dataset for NLP (16000 зразків)	56.99 %, 0.081с	83.09 %, 0.83 с	85.56 %, 13.35 с	87.75 %, 14.261 с 1 NB, 5 SVM, 13 RNN

Точність класифікації після тренування на різних наборах даних з kaggle.com та час витрачений на тренування

Висновки

Створено систему класифікації текстів

Навчено моделі для класифікації за 4-ма категоріями

Розроблено графічний застосунок, який дозволяє

- використовувати вже навчені моделі
- натренувати моделі на власних текстах для інших потреб

Протестовані результати тренування та класифікації для найбільш поширених задач

Є перспективи для підвищення точності моделей шляхом подальших експериментів, а також для покращення зручності використання графічного інтерфейсу

Дякую за увагу!