https://doi.org/10.31649/1997-9266-2025-180-3-133-139

УДК 004.4

М. О. Литвин[1]
Л. М. Олещенко[1]

АВТОМАТИЗОВАНИЙ ПІДХІД ДО ДАТУВАННЯ АНГЛОМОВНОГО ТЕКСТУ З ВИКОРИСТАННЯМ ТРАНСФОРМЕРНИХ НЕЙРОННИХ МЕРЕЖ

[1]Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

Розглянуто наявні методи датування тексту за допомогою нейронних мереж, їхні переваги та недоліки. Датування тексту є актуальною задачею в таких сферах, як історія, архівознавство, лінгвістика та криміналістика, оскільки точне визначення часу створення документа дозволяє підтвердити його достовірність, встановити авторство та виявити підробки. Проте традиційні методи, основані на стилометричних або статистичних підходах, мають обмежену точність, особливо для великих обсягів текстових даних. Авторами запропоновано автоматизований підхід до датування англомовного тексту з використанням трансформерних нейронних мереж, який дозволяє визначати приблизне десятиліття написання фрагмента тексту з точністю до 30 років на рівні 85 % у проміжку XV—XX ст. Це перевищує результати аналогічних методів, що працюють з англомовними текстами. Основна ідея запропонованого підходу полягає у використанні принципів передавального навчання для додатково адаптованої до конкретного завдання та попередньо навченої трансформерної нейронної мережі, оптимізованої для класифікації текстових фрагментів за десятиліттями. Однією з ключових переваг запропонованого підходу є застосування трансформерної архітектури, що завдяки механізму уваги враховує складні зв'язки між частинами тексту. Іншою важливою перевагою є використання передавального навчання, що значно зменшує затрати часу та обчислювальних ресурсів у порівнянні з безпосереднім навчанням моделі. Реалізація запропонованого підходу виконувалася мовою Python з використанням бібліотек "transformers" для навчання та тестування нейронної мережі, "datasets" для роботи з набором даних та "numpy" для обчислень. Результати експериментів продемонстрували високу точність підходу: 86 % з точністю до 30 років та 73 % з точністю до 20 років на тестовому наборі даних. Для XIX та XX століть точність досягала 89 % та 90 % відповідно, тоді як для попередніх століть точність була нижчою і становила близько 30 %. Також у дослідженні розглянуто можливість виділення ознак належності тексту певному періоду, через виділення слів з найбільшим показником уваги. Подальші дослідження спрямовані на підвищення точності для малопредставлених у навчальному наборі періодів шляхом розширення та вдосконалення корпусу даних. Додаткові поліпшення можливі через оптимізацію гіперпараметрів моделі та тестування інших архітектур нейронних мереж. Іншим напрямком подальших досліджень є пошук способів виділення лінгвістичних чи стилістичних ознак належності текстів певному періоду, задля можливості інтерпретації результатів роботи нейронної мережі користувачами. Запропонований підхід може бути використаний у таких сферах, як історичні дослідження, аналіз автентичності документів, виявлення плагіату, літературознавство та криміналістика.

Ключові слова: програмна обробка природного мовлення (NLP), машинне навчання, трансформерні нейронні мережі (TNN), передавальне навчання, BERT, датування тексту, стилометрія, аналіз історичних текстів.

Вступ

Історики, архівісти, лінгвісти та інші фахівці часто потребують датування документів для визначення їхньої автентичності та авторства. Це складне завдання, оскільки тексти можуть не містити чіткої дати, бути частковими або сфальсифікованими. Традиційний підхід передбачає експертний аналіз, що є трудомістким, суб'єктивним і потребує доступу до історичних джерел. Наявні програмні рішення обмежені мовою та часовим проміжком, а багато з них використовують застарілі нейромережеві архітектури. Наявні програмні рішення мають свої обмеження. Кожне рішення

працює лише з певною мовою і на певному часовому проміжку. Також більшість рішень має застарілу архітектуру нейронних мереж і не використовують найпристосованішу трансформерну архітектуру для обробки природної мови.

Метою роботи є дослідження наявних методів датування тексту на основі нейронних мереж та машинного навчання та пропозиція автоматизованого підходу з використанням трансформерної нейронної мережі, який дозволить точніше визначати період написання англомовного тексту ніж наявні аналоги, з точністю до декількох десятиліть. Точність запропонованого підходу протестовано за декількома метриками. Запропонований підхід з покращеною точністю може бути використаним у таких сферах людської діяльності, як: історія, архівна справа, лінгвістика, літературознавство, визначення авторства, визначення достовірності документів, криміналістика тощо.

Результати дослідження

Найближчим аналогом цього є проєкт Ithaca — програми на основі глибоких трансформерних нейронних мереж для відновлення та визначення місця і часу написання давньогрецьких інскрипцій. Її натреновано на публічному наборі даних The Packard Humanities Institute's Searchable Greek Inscriptions, і вона демонструє точність, що перевершує результати експертів [1]. Єдине відоме рішення для англійської мови — NeuralDater. В його основі лежать графові згорткові нейромережі, які менше пристосовані для задач оброблення природної мови (NLP), ніж трансформерні. На відміну від трансформерів, згорткові нейромережі не мають механізму уваги, що ускладнює збереження довготривалих залежностей. NeuralDater натреновано на наборі даних газетних текстів Gigaword corpus, з охопленням періоду 1987—2010 років [2].

Інше дослідження щодо датування середньовічних шведських манускриптів за допомогою згорткових нейромереж, натренованих на фотокопіях манускриптів із Svenskt Diplomatariums huvudkartotek. Варто зазначити, що це рішення працює із зображеннями, а не з текстом, що обмежує його універсальність [3].

До того ж існує публікація про глибоку нейронну мережу прямого поширення, розроблену для датування текстів санскритом. Це складна задача, оскільки точні дати багатьох санскритських текстів залишаються дискусійними. Модель навчено на текстах із Digital Corpus of Sanskrit, і, незважаючи на неоптимальну архітектуру, вона демонструє точність, близьку до експертної [4].

З огляду на зазначені рішення, можна побачити відсутність задовільного інструмента для датування англомовних текстів за допомогою нейромереж. Лише Ithaca використовує сучасну трансформерну архітектуру, але вона орієнтована на давньогрецькі інскрипції. NeuralDater — єдине рішення для англомовних текстів, обмежене часовим періодом і менш ефективною архітектурою.

Рішення для шведських манускриптів працює із зображеннями, що обмежує його застосування. Тим не менш, ці підходи демонструють точність, порівнянну або вищу за експертну. Тому розробка сучасного рішення на основі нейромереж для датування англомовних текстів є актуальною.

Таблиця 1

Порівняння наявних програмних рішень

Назва	Ithaca	NeuralDater	Датування шведських манускриптів	Датування санскриту
Архітектура нейромережі	Transformer	GCNN	CNN	Deep feedforward
Мови тексту	Давньогрецька	Англійська	Шведська, латина	Санскрит
Часовий період	800 д.н.е—800 н.е	1987—2010	817—1540	XV ст. д.н.е—XXI ст. н.е
Розмір датасету	78608	~650000	10992	242
Джерело даних	Інскрипції	Газети	Манускрипти	Релігійні, історичні, художні тексти
Дані, що приймаються	Текст	Текст	Зображення	Текст
Спеціалізованість на датуванні	—	+	+	+
Точність	± 29,3 роки (перевершує експертів ±144,4 роки)	61,5 %, ± 0,87 роки	± 20 років (близько до результатів експертів)	56 %, ± 135,5 років (близько до результатів експертів)

Запропонований підхід використовує трансформерні нейронні мережі (Transformer Neural Network, TNN) та процес передавального навчання (transfer learning). Трансформери — це тип моде-

лей глибокого навчання, представлених у статті «Attention is All You Need» у 2017 році [5]. Вони розроблені для обробки послідовних даних і стали основою багатьох сучасних моделей обробки природної мови (NLP). Трансформерна модель повністю покладається на механізм самоуваги (self-attention), щоб створити глобальні залежності між вхідними та вихідними даними. Це усуває потребу в рекурентних та згорткових шарах, через що вона стає придатнішою до паралелізації та ефективнішою для навчання порівняно з іншими архітектурами, такими як рекурентні та згорткові нейронні мережі. Кожен вхідний елемент (токен) може «звертати увагу» на всі інші елементи в послідовності, дозволяючи моделі розглядати увесь контекст у кожній позиції. Це реалізовано за допомогою масштабованого скалярного добутку.

Оригінальна TNN-модель використовує архітектуру енкодера-декодера, де енкодер обробляє вхідну послідовність, а декодер генерує вихідну послідовність. Кожний з них складається з кількох рівнів мереж уваги та шарів прямого поширення [5].

На рис. 1 показано архітектуру трансформера, який складається з двох основних компонентів: енкодера (зліва) та декодера (справа), кожний з яких містить кілька рівнів (Nx) для обробки послідовних даних. Енкодер приймає вхідну послідовність, яку спочатку перетворюють у векторні подання через вбудовування (embedding).

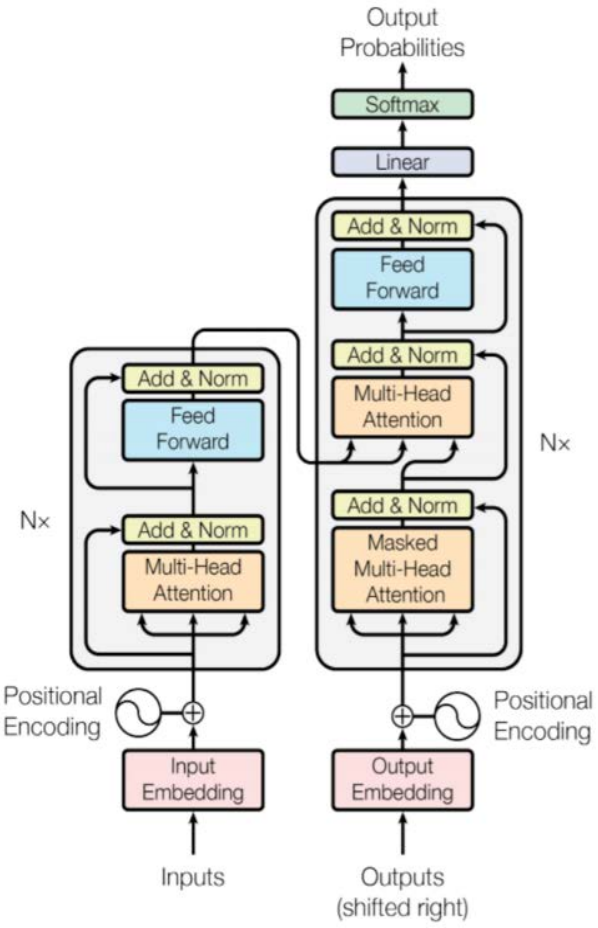

Рис. 1. Трансформерна архітектура нейронної мережі

Потім додається позиційне кодування (positional encoding), що допомагає враховувати порядок слів у послідовності. Далі дані проходять через кілька рівнів механізму Multi-Head Attention, який дозволяє кожному токену взаємодіяти з усіма іншими вхідними токенами. Після цього застосовується нормалізація (Add & Norm) і повнозв'язний шар (Feed Forward) для додаткової обробки. Декодер отримує вихідну послідовність, яку спочатку також перетворюють через вбудовування та позиційне кодування. Перший шар — це маскована мультиувага (Masked Multi-Head Attention), яка обмежує доступ моделі до майбутніх токенів, забезпечуючи автогресивне генерування тексту. Потім модель отримує «увагу» до виходу енкодера, що дозволяє їй використовувати інформацію про вхідні дані. Як і в енкодері, після кожного шару уваги застосовуються нормалізація та Feed Forward. Після проходження через всі рівні декодера вихідна послідовність передається у лінійний шар і Softmax, що генерує ймовірності вибору кожного слова у вихідному тексті. Ця архітектура дозволяє ефективно працювати з довгими текстами, забезпечує паралельну обробку та гнучкість у завданнях обробки природної мови.

Передавальне навчання (transfer learning) — це метод машинного навчання, що передбачає використання вже навченої моделі для нового, але пов'язаного завдання. Коли така модель адаптується до іншого завдання, процес називається тонким налаштуванням (fine-tuning). Це дозволяє використовувати попередньо навчені моделі, що тренувалися на великих наборах даних, щоб підвищити продуктивність і скоротити час навчання для нових задач з обмеженим обсягом даних. Такий підхід широко застосовується у глибокому навчанні, зокрема в комп'ютерному зорі та обробці природної мови, через що він є ефективним для задачі датування текстів. У запропонованому підході використано попередньо навчена трансформерна модель BERT (Bidirectional Encoder Representations from Transformers), яку надалі додатково навчено на специфічних даних для виконання конкретного завдання. Це модель, що навчається контекстному представленню слів, аналізуючи як лівий, так і правий контекст.

Модель BERT попередньо натреновано на великих наборах даних, включно з Toronto Book Corpus і Вікіпедією, для виконання задач передбачення замаскованих токенів (MLM) та передба-

чення наступного речення (NSP) [6]. Головна перевага моделі BERT полягає у формуванні глибоких двонаправлених представлень з непозначеного тексту, що дозволяє застосовувати її до широкого спектра NLP-завдань без значних змін архітектури. Для навчання BERT на задачі датування текстів використано набір даних «Project Gutenberg — English Language eBooks» [7], що містить тексти та метадані 48284 англомовних книг з відкритої бібліотеки Project Gutenberg. Він включає такі поля, як: text (повний текст твору), source (значення gutenberg для всіх книг) і METADATA (відомості про мову, автора, рік видання тощо).

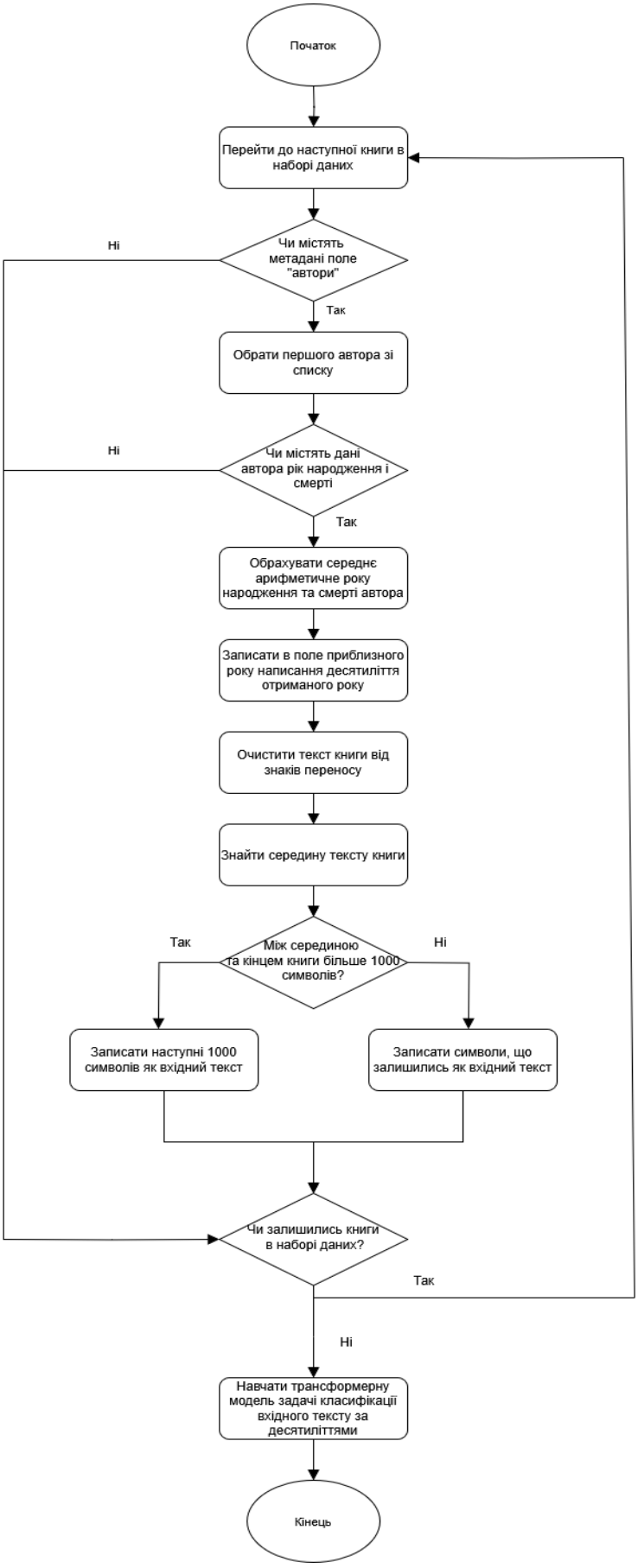

Рис. 2. Блок-схема алгоритму запропонованого підходу

Оскільки точна дата написання текстів у наборі відсутня, необхідно було провести попередню обробку даних. Цей процес включав очищення текстів від символів переносу, вибір фрагментів по 1000 символів із середини кожного твору та визначення приблизного року написання як середнього між роками народження та смерті автора. Дані, де це визначити неможливо, відсіювалися. Далі кожен рік співвідносився з відповідним десятиліттям. Після обробки набір даних розподілено на три підмножини: тренувальну (28481 запис), валідаційну (3553 записів) та тестову (3565 записів). Модель налаштовано для класифікації, де фрагменти по 1000 символів слугували вхідними даними, а десятиліття — класами.

Програмна реалізація запропонованого підходу виконувалася мовою Python з використанням бібліотек “transformers” для навчання та реалізація нейронної мережі, “datasets” для роботи з набором даних та “numpy” для обчислень. Бібліотека “transformers” забезпечує доступ до попередньо навчених моделей, таких як BERT, спрощує їх використання та інтеграцію в різні NLP-завдання.

Блок-схему алгоритму запропонованого підходу показано на рис. 2. Для оцінки точності роботи навченої нейромережі використовувалися такі метрики: точність прогнозування в межах певної кількості років. Цей показник розраховується як відсоток фрагментів тексту, для яких передбачене десятиліття написання, відрізня-

ється від реального не більше ніж на задану кількість років, від загальної кількості фрагментів, для яких зроблено передбачення. Розглянуто показники для відхилень у 30, 20, 10 та 0 років. Додатково аналізувалася точність в межах 30 років у розрізі десятиліть, де цей показник визначався окремо для кожного століття. Після трьох епох навчання нейромережа досягла 85 % точності у визначенні десятиліття написання тексту з відхиленням до 30 років на валідаційному наборі даних. Точність в межах 20 років становила 73 %. Оскільки точність зростала з кожною епохою, процес навчання можна вважати успішним. Найвищі показники спостерігалися для XIX та XX століть, де точність досягала 89 %. Це пояснюється тим, що найбільша кількість текстів у наборі даних належить саме до цих періодів. Точність роботи навченої нейронної мережі перевірили на тестовій частині набору даних, яка містить 3565 записів про книги та виділена на етапі підготовки даних. Для оцінки точності використовувалися ті ж метрики, що й під час навчання нейромережі, описані вище.

Результати тестування показують, що навчена нейронна мережа демонструє високу точність у визначенні приблизної дати написання тексту. Загальна точність визначення десятиліття написання з похибкою до 30 років складає 86 %, що свідчить про ефективність використаного підходу. Зменшення діапазону допустимої похибки спричиняє поступове зниження точності: для допуску 20 років точність становить 73 %, для 10 років — 53 %, а для абсолютно точного передбачення року (0 років похибки) — 21 %. Аналізуючи точність за століттями, можна помітити, що найкращі результати отримані для текстів XIX та XX століть — 89 % та 90 % відповідно. Це пояснюється великою кількістю текстових даних цього періоду в наборі, що сприяло кращому навчанню моделі. У порівнянні, точність визначення текстів XVIII століття значно нижча — лише 29 %, а для XVII та XVI століть показники однакові — по 27 %. Це свідчить про брак даних для цих періодів, що ускладнює коректне навчання нейромережі та її здатність передбачати дату написання текстів ранніх століть. Отримані результати підтверджують, що запропонована модель добре працює з текстами новітнього періоду, проте для підвищення її точності щодо більш ранніх століть необхідно розширити набір навчальних даних, включивши більше текстів XVI—XVIII століть (табл. 2).

Таблиця 2

Порівняння розробленого рішення з існуючими

Назва	Розроблене рішення	Ithaca	NeuralDater	Датування шведських манускриптів
Архітектура нейромережі	Transformer	Transformer	GCNN	CNN
Мови тексту	Англійська	Давньогрецька	Англійська	Шведська, латина
Часовий період	XVI—XX ст.	800 д.н.е —800 н.е	1987—2010	817—1540
Розмір датасету	35599	78608	~650000	10992
Джерело даних	Книги у вільному доступі	Інскрипції	Газети	Манускрипти
Дані, що приймаються	Текст	Текст	Текст	Зображення
Спеціалізованість на датуванні	+	—	+	+
Точність	86 %, ± 30 років 73 %, ± 20 років	± 29,3 роки (перевершує експертів ± 144,4 роки)	61.5%, ± 0,87 роки	± 20 років (близько до результатів експертів)

Результати тестування точності розробленої нейронної мережі співвідносні з показниками іншої трансформерної моделі — Ithaca, яка використовується для датування давньогрецьких інскрипцій. Проте у випадку англійської мови та настільки широкого часового діапазону (XVI—XX ст.) отриманий результат є унікальним.

Інше відоме рішення для датування англомовних текстів — NeuralDater — демонструє нижчу точність і охоплює значно вужчий часовий період (1987–2010 роки).

На початку дослідження припущено, що результат передбачення датування тексту трансформерною нейронною мережею можна буде пояснити, виділивши лінгвістичні ознаки у вигляді слів, які найбільше вплинули на прийняття рішення мережею. Основою для цього мав слугувати механізм уваги трансформерних нейронних мереж. В процесі обробки тексту моделлю кожному вхідному токену ставиться у відповідність значення уваги, що демонструє наскільки сильні його контекстуальні зв’язки з іншими токенами у тексті. Припускалося, що слова, яким відповідають

токени з більшим значенням уваги, будуть більше впливати на результат передбачення десятиліття написання. Проте виявлено, що частіше високі значення уваги отримують токени, що сильно пов'язані з іншими, але не несуть корисної інформації для людської інтерпретації, такі як прийменники та знаки пунктуації. З цього дослідження можна зробити висновок, що передбачення десятиліття написання відбувається скоріше за загальним стилем написання, ніж за конкретними словами. Тому реалізація виділення ознак належності тексту певному періоду потребує іншого вирішення, до прикладу, обчислення впливу цілих груп токенів, а не одиничних токенів.

Висновки

Розроблена трансформерна нейронна мережа здатна визначати приблизне десятиліття написання англомовного тексту з точністю до 30 років та рівнем правильності 85 % у періоді XV—XX століть. Це перевершує результати аналогічних відомих моделей, що працюють з англомовними текстами. Проте варто зазначити, що використаний набір даних має обмеження, оскільки точний рік написання текстів у ньому вказано приблизно.

Перевагами запропонованого підходу є здатність моделі автоматично опрацьовувати велику кількість текстових даних без потреби в ручному виборі ознак, а також використання сучасної архітектури трансформерів, яка дозволяє ефективно враховувати контекст слів у тексті. До того ж реалізована модель є універсальною — її можна адаптувати для інших завдань історичної обробки текстів, таких як визначення авторства чи стилістичний аналіз. Попри складність інтерпретації результатів через обмеженість механізму уваги в поясненні прийнятих рішень, запропонований підхід відкриває перспективи для подальших досліджень. Зокрема, доцільно розробити нові методи інтерпретації, основані на аналізі впливу груп токенів або стилістичних шаблонів, що дозволить краще зрозуміти, які саме лінгвістичні ознаки сигналізують про часову належність тексту.

Подальші дослідження можуть зосередитися на оптимізації архітектури мережі та її гіперпараметрів, а також збільшенні кількості епох навчання для підвищення точності. Розширення набору даних, зокрема, додавання більшої кількості творів, написаних до XIX століття, допоможе покращити результати для ранніх періодів. Іншим важливим напрямом досліджень є пошук способів виділення лінгвістичних чи стилістичних ознак належності текстів певному періоду, задля покращення можливості інтерпретації результатів роботи нейронної мережі людиною. До того ж варто розглянути використання інших трансформерних моделей, здатних обробляти довші текстові фрагменти.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] Y. Assael, T. Sommerschield, et al, "Restoring and attributing ancient texts using deep neural networks," *Nature 603*, pp. 280-283, 2022. https://doi.org/10.1038/s41586-022-04448-z .

[2] Shikhar Vashishth, Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar. "Dating Documents using Graph Convolution Networks," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, vol. 1 (Long Papers), pp. 1605-1615, Melbourne, Australia. Association for Computational Linguistics. 2018. https://doi.org/10.18653/v1/P18-1149 .

[3] Wahlberg, Fredrik & Wilkinson, Tomas & Brun, Anders, *Historical Manuscript Production Date Estimation Using Deep Convolutional Neural Networks*, 2016. https://doi.org/10.1109/ICFHR.2016.0048 .

[4] O. Hellwig, "Dating Sanskrit texts using linguistic features and neural networks," 2019. [Електронний ресурс]. Режим доступу: https://www.academia.edu/53885816/Dating_Sanskrit_texts_using_linguistic_features_and_neural_networks.3073703.

[5] Ashish Vaswani, et al., "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, pp. 6000-6010, 2017. [Electronic resource]. Available: https://dl.acm.org/doi/10.5555/3295222.3295349 .

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018. https://doi.org/10.48550/arXiv.1810.04805 .

[7] Project Gutenberg — English Language eBooks. [Electronic resource]. Available: https://huggingface.co/datasets/sedthh/gutenberg_english .

Рекомендована кафедрою автоматизації та інтелектуальних інформаційних технологій ВНТУ

Стаття надійшла до редакції 14.04.2025

Литвин Михайло Олегович — студент факультету прикладної математики, e-mail: litvinka42@gmail.com ;

Олещенко Любов Михайлівна — канд. техн. наук, доцент, доцент кафедри програмного забезпечення комп'ютерних систем, e-mail: oleshchenkoliubov@gmail.com .

Національний технічний університет «Київський політехнічний інститут імені Ігоря Сікорського», Київ

M. O. Lytvyn[1]
L. M. Oleshchenko[1]

Automated Approach for Dating English Text Using Transformer Neural Networks

[1]National Technical University of Ukraine “Igor Sikorsky Kyiv polytechnic institute”

The paper examines the existing methods of text dating using neural networks, highlighting their advantages and limitations. Text dating is a crucial task in fields such as history, archival studies, linguistics, and forensic science, as accurately determining the creation time of a document can help verify its authenticity, establish authorship, and detect forgeries. However, traditional methods based on stylometric or statistical approaches often lack accuracy, especially when dealing with large volumes of text data. This study proposes an approach for dating English-language texts using transformer neural networks. The model achieves an accuracy of 85 % within a 30-year range for texts written between the 15th and 20th centuries, outperforming existing models applied to English text. The core idea of the proposed automated approach is to utilize transfer learning to fine-tune a pre-trained transformer neural network, optimizing it for the classification of text fragments by decade. One key advantage of this approach is the use of transformer architecture, which, through the self-attention mechanism, effectively captures complex relationships within a text. Another significant benefit is the application of transfer learning, which reduces training time and computational resources compared to training a model from scratch. The approach was implemented in Python using the transformers libraries for training and testing the neural network, datasets for working with the dataset, and numpy for the calculations. Experimental results demonstrated high accuracy: 86 % within a 30-year range and 73 % within a 20-year range on the test dataset. For the 19th and 20th centuries, the model achieved an accuracy of 89% and 90%, respectively, while accuracy for earlier centuries was lower, averaging around 30%. The research also examines the possibility of identifying features that indicate a text's association with a specific period by extracting words with the highest attention scores. Future research will focus on improving the accuracy for underrepresented historical periods by expanding and refining the dataset. Further enhancements may be achieved by optimizing model hyperparameters and experimenting with alternative neural network architectures. Another direction for future research is to explore methods for identifying linguistic or stylistic features that mark texts as belonging to a certain historical period, in order to make the neural network's results more interpretable for the user. The proposed approach has potential applications in historical research, document authentication, plagiarism detection, literary studies, and forensic analysis.

Keywords: software natural language processing (NLP), machine learning, transformer neural networks (TNN), transfer learning, BERT, text dating, stylometry, historical text analysis.

Lytvyn Mykhailo O. — Student of the Department of Applied Mathematics, e-mail: litvinka42@gmail.com ;

Oleshchenko Liubov M. — Cand. Sc. (Eng.), Associate Professor of the Chair of Computer Systems Software, e-mail: oleshchenkoliubov@gmail.com