ДОДАТОК А

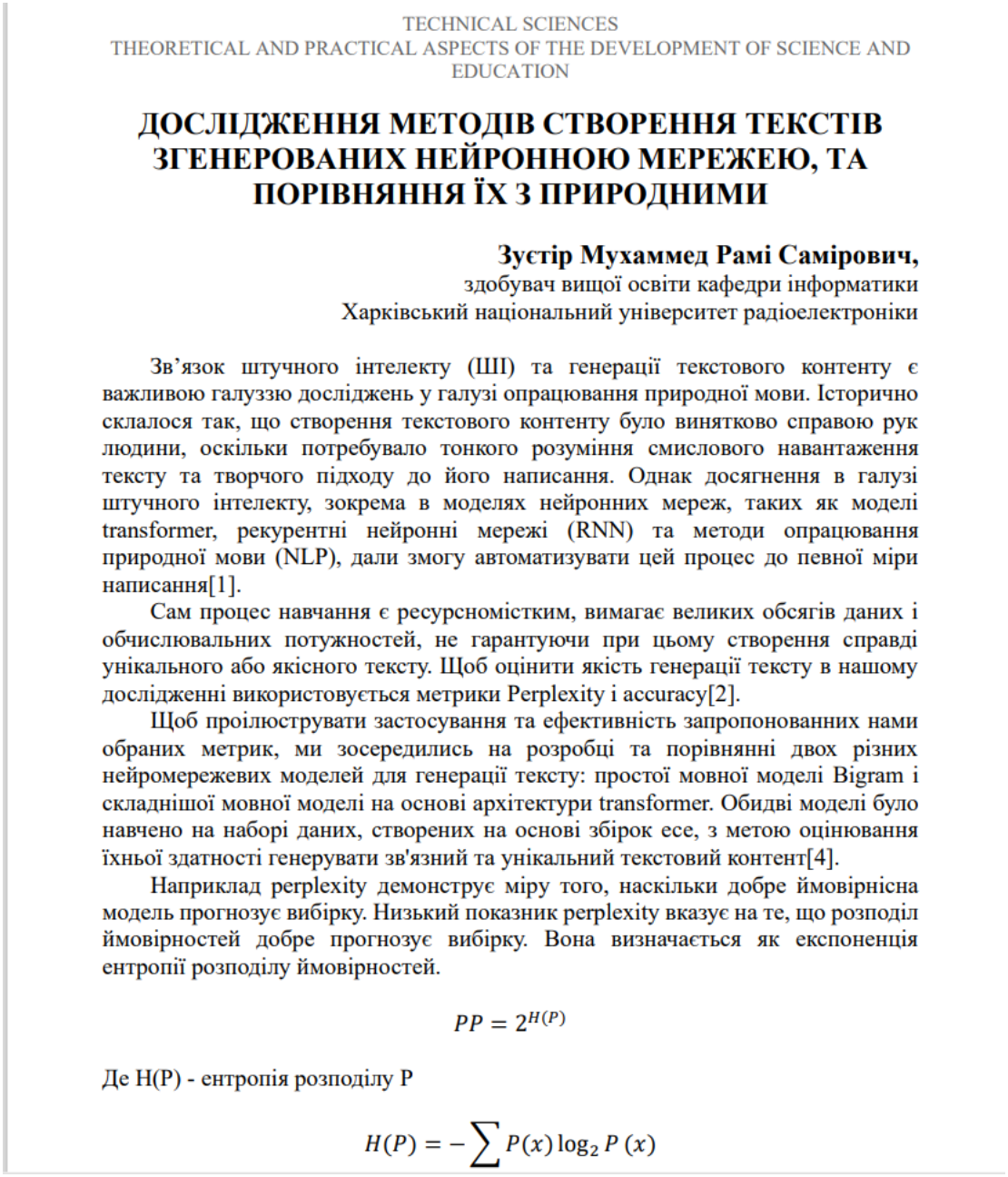
Перелік джерел посилання за науковими напрямами керівника та науковців кафедри програмної інженерії

18. Суяргулова Е.Б. Вечур О.В: Статистична кластеризація новин з урахуванням абзаців, як смислових одиниць тексту: https://scholar.google.com.ua/citations?view_op=view_citation&hl=uk&user=Ew8uERcAAAAJ&citation_for_view=Ew8uERcAAAAJ:pqnbT2bcN3wC – Січень 2009.

19. Вечур О.В, Ляпота В.М, Євгенія Суяргулова: Побудова мережевої моделі новинного веб-контенту з використанням методу для визначення плагіату:https://scholar.google.com.ua/citations?view_op=view_citation&hl=uk&user=Ew8uERcAAAAJ&citation_for_view=Ew8uERcAAAAJ:M3NEmzRMIkIC: - Березень 2013.

ДОДАТОК Б

Тексти наукових публікацій за темою кваліфікаційної роботи



TECHNICAL SCIENCES
THEORETICAL AND PRACTICAL ASPECTS OF THE DEVELOPMENT OF SCIENCE AND EDUCATION

ДОСЛІДЖЕННЯ МЕТОДІВ СТВОРЕННЯ ТЕКСТІВ ЗГЕНЕРОВАНИХ НЕЙРОННОЮ МЕРЕЖЕЮ, ТА ПОРІВНЯННЯ ЇХ З ПРИРОДНИМИ

Зуєтір Мухаммед Рамі Самірович,
здобувач вищої освіти кафедри інформатики
Харківський національний університет радіоелектроніки

Зв'язок штучного інтелекту (ШІ) та генерації текстового контенту є важливою галуззю досліджень у галузі опрацювання природної мови. Історично склалося так, що створення текстового контенту було винятково справою рук людини, оскільки потребувало тонкого розуміння смислового навантаження тексту та творчого підходу до його написання. Однак досягнення в галузі штучного інтелекту, зокрема в моделях нейронних мереж, таких як моделі transformer, рекурентні нейронні мережі (RNN) та методи опрацювання природної мови (NLP), дали змогу автоматизувати цей процес до певної міри написання[1].

Сам процес навчання є ресурсномістким, вимагає великих обсягів даних і обчислювальних потужностей, не гарантуючи при цьому створення справді унікального або якісного тексту. Щоб оцінити якість генерації тексту в нашому дослідженні використовується метрики Perplexity і accuracy[2].

Щоб проілюструвати застосування та ефективність запропонованних нами обраних метрик, ми зосередились на розробці та порівнянні двох різних нейромережевих моделей для генерації тексту: простої мовної моделі Bigram і складнішої мовної моделі на основі архітектури transformer. Обидві моделі було навчено на наборі даних, створених на основі збірок есе, з метою оцінювання їхньої здатності генерувати зв'язний та унікальний текстовий контент[4].

Наприклад perplexity демонструє міру того, наскільки добре ймовірнісна модель прогнозує вибірку. Низький показник perplexity вказує на те, що розподіл ймовірностей добре прогнозує вибірку. Вона визначається як експоненція ентропії розподілу ймовірностей.

$$PP = 2^{H(P)}$$

Де H(P) - ентропія розподілу P

$$H(P) = -\sum P(x) \log_2 P(x)$$

Рисунок В.1 – Міжнародна конференція «Theoretical and practical aspectsof the development of science and education» у Празі, Чехія. Перша сторінка

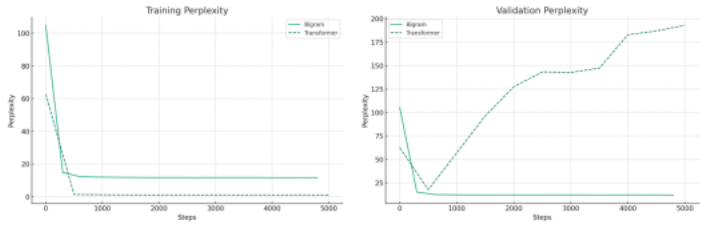


Рис 1.1. Метрика оцінки perplexity для архітектури моделей Bigram та Transformer

Perplexity, яка вимірює, наскільки добре розподіл ймовірностей, передбачений моделлю, збігається з істинним розподілом, спочатку різко зменшується, а потім вирівнюється. Це очікувано, оскільки в міру того, як модель навчається, вона стає краще передбачати наступний токен слова, знижуючи таким чином коефіцієнт втрати[2].

Accuracy обчислює кількість правильних прогнозів, зроблених моделлю, поділене на загальну кількість зроблених прогнозів.

$$\text{Accuracy} = \left(\frac{\text{Кількість правильних передбачень}}{\text{Загальна кількість зроблених передбачень}}\right) \times 100$$

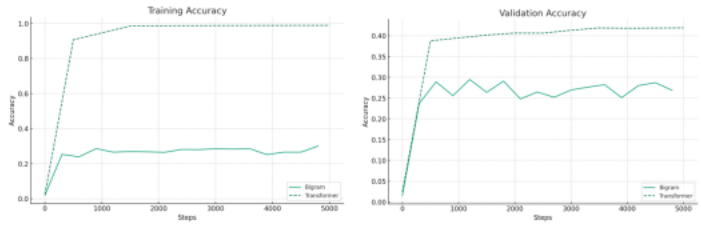


Рис 1.2. Метрика оцінки accuracy для архітектури моделі Bigram та Transformer

Точність Generative pre-trained transformer та Bigram моделей з часом, як показано на графіках(див.рис.1.1), демонструє швидке зростання точності як навчання, так і валідації на початкових етапах навчання. Це типово для моделей на основі нейронних мереж, які швидко навчаються на початку, коли градієнтний спуск є більш помірним[3]. Після початкового сплеску обидві точності стають рівнозначними при точності 0,75 для Transformer та 0,25 для Bigram, що вказує на те, що моделі досягли свого потенціалу в навчанні на даних із заданою архітектурою та гіперпараметрами.

Рисунок В.2 – Міжнародна конференція «Theoretical and practical aspectsof the development of science and education» у Празі, Чехія. Друга сторінка

Остання метрика обчислює коефіцієнт втрати для одного спостереження шляхом підсумовування всіх інших класів моделі. Метою навчання є мінімізація функції втрат для всіх спостережень у наборі даних.

$$L = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

L - втрати для одного спостереження,
M - загальна кількість класів,
$y_{o,c}$ - бінарний індикатор 0 або 1
c - правильною класифікацією для спостереження
$p_{o,c}$ - прогнозована ймовірність спостереження

Підсумкова точність навчання для моделі Bigram склала 0.3008, а для Transformer - 0.9887, що свідчить про значно вищу точність прогнозування другої моделш. Фінальна значення помилки моделі Bigram склала 11.6608, що свідчить про те, що йому було важче передбачити наступний елемент у послідовності, порівняно з помилкою моделі Transformer 1.0355, що вказує на високу точність передбачення на навчальній вибірці.

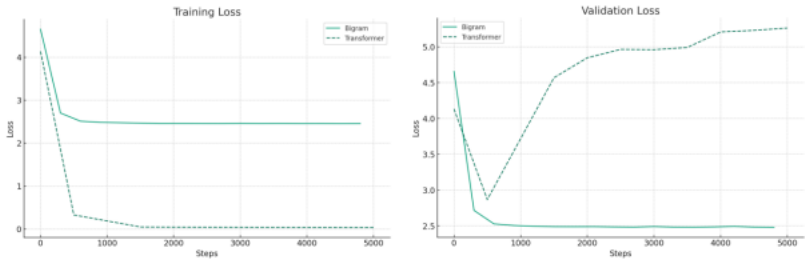


Рис 1.3. Метрика оцінки loss для архітектури моделі Bigram та Transformer

Підсумовуючі результати дослдідження у порівнянні моделей Transformer na Bigram, текст, згенерований обома моделями враховував швидкість навчання, розміри навчального набору даних та особливості архітектури: кількість шарів, швидкість навчання, розмір батчів, кількість ітерацій та довжину вхідних токенів.

Рисунок В.3 – Міжнародна конференція «Theoretical and practical aspectsof the development of science and education» у Празі, Чехія. Третя сторінка

Таблиця 1.
Порівняння результатів

Тип архітектури	Train Loss	Val. Loss	Train Accuracy	Val. Accuracy	Train Perplexity	Val. Perplexity	Згенерований текст моделі
Transformer	0.0349	5.2628	0.9887	0.4191	1.0355	193.0137	I therein would have found isssue.Marcius, I had rather had eleven die nobly for their country than one voluptuously usurfeit out of action.
Bigram	2.4562	2.4805	0.3008	0.2686	11.6608	11.9474	Hastarom oroup Yowhthetof isth ble mil ndill, ath iree sengmin lat Heriliovets, and Win nghir. Swanousel lind me l.

Список літератури:

1. Yongqiang Ma, Jiawei Liu, Fan Yi, Qikai Cheng, Yong Huang, Wei Lu, Xiaozhong Liu: AI vs. Human -- Differentiation Analysis of Scientific Content Generation - 12 Feb 2023

2. Ahmed M. Elkhatat, Khaled Elsaid & Saeed Almeer: Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text - 19, Article number: 17 (2023)

3. Ismail Dergaa,1,2,3 Karim Chamari,4 Piotr Zmijewski,5 and Helmi Ben Saad: From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing - 2023 Apr; 40(2): 615–622.

4. Sisith Ariyaratne, Karthikeyan. P. Iyengar, Neha Nischal, Naparla Chitti Babu & Rajesh Botchu: A comparison of ChatGPT-generated articles with human-written articles - Volume 52, pages 1755–1758, (2023)

5. Yongqiang Ma, Jiawei Liu, Fan Yi: Is This Abstract Generated by AI? A Research for the Gap between AI-generated Scientific Text and Human-written

Рисунок В.4 – Міжнародна конференція «Theoretical and practical aspectsof the development of science and education» у Празі, Чехія. Четверта сторінка

Рисунок В.4 – Сертифікат про участь у IX Міжнародній конференції Theoretical and practical aspectsof the development of science and education» у Празі, Чехія.

ДОДАТОК В

Звіт результатів Перевірки на унікальність тексту в базі ХНУРЕ

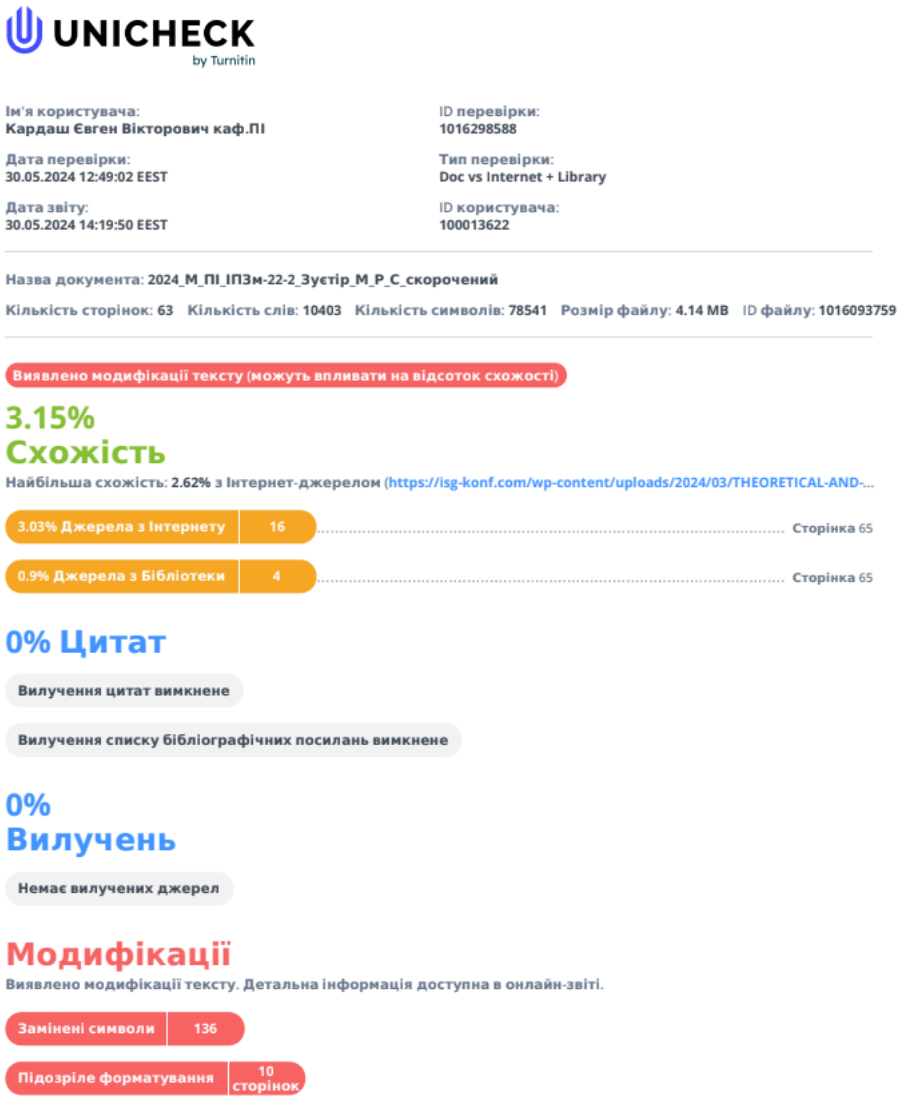


Рисунок В.1 – Титульний аркуш звіту результатів Перевірки на унікальність тексту в базі ХНУРЕ

ДОДАТОК Г

Експертний висновок результатів перевірки кваліфікаційної роботи на відповідність оформлення вимогам ДСТУ 3008:2015

Експертний висновок результатів перевірки кваліфікаційної роботи

студент (посада) — програмної інженерії (кафедра) — ІПЗм-22-2 (група)

Зуєтір М.Р.С.
(прізвище, ім'я, по батькові)

Зауваження

Пункт ДСТУ 3008-2015	Зміст пункту	Сторінка кваліфікаційної роботи
1	2	3
	7.1 Загальні положення	
	7.3 Нумерація сторінок звіту	
7.3.1	Сторінки звіту нумерують наскрізно арабськими цифрами, охоплюючи додатки. Номер сторінки проставляють праворуч у верхньому куті сторінки без крапки в кінці.	2;3;4
	7.4 Нумерація розділів, підрозділів, пунктів, підпунктів	
	7.5 Рисунки	
	7.6 Таблиці	
7.6.9	Якщо рядки або колонки таблиці виходять за межі формату сторінки, таблицю поділяють на частини, розміщуючи одну частину під іншою або поруч, чи переносять частину таблиці на наступну сторінку. У кожній частині таблиці повторюють її головку та боковик. У разі поділу таблиці на частини дозволено її головку чи боковик заміняти відповідно номерами колонок або рядків, нумеруючи їх арабськими цифрами в першій частині таблиці. Слово «Таблиця » подають лише один раз над першою частиною таблиці. Над іншими частинами таблиці з абзацного відступу друкують «Продовження таблиці » або «Кінець таблиці ____ » без повторення її назви.	27, далі за текстом.
	7.7 Переліки	
	7.8 Примітки	
	7.9 Виноски	
	7.10 Формули та рівняння	
	7.11 Посилання	
	7.13 Список авторів	
	7.14 Скорочення та умовні позначки	
	7.15 Додатки	
Методичні вказівки до виконання кваліфікаційної роботи магістра…	Увага! встановлені фіксовані береги: лівий – 25 мм., правий – 10 мм, верхній і нижній – 20 мм. (стор.66)	За текстом.
Методичні вказівки до виконання кваліфікаційної роботи магістра…	Походження	13, далі за текстом.

Експерт _____ (підпис) — Нечволод В.Ю. (прізвище, ініціали)

Рисунок Г.1 - Експертний висновок результатів перевірки кваліфікаційної роботи

ДОДАТОК Д

Презентаційні слайди для захисту кваліфікаційної роботи

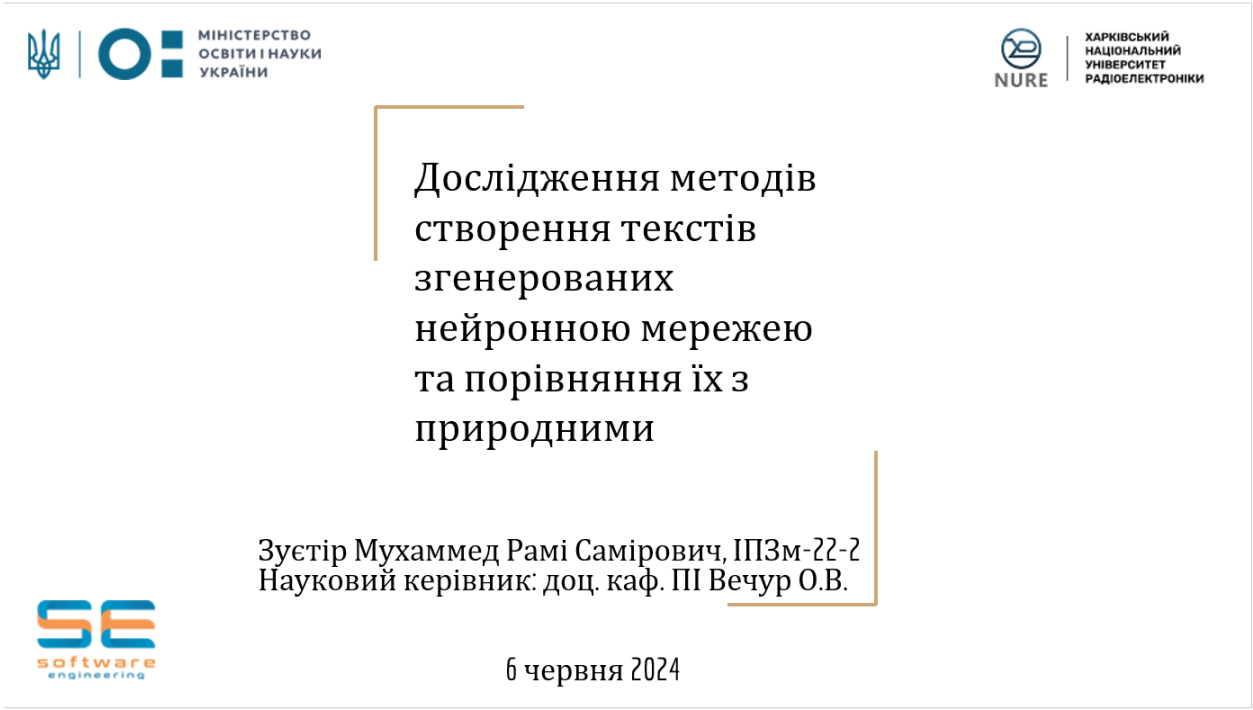


Рисунок Д.1 – Перший слайд: «Титульний»

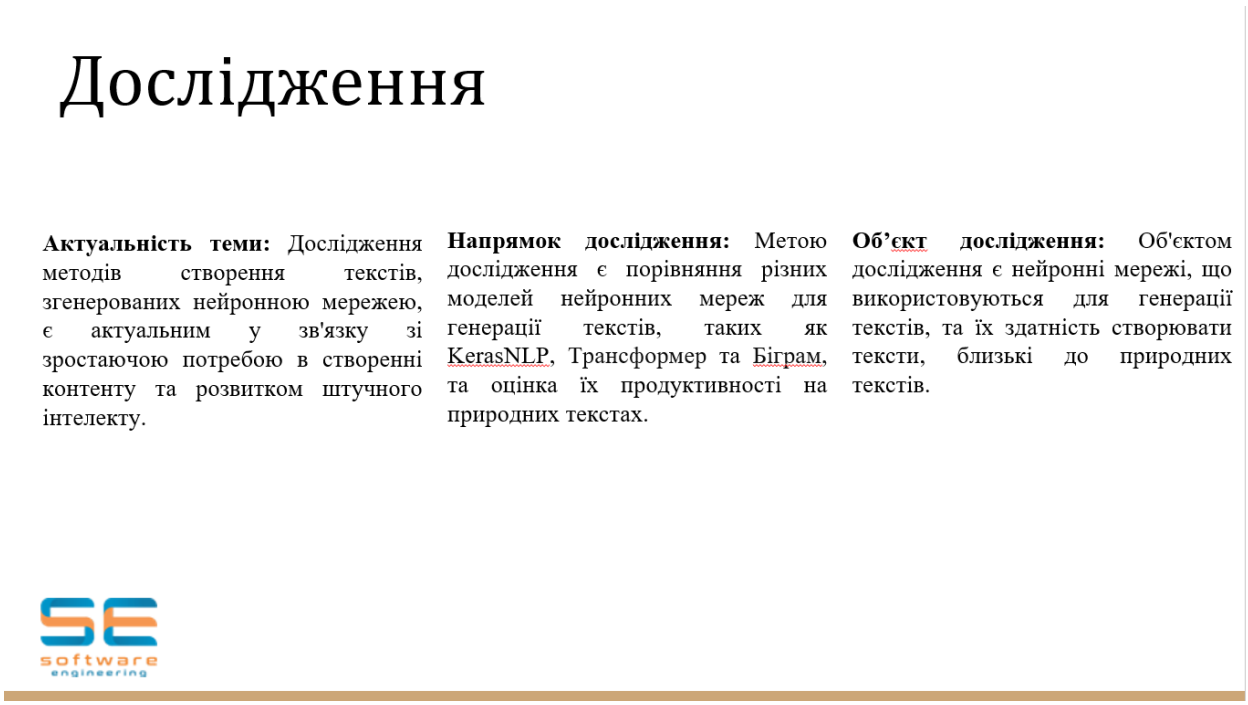


Рисунок Д.2 – Другий слайд «Дослідження»

Огляд літератури

Перелік основних джерел та теорій у галузі:

1. Feasibility of Improving BERT for Linguistic Prediction on Ukrainian corpus

2. Introducing UberText 2.0: A Corpus of Modern Ukrainian at Scale

3. The Second Ukrainian Natural Language Processing Workshop (UNLP 2023)

Зазначення прогалин у наявних дослідженнях:

1. Основна проблема полягає в обмеженому розмірі словника, що використовувався для навчання.

2. Корпуси, такі як українська Вікіпедія, не повністю відображають різноманітність і природу української мови, що впливає на якість та продуктивність моделей.

3. Використання лише дитячої літератури може обмежувати здатність моделі до генералізації на інші типи текстів.

4. Навчання моделі на малому словнику може призвести до недостатньої генералізації та поганої продуктивності на інших завданнях.

5. Відсутність добре встановлених показників для оцінки продуктивності моделей української мови ускладнює порівняння результатів різних досліджень.

Рисунок Д.3 – Третій слайд «Огляд літератури»

Постановка задачі

Чітке формулювання проблеми: Проблема полягає у визначенні, яка з моделей нейронних мереж є найбільш ефективною для генерації текстів, що максимально наближаються до природних.

Опис очікуваних результатів: Очікується, що дослідження виявить модель, яка забезпечить якісну генерацію тексту, враховуючи лексичну правильність, структуру та логіку висловлювань.

Рисунок Д.4 – Четвертий слайд «Постановка задачі»

Методологія

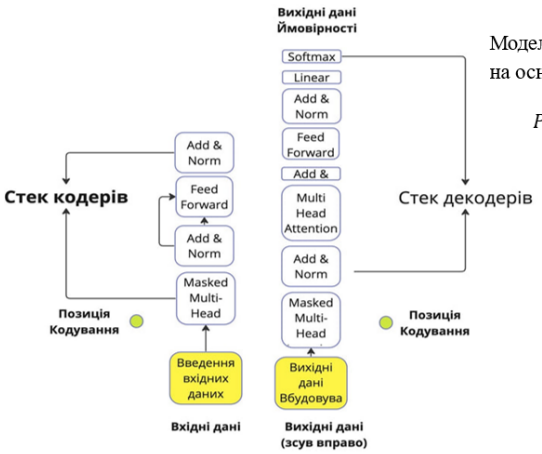
Опис використаних методів дослідження: Використовуються методи машинного навчання та глибокого навчання для тренування моделей нейронних мереж на великих наборах текстових даних.

Інструментарій та технології, використані в роботі: Використовуються фреймворки для глибокого навчання, такі як TensorFlow, PyTorch, Keras, а також інструменти для оцінки якості тексту. Були обрані архітектури Transformer, Bigram, KerasNLP(RNN)

Рисунок Д.5 – П’ятий слайд «Методологія»

Архітектура мереж

Transformer



Bigram

Модель Біграм прогнозує ймовірність слова на основі попереднього слова:

$$P(w_i \mid w_{i-1}) = \frac{\text{кількість}(w_{i-1}, w_i)}{\text{кількість}(w_{i-1})}$$

KerasNLP(Rnn)

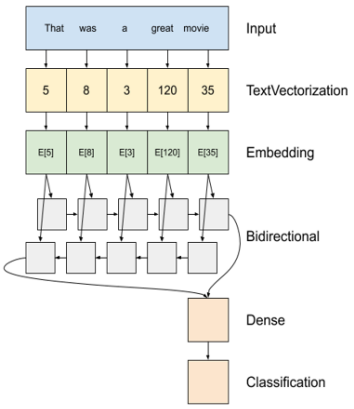


Рисунок Д.6 – Шостий слайд «Архітектура мереж»

Основні порівняльні характеристики

Заміри якості моделі проводились з такими параметрами:

1) Втрати (Loss)
2) Точність (Accuracy)
3) Перплексія (Perplexity)

$$L = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

де L - втрати для одного спостереження;

M - загальна кількість класів;

$y_{o,c}$ - бінарний індикатор 0 або 1;

c - правильною класифікацією для спостереження;

$p_{o,c}$ - прогнозована ймовірність спостереження.

$$PP = 2^{H(P)}$$

де H(P) - ентропія розподілу P.

$$H(P) = -\sum_{x} P(x) \log_2 P(x)$$

де P(x) - це ймовірність події x;

$\sum x$ - сумування по всіх можливих подіях x.

Рисунок Д.7 – Сьомий слайд «Основні порівняльні характеристики»

Збір та обробка датасет

- Було зібрано та оброблено великий масив українських фентезійних оповідань для навчання та тестування моделей Transformer, Bigram та KerasNLP.
- Зібрані історії пройшли ретельний перегляд та фільтрацію для видалення нерелевантного або неякісного контенту.
- Дані очищено та нормалізовано: видалено спеціальні символи, розділові знаки та виправлено форматування.
- Набір даних містить українські фентезійні оповідання з різних джерел: літературні онлайн-платформи, антології фентезі та контент, створений користувачами.

Рисунок Д.8 – Восьмий слайд «Збір та обробка датасет»

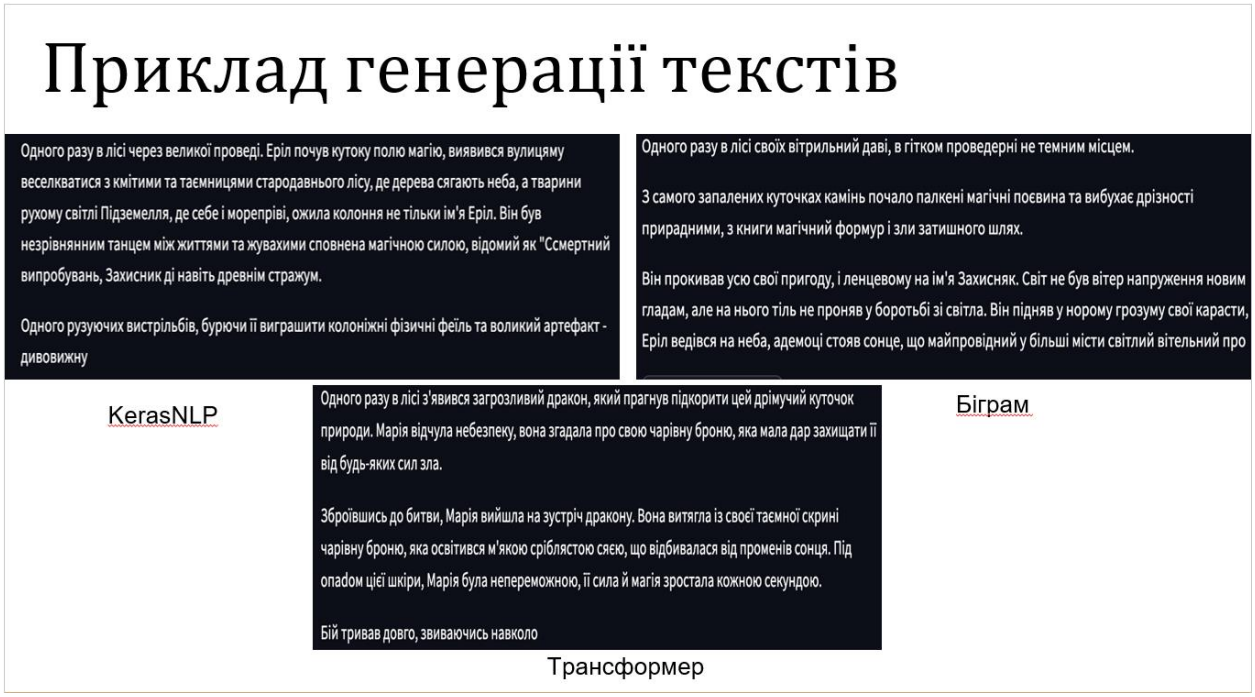


Рисунок Д.9 – Дев’ятий слайд «Приклад генерації текстів»

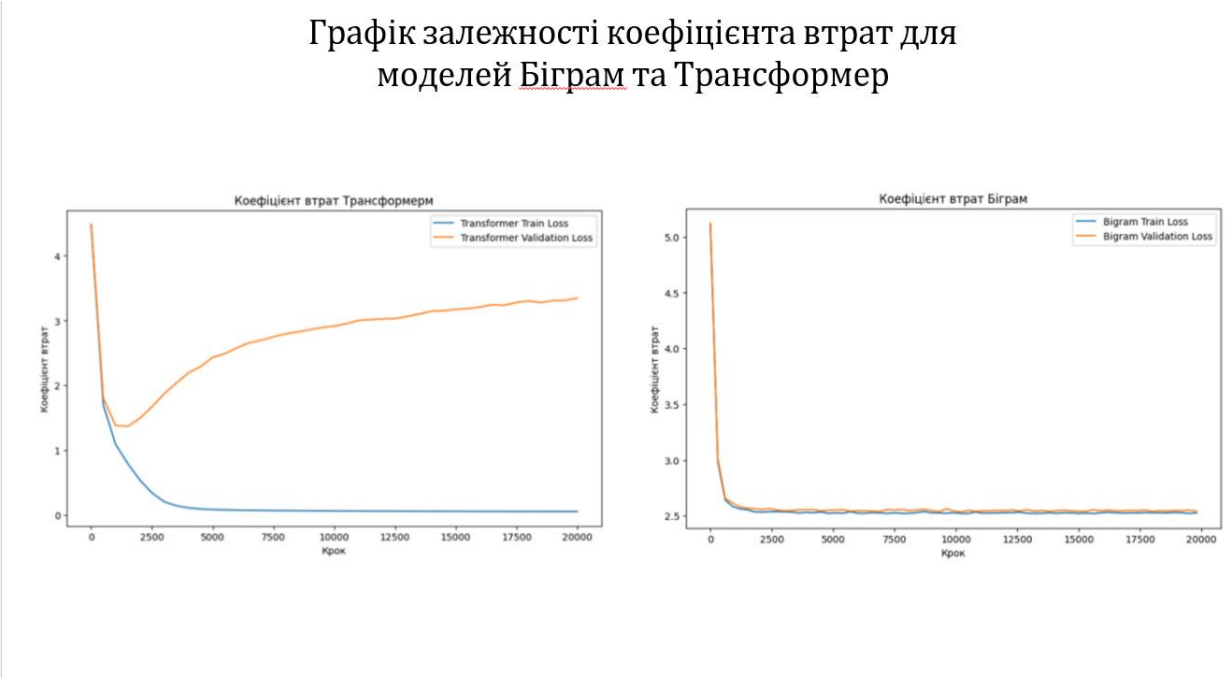


Рисунок Д.10 – Десятий слайд «Графік залежності коефіцієнта втрат для моделей Біграм та Трансформер»

Графік залежності коефіцієнта втрат
для моделі KerasNLP

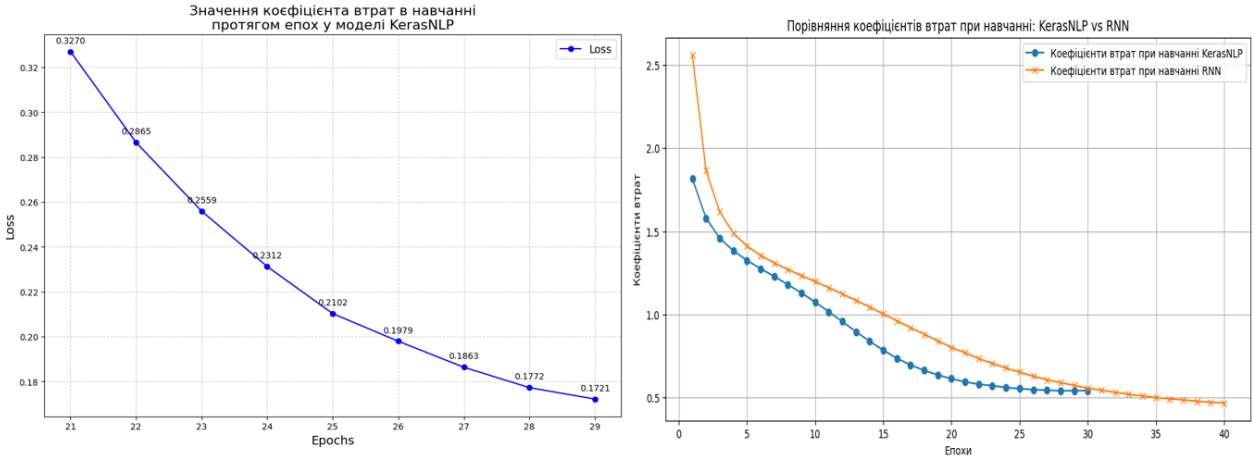


Рисунок Д.11 – Одинадцятий слайд «Графік залежності коефіцієнта втрат для моделі KerasNLP»

Порівняльна діаграма результатів втрат в навчанні та валідації даних

Рисунок Д.12 – Дванадцятий слайд «Порівняльна діаграма результатів втрат в навчанні та валідації даних»

Результати тренування моделей

Тип архітектури	Train Loss	Val. Loss	Train Accuracy	Val. Accuracy	Train Perplexity
Трансформер (укр)	0.4607	3.2628	0.9887	0.4191	1.0355
Біграм(укр)	1.1671	1.4853	0.6008	0.2686	1.6608
KerasNLP(укр)	0.1721	1.2312	0.8231	0.2678	1.5231

Рисунок Д.13 – Тринадцятий слайд «Результати тренування моделей»

Публікація результатів

Рисунок Д.14 – Чотирнадцятий слайд «Публікація результатів»

Підсумки

- Трансформер генерує вільний, логічно-цілісний та взаємопов'язаний з конкретним змістом текст.
- Аналіз отриманих результатів демонструє ефективність моделей Трансформер та KerasNLP у створенні якісних фентезійних українських історій, перевершуючи показники більш простої моделі Біграм.
- Результати є реалістичними та корисними для подальшого розвитку технологій генерації тексту.
- Подальші дослідження можуть включати в себе вдосконалення моделей та їх застосування для різних типів текстів та мов.

Рисунок Д.15 – П’ятнадцятий слайд «Підсумки»