ДОДАТОК А

Графічний матеріал кваліфікаційної роботи

Харківський національний університет радіоелектроніки
кафедра ЕОМ

АНАЛІЗ ВПЛИВУ ВИКОРИСТАННЯ КОНТЕКСТУАЛЬНИХ ЕМБЕДИНГІВ НА ТОЧНІСТЬ КЛАСИФІКАЦІЇ ТЕКСТУ

Кваліфікаційна робота
Другий рівень (магістр)

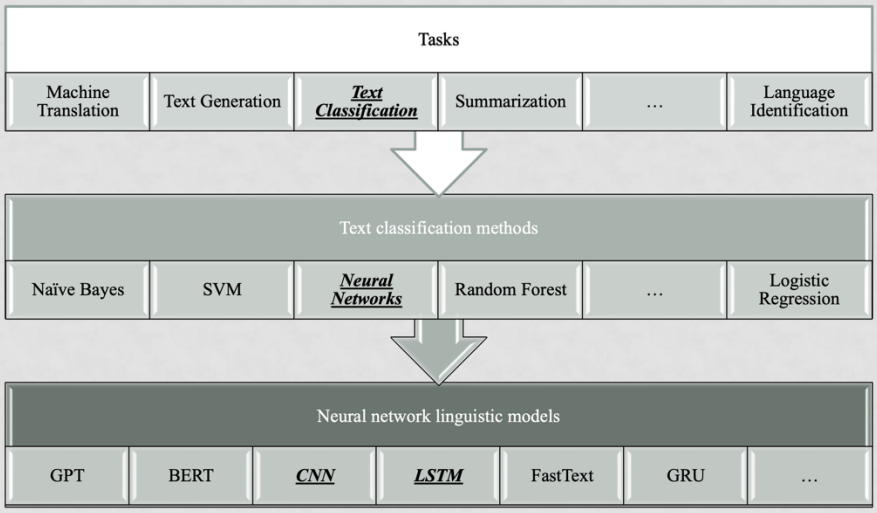
Автор
Воропаєва К.А.
ст. гр. КСМм-22-2

Керівник
Барковська О.Ю.
доц. каф. ЕОМ

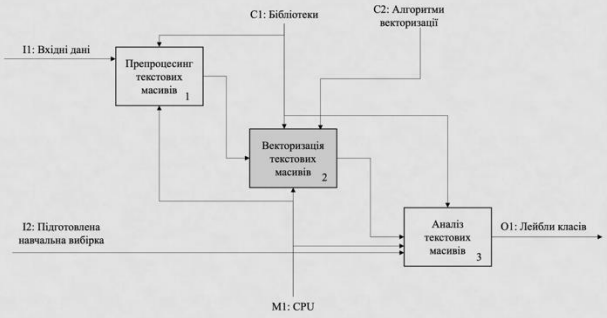
АКТУАЛЬНІСТЬ КВАЛІФІКАЦІЙНОЇ РОБОТИ

- Зростання обсягів текстової інформації в цифровому вигляді.
- Вимоги до швидкості та ефективності обробки тексту в реальному часі.
- Важливість класифікації тексту для підвищення ефективності пошукових систем.
- Розширення можливостей застосування в різних галузях, включаючи медицину та юриспруденцію.

ОГЛЯД ПРОБЛЕМНОЇ ОБЛАСТІ ТЕМИ ДОСЛІДЖЕННЯ

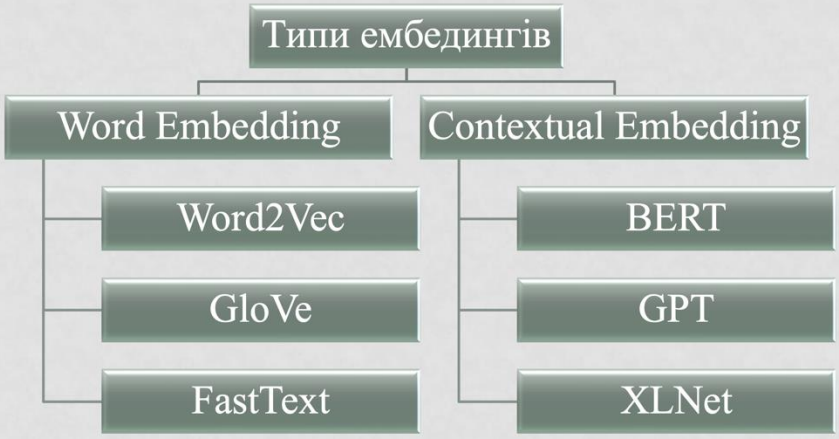


УЗАГАЛЬНЕНИЙ ПІДХІД НЕЙРОМЕРЕЖЕВОЇ КЛАСИФІКАЦІЇ ТЕКСТОВИХ МАСИВІВ



Характеристика	LSTM Model	CNN Model
Основна архітектура	Рекурентна нейронна мережа зі шаром LSTM	Згорткова нейронна мережа зі шарами Conv1D
Типові завдання	Аналіз послідовних даних, текстові дані	Аналіз послідовних даних, зображення
Особливості	Враховує контекст залежностей слів	Визначає локальні шаблони в даних
Алгоритми навчання	Зворотнє поширення помилки, оптимізатор Adam	Зворотнє поширення помилки, оптимізатор Adam
Функції активації	Tanh, Sigmoid	ReLU
Використання пам'яті	Використовує короткотермінову та довготермінову пам'ять	Не використовує пам'ять
Застосування	Послідовний аналіз тексту, мовний переклад	Зображення, відеоаналіз
Бібліотеки реалізації	TensorFlow, Keras	TensorFlow, Keras

РОЗГЛЯНУТІ МЕТОДИ ВЕКТОРИЗАЦІЇ ТЕКСТОВИХ МАСИВІВ ДЛЯ РІЗНИХ ТИПІВ ЕМБЕДИНГІВ



5

МЕТА ТА ЗАДАЧІ КВАЛІФІКАЦІЙНОЇ РОБОТИ

Метою кваліфікаційної роботи є аналіз впливу використання Contextual та Word ембедінгів на точність класифікації текстових масивів.

Задачі кваліфікаційної роботи:

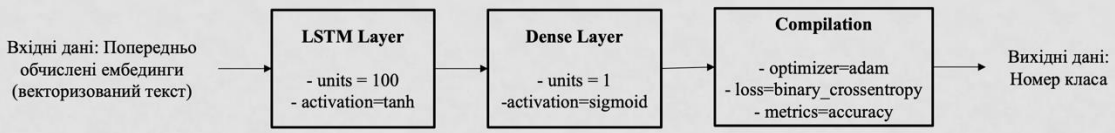
- порівняльний аналіз методів векторизації текстових документів;
- огляд існуючих конвеєрів класифікації тексту, включаючи препроцесінг та безпосередній аналіз;
- дослідження впливу методів векторизації тексту (Word2Vec та GloVe) на точність класифікації текстових масивів в рамках підходу Word Embedding на основі нейромережевої моделі LSTM;
- дослідження впливу методів векторизації тексту (подібні до векторизації в моделях BERT та GPT) на точність класифікації текстових масивів в рамках підходу Contextual Embedding на основі нейромережевої моделі LSTM;
- аналіз точності отриманих результатів.

6

ЗАПРОПОНОВАНИЙ ПОСЛІДОВНІСТЬ РІШЕННЯ ПОСТАВЛЕНОЇ ЗАДАЧІ

7

РОЗРОБЛЕНА АРХІТЕКТУРА LSTM МОДЕЛІ



8

ОТРИМАНІ ПРАКТИЧНІ РЕЗУЛЬТАТИ НА КОЖНОМУ З ЕТАПІВ РОБОТИ СИСТЕМИ

1. Результати передобробки

	review	sentiment
2132	I, too, found "Oppenheimer" to be a brilliant series and one of the finest offerings ever on American PBS. David Suchet was particularly effective as Edward Teller, as I recall, and the overall conception was spectacularly good. The only reason that the series doesn't rate a full 10/10 is for the low-budget production values in some areas. Actual content is absolutely first-rate in my recollection. The Oppenheimer miniseries will be released in the UK on July 31st! It will be a Region 2/PAL set, but it would seem that a Region 1/NTSC set should be soon in the offing. If you have a universal player in the US, you can order the series right now from Amazon UK. http://tinyurl.com/znyyq Huzzah!!	positive

	review	sentiment
2132	i too found oppenheimer to be a brilliant series and one of the finest offerings ever on american pbs david suchet was particularly effective as edward teller as i recall and the overall conception was spectacularly good the only reason that the series doesnt rate a full 1010 is for the lowbudget production values in some areas actual content is absolutely firstrate in my recollectionthe oppenheimer miniseries will be released in the uk on july 31st it will be a region 2pal set but it would seem that a region 1ntsc set should be soon in the offingif you have a universal player in the us you can order the series right now from amazon uk	positive

9

ОТРИМАНІ ПРАКТИЧНІ РЕЗУЛЬТАТИ НА КОЖНОМУ З ЕТАПІВ РОБОТИ СИСТЕМИ

1. Результати нормалізації передобробленого тексту

	review	sentiment
2132	i too found oppenheimer to be a brilliant series and one of the finest offerings ever on american pbs david suchet was particularly effective as edward teller as i recall and the overall conception was spectacularly good the only reason that the series doesnt rate a full 1010 is for the lowbudget production values in some areas actual content is absolutely firstrate in my recollectionthe oppenheimer miniseries will be released in the uk on july 31st it will be a region 2pal set but it would seem that a region 1ntsc set should be soon in the offingif you have a universal player in the us you can order the series right now from amazon uk	positive

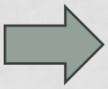
	review	sentiment
2132	found oppenheimer brilliant series one finest offering ever american pb david suchet particularly effective edward teller recall overall conception spectacularly good reason series doesnt rate full lowbudget production value area actual content absolutely firstrate recollectionthe oppenheimer miniseries released uk july region set would seem region set soon offingif universal player u order series right amazon uk	positive

10

ОТРИМАНІ ПРАКТИЧНІ РЕЗУЛЬТАТИ НА КОЖНОМУ З ЕТАПІВ РОБОТИ СИСТЕМИ

1. Результати нормалізації передобробленого тексту

	review	sentiment
2132	i too found oppenheimer to be a brilliant series and one of the finest offerings ever on american pbs david suchet was particularly effective as edward teller as i recall and the overall conception was spectacularly good the only reason that the series doesnt rate a full 1010 is for the lowbudget production values in some areas actual content is absolutely firstrate in my recollectionthe oppenheimer miniseries will be released in the uk on july 31st it will be a region 2pal set but it would seem that a region 1ntsc set should be soon in the offingif you have a universal player in the us you can order the series right now from amazon uk	positive



	review	sentiment
2132	found oppenheimer brilliant series one finest offering ever american pb david suchet particularly effective edward teller recall overall conception spectacularly good reason series doesnt rate full lowbudget production value area actual content absolutely firstrate recollectionthe oppenheimer miniseries released uk july region set would seem region set soon offingif universal player u order series right amazon uk	positive

10

ПЛАНУВАННЯ ЕКСПЕРИМЕНТУ

Номер експерименту	Фактори впливу			Цілі, для яких оцінюються залежності			
	Варіативна кількість епох навчання	Метод векторизації текстових масивів № 1	Метод векторизації текстових масивів № 2	Train accuracy	Validation accuracy	Validation loss	Training time
1	1	+	-	+	+	+	+
2	1	-	+	+	+	+	+
3	2	+	-	+	+	+	+
4	2	-	+	+	+	+	+
…	…	…	…	…	…	…	…
2N-1	N	+	-	+	+	+	+
2N	N	-	+	+	+	+	+

12

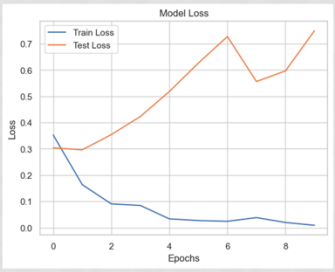
РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ ВПЛИВУ МЕТОДІВ ВЕКТОРИЗАЦІЇ WORD2VEC ТА GLOVE ТА ТОЧНІСТЬ КЛАСИФІКАЦІЇ

Epoch	Word2Vec Accuracy, %	Word2Vec Val Accuracy, %	Word2Vec Val Loss	GloVe Accuracy, %	GloVe Val Accuracy, %	GloVe Val Loss
1	84,94	88,77	0.3037	76,97	82,33	0.4019
2	94,10	88,19	0.2963	83,79	85,41	0.3406
3	96,82	86,80	0.3543	86,53	86,43	0.3227
4	97,02	87,40	0.4230	87,76	87,06	0.3077
5	98,94	87,13	0.5181	88,77	86,61	0.3308
6	99,22	86,47	0.6262	89,54	87,30	0.3056
7	99,26	83,90	0.7275	90,63	87,98	0.3073
8	98,74	85,93	0.5562	91,23	87,73	0.3033
9	99,37	86,55	0.5966	92,12	86,86	0.3273
10	**99,75**	**88,77**	**0.7497**	**93,17**	**87,72**	**0.3166**
15	99,75	86,03	0.8173	93,00	87,21	0.3392

13

АНАЛІЗ ОТРИМАНИХ РЕЗУЛЬТАТІВ ДОСЛІЛЖЕННЯ МЕТОДІВ WORD EMBEDDING

Embedding Type	Word2Vec	GloVe
Number of Epochs	10	**10**
Test Accuracy, %	86,87	**87,72**
Test Loss	0.7497	**0.3166**
ROC-AUC Score	0.9333	**0.9457**
Precision (avg)	0.87	**0.88**
Recall (avg)	0.87	**0.88**
F1-Score (avg)	0.87	**0.88**
Training Time, sec	6340	**3830**



Тенденції втрати при навчанні та тестуванні моделі Word2Vec

Тенденції втрати при навчанні та тестуванні моделі GloVe

14

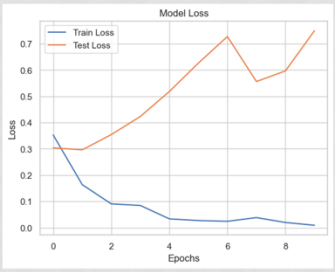
РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ ВПЛИВУ МЕТОДІВ ВЕКТОРИЗАЦІЇ BERT ТА GPT ТА ТОЧНІСТЬ КЛАСИФІКАЦІЇ

Epoch	BERT Accuracy, %	BERT Val Accuracy, %	BERT Val Loss	GPT Accuracy, %	GPT Val Accuracy, %	GPT Val Loss
1	87,62	90,14	0.2601	85,50	88,92	0.2987
2	91,53	90,98	0.2427	88,04	89,27	0.2789
3	93,27	91,45	0.2285	89,31	89,84	0.2623
4	94,52	91,82	0.2204	90,28	90,29	0.2507
5	95,58	92,07	0.2153	91,12	90,65	0.2421
6	96,41	92,31	0.2135	91,76	90,97	0.2365
7	97,12	92,56	0.2141	92,39	91,20	0.2320
8	97,65	92,79	0.2168	92,87	91,43	0.2295
9	**98,14**	**92,97**	**0.2212**	**93,34**	**91,65**	**0.2281**
10	98,91	91,23	0.2919	97,79	89,45	0.3851

15

АНАЛІЗ ОТРИМАНИХ РЕЗУЛЬТАТІВ ДОСЛІЛЖЕННЯ МЕТОДІВ WORD EMBEDDING

Embedding Type	Word2Vec	GloVe
Number of Epochs	10	**10**
Test Accuracy, %	86,87	**87,72**
Test Loss	0.7497	**0.3166**
ROC-AUC Score	0.9333	**0.9457**
Precision (avg)	0.87	**0.88**
Recall (avg)	0.87	**0.88**
F1-Score (avg)	0.87	**0.88**
Training Time, sec	6340	**3830**



Тенденції втрати при навчанні та тестуванні моделі Word2Vec

Тенденції втрати при навчанні та тестуванні моделі GloVe

14

ВИСНОВКИ

- В процесі роботи над кваліфікаційною роботою було проведено дослідження різних підходів до векторизації текстових масивів, в тому числі аналізу Word2Vec та GloVe для Word Embedding, а також BERT і GPT для Contextual Embedding, з метою оцінки їх впливу на точність класифікації текстових масивів.
- На підставі проведеного дослідження було виявлено, що контекстуальні ембедінги, які використовуються в моделях на зразок BERT та GPT, забезпечують вищу точність класифікації порівняно з традиційними методами Word Embedding, такими як Word2Vec і GloVe, особливо в задачах, де важливий глибокий контекстуальний аналіз.
- Було проведено зіставлення результатів між найкращими моделями Word Embedding та Contextual Embedding. Найкращою моделлю Word Embedding, на основі наших критеріїв оцінювання, виявилась GloVe, яка продемонструвала кінцеву точність 87.72%, що було вище, ніж у моделі Word2Vec з точністю 86.87%. У контексті Contextual Embedding, BERT виявився ефективнішим порівняно з GPT, з кінцевою точністю 91.23% проти 89.45% в GPT

17

АПРОБАЦІЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

Стаття у фаховому виданні:

O. BARKOVSKA, K. VOROPAIEVA, O. RUSKIKH Justifying the selection of a neural network linguistic classifier // Innovative technologies and scientific solutions for industries. – 2023. – №. 3 (25). – pp. 5-14.

18