Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 2 (32)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

УДК 303.732.4
DOI: https://doi.org/10.30837/2522-9818.2025.2.151

В. Шкурко А. Поляков

ОРГАНІЗАЦІЯ ПРОГРАМНИХ І НЕЙРОМЕРЕЖЕВИХ АЛГОРИТМІВ МАШИННОГО АНАЛІЗУ ТЕКСТОВИХ ПОВІДОМЛЕНЬ, ПОДАНИХ ПРИРОДНОЮ МОВОЮ

У статті розглянуто питання організації програмних і нейромережевих алгоритмів машинного аналізу текстових даних, поданих природною мовою. Обґрунтовано актуальність завдання оброблення як стислих текстових повідомлень і відгуків, що потребують швидкого оброблення з мінімальними ресурсними витратами, так і складних структурованих документів, які вимагають збереження структурних характеристик і глибокого контекстного аналізу. Проведено комплексний аналіз сучасних методів машинного оброблення текстової інформації, зокрема токенізації, кластеризації, семантико-релевантного пошуку й застосування нейромережевих архітектур. Особливу увагу приділено підходам, що дають змогу оптимізувати обчислювальні витрати без суттєвого зниження якості результатів аналізу, що є критично важливим для роботи в умовах обмежених ресурсів. На основі аналізу розроблено багаторівневу методику організації машинного аналізу текстових даних. Методика передбачає попередню класифікацію текстових масивів за типами документів, групування текстів методом кластеризації для підвищення релевантності оброблення та застосування нейромережевих моделей глибокого навчання. Для глибокого аналізу текстової інформації реалізовано архітектуру на основі двонаправленої рекурентної нейронної мережі (*Bidirectional LSTM*) із використанням регуляризації *Dropout* та механізмів раннього припинення навчання. З метою практичної перевірки запропонованої методики розроблено застосунок для автоматизованого аналізу стислих текстових повідомлень природною мовою. Подано результати навчання моделі, побудовано графіки динаміки зміни функції втрат на тренувальних і валідаційних вибірках, розроблено матриці помилок та візуалізацію результатів прогнозування. Продемонстровано стабільне зниження функції втрат без суттєвого збільшення обчислювальних витрат системи. Запропонована методика може бути застосована в інформаційних системах різного призначення для автоматизованого оброблення текстових повідомлень у режимах з обмеженими ресурсами, а також має перспективи подальшого розвитку в напрямі аналізу мультимодальних даних і впровадження в реальні інформаційно-аналітичні комплекси.

Ключові слова: машинний аналіз; текстові дані; природна мова; програмні алгоритми; нейромережева архітектура.

Вступ

Розвиток інформаційних технологій [1], упровадження нейромережевих алгоритмів глибокого навчання [2, 3] та поширення концепції *Big Data* [4, 5] дали змогу автоматизувати оброблення великих масивів текстових даних, поданих природною мовою. Застосування програмних і нейромережевих алгоритмів машинного аналізу текстів стає важливим завданням як для оброблення складних структурованих документів (наукових, технічних, юридичних текстів), так і для стислих текстових повідомлень і відгуків, що потребують швидкого оброблення в умовах обмежених обчислювальних ресурсів.

Оброблення невеликих текстів потребує уваги до таких особливостей: обмеженого обсягу інформації, високої варіативності лексичних конструкцій та необхідності збереження контексту за мінімальної довжини вхідних послідовностей. А для аналізу складних документів важливо забезпечити оброблення структурних і термінологічних особливостей тексту.

Сучасні системи машинного аналізу текстових даних мають бути адаптивними до різних форматів вхідної інформації – від лаконічних повідомлень до великих структурованих текстів – з огляду на обмеження обчислювальних ресурсів.

Зазначимо, що сучасні апаратно-програмні платформи центрів оброблення даних та мережеві сервіси на їх основі надають широкий інструментарій для автоматизованої роботи з великими текстовими масивами в галузях, що цілковито охоплюють життєдіяльність сучасної людини.

1. Соціальні мережі та пошукові системи, що передбачають семантичний аналіз запитів користувачів, виявлення громадської думки й актуальних трендів, індексацію та класифікацію контенту.

2. Лінгвістичні та філологічні дослідження, що передбачають автоматичний переклад текстів, аналіз стилю та авторства, а також лексикографію та семантичний аналіз.

3. Наука, інженерія, медицина й навчання, що передбачають систематизацію знань і автоматизацію рецензування, формування адаптивних систем

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2025. No. 2 (32)

персоналізованого навчання, виявлення симптомів і автоматизацію діагностики, а також управління проєктами й технічну підтримку.

4. Державне управління та судова практика, що передбачає моніторинг і оцінювання впровадження політик, визначення громадських потреб та пріоритетів, виявлення прецедентів і правових аргументів.

5. Фінансовий аналіз, електронна комерція та маркетинг, що передбачає аналіз і прогнозування ринкових трендів, вивчення поведінкових патернів споживачів, а також персоналізацію рекомендацій.

Багаторівнева класифікація, що може бути побудована на основі відповідного переліку (рис. 1), дає змогу надалі оцінити актуальні завдання щодо побудови цілісної методології організації та адаптації програмних і нейромережевих алгоритмів машинного аналізу текстових даних, поданих природною мовою.

Рис. 1. Багаторівнева класифікація актуальних галузей застосування засобів машинного аналізу текстових даних

Як показав аналіз наукових досліджень, присвячених проблемам упровадження систем автоматизації процедури аналізу великих масивів текстових даних, організація та адаптація відповідних програмних і нейромережевих алгоритмів є комплексним завданням [6–10]. Дослідники зазначають,

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 2 (32) ISSN 2524-2296 (online)

що в розробленні математичної моделі та нейромережевої архітектури, що може лягти в основу алгоритмів машинного аналізу текстових даних, важливо визначити ресурс пам'яті, обчислювальний ресурс і перепускність інформаційних каналів апаратної платформи [6], а отже, в основі проєктування має лежати компроміс між продуктивністю інформаційної системи та рівнем витрат. У впровадженні нейромережевих алгоритмів також мають бути оцінені затрати на навчання великих мовних моделей (*Large Language Models*; LLM) з пріоритизацією важливості взаємодії між відповідними алгоритмами та наявним програмним забезпеченням апаратної платформи [6, 7] згідно з концепцією попередньо тренованої мовної моделі (*Pre-Trained Language Model*; PLM).

Зі свого боку ефективність процедури навчання штучної нейромережі (*Artificial Neural Network*; ANN) основана на підходах, що використовуються в дослідженні методів моделювання мови [7]. Завдання полягає в побудові алгоритмів, здатних із мінімальною затримкою виконувати поточні запити з необхідним рівнем точності, а також піддаватись масштабуванню, адаптації та налаштуванню [6, 7] за умов розширення поставленого набору завдань. Зазначимо, що оптимізація відповідних нейромережевих алгоритмів щодо збільшення продуктивності та зменшення ресурсоємності передбачає низку сучасних підходів, що лежать в основі концепції крайового машинного навчання (*Edge Machine Learning*; EML), зокрема низькорозрядна квантизація (*Low-Bit Quantization*), параметроефективне донавчання (*Parameter-Efficient Fine-Tuning*; PEFT), а також аналіз гіперпараметрів [8, 11]. Крім того, постійне оновлення PLM відповідно до актуальних даних і подолання таких обмежень, як погана інтерпретованість результатів машинного аналізу й потреба в підготовці великих масивів анотованих даних [9], пропонується виконати за допомогою формування нейромережевої архітектури мовної моделі з попереднім навчанням та підсиленням на основі зовнішньої бази знань (*Knowledge-Enhanced Pre-trained Language Models*; KEPLMs) з визначенням методів адаптації ANN [9, 12]. Нарешті, основна увага в галузі машинного аналізу текстової інформації, поданої природною мовою, приділяється моделям трансформерів [10], зокрема таким, що лежать в основі архітектури генеративного трансформера попереднього навчання (*Generative Pre-Trained Transformer*; GPT) та використовується в онлайн-сервісах на базі *ChatGPT*. Підтверджено ефективність застосування моделей змішаної точності (*Mixed-Precision Model*; MPM) для зменшення застосування ресурсу пам'яті, а також збільшення точності та швидкості оброблення вхідних запитів [10, 12–14]. Необхідно зазначити, що з метою адаптації LLM для виконання практичних завдань, зокрема аналізу фондових ринків, необхідно розширити інструментарій нейромережевого аналізу за допомогою впровадження в межах інформаційної системи різних форматів даних, що є нерозв'язаною частиною загального дослідження.

Окремо варто наголосити, що в багатьох випадках використання повноцінних обчислювальних потужностей для машинного аналізу текстових повідомлень є недоступним через обмеження ресурсів кінцевих пристроїв або інфраструктури. Тому особлива увага в дослідженні приділяється підходам до розроблення систем, здатних працювати в середовищах з низькою обчислювальною потужністю без суттєвого зниження точності аналізу.

Мета роботи полягає в розробленні комплексної методики організації програмних і нейромережевих алгоритмів машинного аналізу текстових даних природною мовою з огляду на обмеження ресурсів та особливості оброблення різних форматів текстової інформації.

Для досягнення поставленої мети необхідно розв'язати такі ***завдання***:

– проаналізувати сучасні методи машинного аналізу текстових даних, поданих природною мовою, зважаючи на їх переваги та обмеження;

– розробити концепцію семантико-релевантного пошуку для оброблення стислих і великих текстів;

– побудувати узагальнену схему організації інформаційної системи машинного аналізу текстів;

– визначити особливості адаптації програмних алгоритмів для стислих текстових повідомлень;

– дослідити можливості застосування нейромережевих архітектур для глибокого оброблення текстів;

– мінімізувати обчислювальні витрати інформаційної системи без суттєвого зниження якості аналізу текстових даних.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online)
Innovative technologies and scientific solutions for industries. 2025. No. 2 (32)

1. Постановка завдання машинного аналізу масивів текстових даних, поданих природною мовою

Зростання обсягів текстових повідомлень природною мовою, що потребують автоматизованого оброблення, зумовлює необхідність створення систем машинного аналізу тексту. Текстові масиви містять як стислі повідомлення (відгуки, запити, коментарі), так і структуровані документи (наукові публікації, технічну, юридичну та фінансову документацію, новинні статті), кожен з яких має особливості подання та вимоги до оброблення.

Постановка завдання машинного аналізу текстових даних полягає в побудові системи, що забезпечує:

– оброблення великих масивів текстів з огляду на їх контекст і семантичні зв'язки;

– підтримку оброблення як стислих, так і складних текстових форматів;

– оптимізацію обчислювальних ресурсів у процесі збереження точності результатів аналізу;

– мінімізацію часу оброблення та обсягу пам'яті, необхідної для роботи із текстовими масивами.

Основні складнощі розв'язання завдання пов'язані з експоненційним зростанням обсягів текстових повідомлень, необхідністю уваги до супутніх контекстних характеристик текстів та потребою підтримки різних форматів структурованої інформації.

Зменшення вимог щодо обчислювального ресурсу, перепускності інформаційних каналів та ресурсу пам'яті апаратної платформи одночасно зі збільшенням продуктивності інформаційної системи машинного аналізу текстових даних може бути виконано завдяки впровадженню концепції семантико-релевантного пошуку [15–18], що ґрунтується на визначенні контекстного значення окремих складників вхідного запиту. Упровадження зазначеної концепції в цьому разі може базуватись на алгоритмах, що визначаються мінімальним навантаженням на ресурси апаратної платформи та реалізації водночас розширеного функціоналу, як це показано на діаграмі (рис. 2).

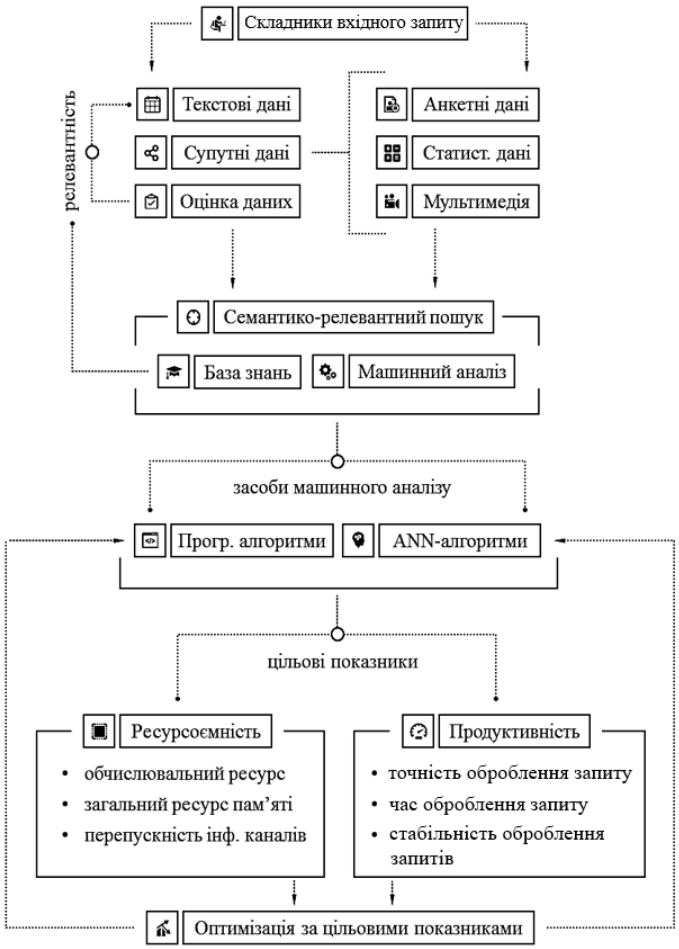

Рис. 2. Узагальнена схема організації системи машинного аналізу текстових даних на основі семантико-релевантного пошуку

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 2 (32)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

Відповідно до схеми першим етапом проєктування системи машинного аналізу текстових повідомлень на основі семантико-релевантного пошуку є визначення складників вхідного запиту, а саме текстової інформації з оцінкою її релевантності та супутніх даних, що містять такі категорії, як анкетні, статистичні та мультимедійні дані. Семантико-релевантний пошук проводиться на основі бази знань та машинного аналізу із застосуванням нейромережевих та програмних алгоритмів. Оптимізація та налаштування сервісу проводиться відповідно до набору цільових показників точності та швидкодії в процесі оброблення вхідних запитів, що визначаються на кількісному рівні (як на етапі проєктування способом розрахунку згідно з математичною моделлю, так і на етапі експлуатації програмної системи).

Організація системи машинного аналізу текстових повідомлень на основі семантико-релевантного пошуку, отже, основана на отриманні максимально широкого набору даних відповідно до складників вхідного запиту, а також мети оброблення інформації для досягнення релевантного результату. У низці випадків збільшення продуктивності машинного аналізу за умов зменшення навантаження на ресурс апаратної платформи можливе завдяки долученню в аналіз фіксованої структури документа, властивої для відповідної категорії текстових даних.

У межах дослідження пропонується виокремити категорії організації масивів текстової інформації, що є найбільш актуальними для машинного аналізу.

1. Наукові публікації, структура яких зазвичай містить послідовний набір таких текстових блоків: назва публікації; автори; перелік наукових установ; тези; ключові слова; вступ; аналіз публікацій у профільних виданнях; основна частина; висновки; перелік джерел посилання; назва видання, у якому розміщена публікація.

2. Технічна документація, структура якої зазвичай містить послідовний набір таких текстових блоків: назва документа; назва організації; автори; дата випуску; вступ; зміст; основна частина (загальні відомості, опис продукту, інструкції з використання, технічні характеристики, процедури обслуговування та ремонту); безпека та попередження; додатки.

3. Юридична документація, структура якої зазвичай передбачає послідовний набір таких текстових блоків: назва документа; перелік авторів / учасників; дата документа; преамбула; основна частина (визначення термінів, умови, зобов'язання сторін, санкції та відповідальність); завершальні положення; додаткові умови; підписи сторін.

4. Фінансова документація, структура якої зазвичай містить послідовний набір таких текстових блоків: назва документа; організація; перелік відповідальних осіб; дата звіту; вступ; основна частина (баланс, звіт про прибутки та збитки, звіт про рух грошових коштів, звіт про зміни у власному капіталі); примітки до фінансової звітності; аудиторський висновок; додатки.

5. Новинні статті, структура яких зазвичай передбачає послідовний набір таких текстових блоків: заголовок; автор; дата публікації; короткий опис; вступ; основна частина (факти, цитати, події); додаткова інформація; підсумки; джерела.

6. Стислі повідомлення із соціальних мереж, відгуків, коментарів, які часто мають завуальований або емоційно забарвлений контекст, містять скорочення, сленг або неструктуровану інформацію. Структура таких текстів, як правило, є фрагментарною та мінімально формалізованою: автор повідомлення, дата або час публікації, текст повідомлення, іноді – мультимедійні вкладення або реакції інших користувачів. Особливу складність для машинного аналізу становить необхідність розпізнавання прихованого змісту, емоційної тональності та інтерпретації контексту за мінімальної кількості лінгвістичних підказок.

Наявність фіксованої структури масиву текстових даних, що підлягає аналізу, надає додаткові можливості для пошуку компромісу між продуктивністю системи машинного аналізу, часом оброблення вхідного запису та навантаженням на ресурс апаратної платформи. Так, для категорії "Наукові публікації" такі структурні елементи, як "назва видання, у якому розміщена публікація", "перелік авторів" і "перелік наукових установ" дає змогу оцінити релевантність і достовірність поданої інформації без необхідності детального аналізу всього текстового масиву, а структурні елементи "тези" та "висновки" можуть бути використані для зменшення загального обсягу даних, що підлягає аналізу без суттєвого зменшення точності оброблення вхідного запиту. Це сприяє значному послабленню навантаження на обчислювальний ресурс, перепускність інформаційного каналу та ресурс пам'яті.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2025. No. 2 (32)

Розроблення системи машинного аналізу масивів текстових даних природною мовою передбачає розв'язання таких основних завдань:

– побудова концепції семантико-релевантного пошуку текстової інформації, яка базується на оцінці смислової близькості між текстовими блоками та контекстних властивостей стислих повідомлень;

– упровадження програмних і нейромережевих алгоритмів попереднього оброблення, токенізації, кластеризації та глибокого аналізу текстів для підвищення якості інтерпретації змісту даних;

– розроблення підходів до оптимізації машинного аналізу текстової інформації з огляду на обмеження обчислювальних ресурсів системи (процесорний час, обсяг пам'яті, пропускна здатність каналів оброблення запитів);

– визначення критеріїв якості аналізу текстових повідомлень за кількісними показниками, такими як точність класифікації, середня похибка прогнозування та обсяг використаних ресурсів.

Запропонований підхід дає змогу адаптувати машинний аналіз до особливостей як стислих неструктурованих повідомлень, так і великих структурованих документів, що розширює можливості побудови гнучких і продуктивних інформаційних систем оброблення текстових даних природною мовою.

2. Методика організації програмного оброблення текстових даних із використанням нейромережевих моделей

З метою побудови системи машинного аналізу текстової інформації природною мовою, що поєднує оброблення як стислих повідомлень, так і складних структурованих документів, у роботі запропоновано багаторівневу методику організації програмних і нейромережевих алгоритмів. Методика основана на послідовному застосуванні етапів класифікації текстів, попереднього оброблення, кластеризації за семантичними ознаками, релевантного пошуку та глибокого аналізу текстових даних із використанням нейронних мереж. У цьому разі особливу увагу приділено мінімізації обчислювальних витрат без суттєвого зниження точності аналізу. Належна адаптація програмних алгоритмів для організації системи машинного аналізу текстових повідомлень дає змогу зменшити навантаження на ресурс апаратної платформи та збільшити швидкодію оброблення вхідної інформації.

2.1. Етапи організації машинного аналізу текстових даних

У межах цього дослідження пропонуємо структурування процесу аналізу текстових даних у вигляді послідовності етапів, кожен з яких розв'язує конкретне завдання попереднього оброблення, класифікації або глибокого аналізу текстових масивів.

Методика передбачає п'ять основних етапів.

1. Аналіз структури вхідних текстових даних, де аналізується характер вхідної інформації з метою її класифікації за типами документів. Береться до уваги різниця між стислими повідомленнями (соціальні мережі, відгуки) та великими структурованими документами (наукові статті, технічна й фінансова документація). Визначення типу тексту дає змогу попередньо обрати відповідні підходи до його оброблення, оптимізувати побудову моделі оброблення інформації та сформувати очікувану структуру вихідних результатів. Для кожної категорії текстових повідомлень формується типова модель структури, що зважає на формат подання, наявність метаданих і можливі супутні елементи (заголовки, авторські поля, дати). У межах запропонованої методики тексти поділяються на основі їх функціонального призначення, що наведено в табл. 1.

Таблиця 1. *Класифікація типів текстових документів*

№	Тип документа	Характеристика	Приклади
1	Наукові публікації	Використання термінології, структурованість	Статті, монографії
2	Технічна документація	Регламентовані формати опису процесів	Інструкції, технічні паспорти
3	Юридична документація	Формалізовані структури, правові терміни	Контракти, договори, рішення суду
4	Фінансова документація	Числова інформація, аналітичні висновки	Баланси, фінансові звіти
5	Новинні статті	Динамічні, оперативні тексти	Новини, репортажі
6	Стислі повідомлення з соціальних мереж, відгуків, коментарів	Неструктуровані тексти, емоційна забарвленість, лаконічна форма	Відгуки, дописи в соцмережах, коментарі

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 2 (32)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

2. Попереднє оброблення текстових масивів, де здійснюється токенізація текстових даних (поділ тексту на окремі значущі одиниці: слова, підслова або символи). Для великих за обсягом документів особливу увагу приділяють збереженню логічних розділів і контексту між абзацами, тоді як для стислих повідомлень важливішою є нормалізація тексту та вилучення шуму. Паралельно застосовуються методи попереднього виділення ключових слів і тематичних груп завдяки частотного аналізу лексем. Для підвищення релевантності оброблення застосовується кластеризація текстів за змістовими ознаками, що дає змогу згрупувати тексти в тематично однорідні підмножини, мінімізуючи складність подальшого аналізу.

3. Семантико-релевантний аналіз текстових даних, де реалізується процедура визначення змістової близькості текстових елементів на основі семантичних ознак. Оброблення здійснюється із застосуванням концепції семантико-релевантного пошуку, що передбачає формування простору ознак для текстових блоків та оцінювання їх відносної релевантності до запиту чи бази знань. Релевантність розраховується як за класичними статистичними методами (наприклад, TF-IDF або BM25), так і за допомогою векторного подання текстів (Word2Vec, GloVe, BERT-ембединги). Семантична релевантність визначає пріоритетність текстів для подальшого глибокого аналізу.

4. Глибокий аналіз за допомогою нейромережевих моделей, де в запропонованому рішенні використовується архітектура двонаправленої рекурентної нейронної мережі (*Bidirectional LSTM*), що дає змогу моделювати залежності між словами як у прямому, так і у зворотному напрямках. Застосування двонаправленого оброблення особливо актуальне для невеликих за обсягом текстів, де кожен елемент має вагоме контекстне значення. Модель налаштовується на розв'язання завдання класифікації текстів за тональністю або іншими ознаками за мінімальної кількості епох і обмеженого розміру вхідного словника.

5. Оптимізація обчислювальних ресурсів на всіх етапах оброблення, де мінімізації витрат ресурсів інформаційної системи реалізуються за допомогою використання методів регуляризації моделі (*Dropout, Batch Normalization*), раннього припинення навчання (*EarlyStopping*), адаптивного регулювання розміру пакетів даних. Також передбачено обмеження обсягу словника ознак і скорочення довжини послідовностей для текстів з низькою інформаційною насиченістю. Сукупність зазначених заходів сприяє зменшенню обсягів споживаної оперативної пам'яті та скороченню часу оброблення без суттєвого зниження точності результатів.

Запропоновані етапи дають змогу формалізувати оброблення текстової інформації природною мовою таким чином, щоб поєднати адаптивність системи до різних типів текстових масивів з оптимальним використанням обчислювальних ресурсів. Це забезпечує можливість застосування системи в реальних умовах, зокрема в середовищах з обмеженими ресурсами або для оброблення великих обсягів текстових даних.

2.2. Використання кластеризації та токенізації під час попереднього оброблення текстових даних

На етапі попереднього оброблення та структуризації великих масивів текстової інформації найбільшу ефективність відповідно до швидкодії та мінімізації використання ресурсів демонструють такі групи методів:

– методи кластеризації масивів текстових повідомлень, а саме кластеризація зв'язності [19], кластеризація методом *k*-середнього [20], зокрема кластеризація із середнім зсувом [21], а також кластеризація за схемою DBSCAN [22];

– методи токенізації [23, 24], зокрема токенізація на основі слів, токенізація на основі підслів і токенізація на основі символів;

– методи ймовірнісної класифікації складників масивів текстових повідомлень із застосуванням моделі гауссової суміші [25].

Кластеризація як група методів поділу масиву текстових даних на підмножини (кластери) визначається найбільш широким функціоналом, гнучкістю та масштабованістю. У межах дослідження пропонуємо поділити методи кластеризації на жорсткі, у межах яких ділянки кластерів не перетинаються, і м'які, де окремі елементи текстового блоку можуть одночасно належати до кількох кластерів [19–23]. Базовими підходами кластеризації, що використовується в межах концепції семантико-релевантного пошуку, є визначення ділянок кластерів на основі онтології семантичного словника, що адаптується для текстової інформації, поданої природною мовою [5, 19], та виділення ключових слів, що за типовим сценарієм визначаються програмними алгоритмами відповідно

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2025. No. 2 (32)

до показника частоти вживання зазначених термінів і структури документа [30]. Водночас зауважимо, що нині під час роботи з текстовими даними найбільш широко використовуються алгоритми кластеризації, основані на методі *k*-середнього [20–22]. Згідно із зазначеним методом для кожної ділянки кластера з огляду на середнє арифметичне ключових ознак його елементів циклічно розраховується центроїд. Кількість циклів залежить від стійкості центроїда, а також вимог щодо мінімізації дисперсії між елементами в середині кластера та максимізації дисперсії складників між окремими кластерами (рис. 3). Відповідні алгоритми кластеризації визначаються мінімальним навантаженням на ресурс апаратної платформи, але мають низку недоліків, що здебільшого пов'язано з випадковим характером початкового вибору центроїдів ділянок кластерів [20–22]. Розв'язання зазначеної проблеми полягає в упровадженні методів кластеризації із середнім зсувом як непараметричного аналізу простору ознак масиву текстових даних із визначенням місця розташування ділянок високої щільності ймовірності [21].

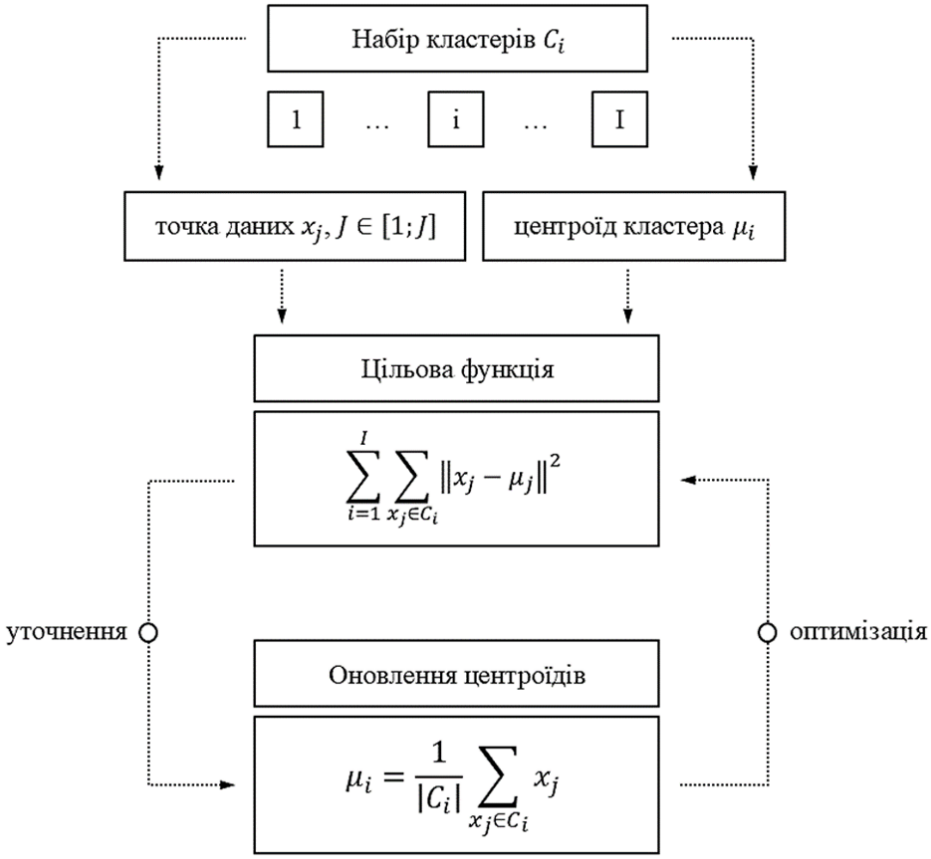

Рис. 3. Схема кластеризації на основі методу *k*-середнього

Важливо зауважити, що програмні алгоритми на основі методів кластеризації із середнім зсувом, зокрема такі, що базуються на схемі DBSCAN *(Density-Based Spatial Clustering of Applications with Noise)* можуть також ефективно застосовуватись у кластерному аналізі супутніх даних [22]. Відповідна схема кластеризації передбачає визначення таких категорій (рис. 4):

– ділянка щільності, що встановлюється радіусом;

– показник, який визначає мінімальну кількість точок у ділянці щільності, що є необхідною для формування точки ядра;

– точки ядра, що встановлюються за допомогою показника;

– пригранічні точки, що знаходять у ділянці щільності принаймні однієї точки ядра;

– шумові точки, що лежать поза ділянки щільності всіх точок ядра.

Недоліки методів кластеризації із середнім зсувом полягають у невизначеності низки елементів,

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 2 (32) ISSN 2524-2296 (online)

що лежать на межі ділянок кластерів і зазвичай належать до викидів. Це виконується за допомогою вибору кластеризації на підставі зв'язності між елементами [19], алгоритми на основі якої формують ділянки кластерів у вигляді деревоподібної структури, де кожен кластер поданий у вигляді вузла загальної ієрархічної структури. У цьому разі для алгоритму машинного аналізу текстової інформації може бути обрана метрика для обчислення відстані між елементами масиву, що надалі відіграє ключову роль у кластеризації. Методи кластеризації зв'язності застосовуються для ідентифікації та класифікації параметрів текстових даних, водночас відповідна архітектура є достатньо гнучкою для модифікацій та масштабування.

Крім того, у використанні методів імовірнісної класифікації застосовується ранжування відповідно до оцінки за критерієм максимальної схожості складників (*Estimation Maximization*; EM) з використанням імовірнісної моделі гауссової суміші (*Gaussian Mixture Models*, GMM). Ці моделі передбачають, що дані можуть бути подані як комбінація кількох гауссових розподілів, кожен з яких відповідає окремій групі або кластеру. На відміну від детермінованих підходів, імовірнісні моделі беруть до уваги невизначеність та розподіл даних у просторі. Перелічимо основні етапи роботи.

1. Ініціалізація параметрів, де задаються початкові значення параметрів, таких як середнє значення та коваріаційна матриця для кожного гауссового компонента.

2. Розрахунок очікувань, де для кожного зразка даних обчислюється ймовірність належності до кожного з компонентів гауссової суміші. Ці ймовірності використовуються для розрахунку очікуваного значення функції правдоподібності. Замість жорсткої класифікації, де кожен зразок чітко належить до одного кластера, ймовірнісна класифікація дає змогу розподілити зразки між кількома кластерами на основі ймовірності.

3. Максимізація параметрів, де після розрахунку очікуваних значень на основі отриманих імовірностей оновлюються параметри моделі (середнього, коваріаційної матриці і вагових коефіцієнтів для кожного компонента). Мета полягає в максимізації функції правдоподібності на основі оновлених параметрів.

4. Циклічне повторення кроків 1–3 доти, доки функція правдоподібності не досягне збіжності, тобто коли подальші зміни параметрів стануть незначними. Цей підхід дає змогу поступово покращувати модель, підлаштовуючи її під дані.

Алгоритми ймовірнісної класифікації використовуються з метою пошуку максимальної схожості відповідно до параметрів імовірнісної моделі. В основі цього підходу лежить припущення, що вихідні значення цілковито залежать від набору прихованих змінних. Виконання процедури ймовірнісної класифікації передбачає циклічне визначення очікуваного значення функції правдоподібності та оцінювання показника максимальної схожості до отримання збіжності згідно з поставленими умовами. Імовірнісна класифікація найбільш ефективна в роботі з текстовими повідомленнями, що належать до декількох груп (м'яка класифікація). Порівняно із середнім зсувом відповідні алгоритми є більш гнучкими й масштабованими, а також не обмеженими метрикою класичних методів кластеризації.

Нарешті, токенізація масивів текстової інформації використовується для збільшення ефективності подальшого застосування нейромережевих алгоритмів як на етапі навчання, так і на етапі машинного аналізу [23, 24]. Залежно від обраного підходу для слів, підслів або символів надаються показники, подані на кількісному рівні, що визначають положення складника в текстовому блоці, а також наявність відповідного елемента в словнику.

Важливо зауважити, що ефективність токенізації елементів залежить від однорідності тексту, а отже, машинний аналіз має проводитись не для всього масиву даних, а для структурних складників. Це дає змогу зменшити розмір набору токенів і ефективно адаптувати модель відповідно до вхідного запиту.

Зі свого боку програмні алгоритми, що можуть бути використані на основному етапі машинного аналізу текстових даних, містять такі категорії:

– кількісна оцінка важливості окремих термінів [26, 27] у масиві текстових даних (*Term Frequency-Inverse Document Frequency*; TF-IDF);

– ранжування складників текстового запиту відповідно до показників релевантності пошуковому запиту [28, 29].

В основі методів TF-IDF [26, 27] лежить розрахунок вагового коефіцієнта, що визначає ступінь важливості окремого терміна в масиві текстових повідомлень, який підлягає машинному аналізу. Відповідна величина розраховується через

ISSN 2522-9818 (print)
ISSN 2524-2296 (online)
Innovative technologies and scientific solutions for industries. 2025. No. 2 (32)

добуток частоти появи терміна в окремому текстовому блоці (T*erm Frequency*; TF) та оберненої частоти для всього масиву текстової інформації (*Inverse Document Frequency*; IDF). Застосування цього підходу сприяє зниженню ваги загальних частовживаних, але малоінформативних слів, наприклад сполучників і прийменників. Програмні алгоритми на основі методу TF-IDF визначаються простотою реалізації за умов постановки завдання початкового аналізу великих обсягів текстових даних, оскільки обчислювальна складність є лінійною щодо розміру загального масиву. Натомість основним недоліком зазначеного методу є обмеження функціональних можливостей. Зазначається, що алгоритми TF-IDF проводять лише лексичний аналіз, тоді як для позиціювання термінів необхідно застосувати інші підходи. Зі свого боку за умов ранжування складників текстового запиту за показниками релевантності пошуковому запиту [28, 29] програмний алгоритм оцінює близькість запиту до масиву текстової інформації без необхідності додаткових розрахунків з метою визначення відповідності між векторними репрезентаціями вхідного запиту й текстових повідомлень. Математичний апарат, що лежить в основі алгоритму, також містить обчислювальні метрики методу TF-IDF, але відповідні вагові коефіцієнти термінів обчислюються не для створення векторного подання документа, а для співвідношення документа й вхідного запиту.

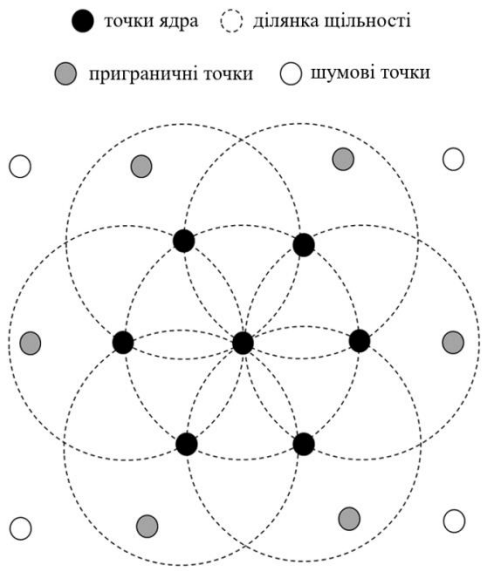

Рис. 4. Кластеризація на основі схеми DBSCAN

Перевага нейромережевих алгоритмів машинного аналізу масивів текстових даних полягає в розширенні функціоналу програмної системи без необхідності розроблення точної математичної моделі мовного корпусу як структурованої збірки текстів, що використовується для семантичного аналізу з метою дослідження мовних закономірностей. Недоліком зазначеного підходу є зростання навантаження на ресурс апаратної платформи й складність упровадження схеми оптимізації програмної системи відповідно до цільових показників продуктивності машинного аналізу [31–39].

2.3. Практична реалізація методики машинного аналізу текстових даних

На основі розробленої багаторівневої методики створено експериментальну програмну систему для автоматизованого машинного аналізу текстової інформації природною мовою для визначення тональності відгуків користувачів. Основною метою практичної реалізації є перевірка працездатності запропонованого підходу на прикладі завдання аналізу стислих текстових повідомлень.

Реалізація передбачає модулі попереднього оброблення текстових даних (токенізація, нормалізація, кластеризація), механізми семантико-релевантного аналізу й нейромережеву модель на основі архітектури *Bidirectional LSTM* для глибокого аналізу змісту повідомлень. Особливу увагу приділено мінімізації ресурсних витрат унаслідок впровадження регуляризації, оптимізації структури мережі та адаптації довжини вхідних послідовностей до характеру оброблюваних текстових даних.

Практична реалізація дала змогу оцінити можливості запропонованої методики в роботі з текстовими масивами різної структури та довжини, а також сформувати рекомендації щодо подальшого вдосконалення систем оброблення природної мови в ресурсно обмежених середовищах.

Створено експериментальну модель глибокого аналізу з використанням архітектури *Bidirectional LSTM* та налаштовано на розпізнавання тональності тексту способом оброблення токенізованих і нормалізованих послідовностей вхідної інформації (рис. 5). Конструкція моделі передбачає векторизацію ознак (*Embedding*), двонаправлену LSTM-структуру для оброблення контекстних залежностей, кілька повнозв’язних шарів для кінцевої регресії та механізми регуляризації для запобігання перенавчанню.

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 2 (32)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

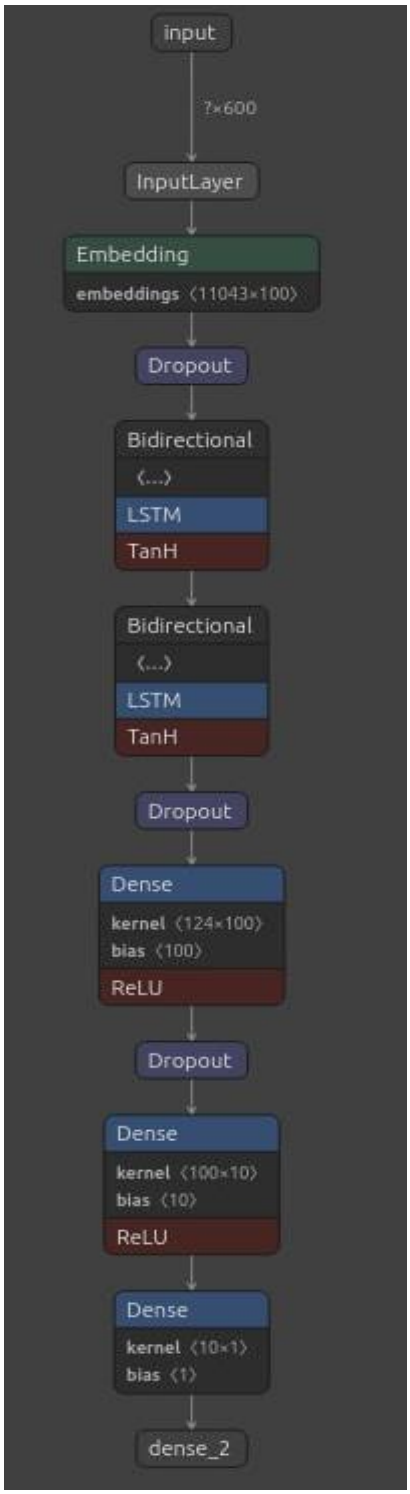

Рис. 5. Налаштування нейронної мережі для розпізнавання тональності тексту

Як показали дослідження, алгоритми на основі нейромережевої архітектури типу генеративного трансформера й моделей, побудованих на її основі, зокрема BERT, GPT та T5, мають найбільш широкий спектр застосувань в аналізі текстових даних. Завдяки механізму *self-attention* трансформери ефективно обробляють великі текстові масиви, зважаючи на контекст кожного терміна в межах всього документа [31–33]. Це дає змогу досягати високої точності у виконанні завдання щодо роботи з природними мовами, таких як переклад, класифікація, витяг релевантної інформації та генерація тексту. Зі свого боку архітектура RNN та сучасні моделі на їх основі, наприклад LSTM та GRU, також широко застосовуються в аналізі текстових повідомлень завдяки своїй здатності обробляти послідовності інформації [34–38]. Алгоритми на їх основі використовуються для завдань, що потребують визначення контексту та порядку слів, зокрема мовне моделювання, машинний переклад, розпізнавання мовлення та класифікація тексту. У цьому разі RNN мають обмеження в обробленні дуже довгих послідовностей через проблеми з градієнтним загасанням.

Нарешті, архітектура CNN більшою мірою асоціюється з обробленням матриці зображення, але вони також ефективні для аналізу текстових даних, поданих природною мовою [37–39]. У текстових завданнях CNN використовують згорткові фільтри, що допомагають виявляти суттєві особливості в тексті, наприклад комбінації слів або фраз, що мають важливе значення для конкретного завдання.

Для підвищення точності розпізнавання тональності стислих текстових повідомлень розроблено процедуру навчання нейронної мережі на основі оптимізованої структури шарів і механізмів регуляризації. Навчання здійснювалося з використанням функції втрат середньоквадратичної помилки (MSE) та оптимізатора *Adam* із початковою швидкістю навчання 0.001.

Для контролю перенавчання моделі застосовано механізм раннього припинення тренування (*EarlyStopping*), а для адаптації швидкості навчання – механізм *ReduceLROnPlateau.* Попередню токенізацію даних проведено з обмеженням розміру словника до 10 000 ознак і фіксацією довжини вхідних послідовностей на рівні 600 токенів.

Динаміка навчання нейронної мережі фіксувалася у вигляді протоколу, що відтворює зміну значень помилки на тренувальній та валідаційній вибірках після кожної епохи навчання.

На рис. 6 наведено фрагмент протоколу тренування, де продемонстровано кількість епох, значення функції помилки (*loss*) для тренувальної вибірки (*loss*), значення функції помилки для валідаційної вибірки (*val_loss*) та швидкість навчання (*lr*).

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2025. No. 2 (32)

```
48/48 [==============================] - 82s 2s/step - loss: 14.4485 - val_loss: 9.9785 - lr: 0.0010
Epoch 2/20
48/48 [==============================] - 79s 2s/step - loss: 9.8172 - val_loss: 7.3289 - lr: 0.0010
Epoch 3/20
48/48 [==============================] - 75s 2s/step - loss: 6.3964 - val_loss: 6.2269 - lr: 0.0010
Epoch 4/20
48/48 [==============================] - 78s 2s/step - loss: 4.4186 - val_loss: 5.5421 - lr: 0.0010
Epoch 5/20
48/48 [==============================] - 69s 1s/step - loss: 3.4891 - val_loss: 5.5777 - lr: 0.0010
Epoch 6/20
48/48 [==============================] - 69s 1s/step - loss: 2.8274 - val_loss: 4.4453 - lr: 0.0010
Epoch 7/20
48/48 [==============================] - 69s 1s/step - loss: 2.4941 - val_loss: 4.4178 - lr: 0.0010
Epoch 8/20
48/48 [==============================] - 68s 1s/step - loss: 2.1417 - val_loss: 4.4678 - lr: 0.0010
Epoch 9/20
48/48 [==============================] - 69s 1s/step - loss: 1.8395 - val_loss: 4.7220 - lr: 0.0010
Epoch 10/20
48/48 [==============================] - 67s 1s/step - loss: 1.7584 - val_loss: 4.7917 - lr: 0.0010
Epoch 11/20
48/48 [==============================] - 70s 1s/step - loss: 3.7593 - val_loss: 8.3930 - lr: 0.0010
Epoch 12/20
48/48 [==============================] - 74s 2s/step - loss: 3.6373 - val_loss: 4.8200 - lr: 0.0010
Epoch 13/20
48/48 [==============================] - 74s 2s/step - loss: 2.3809 - val_loss: 4.9578 - lr: 0.0010
```

Рис. 6. Фрагмент протоколу навчання нейронної мережі з показниками помилки на тренувальній та валідаційній вибірках

Для візуалізації результатів роботи нейронної мережі розроблено інтерфейс з метою виведення результатів прогнозування тональності стислих повідомлень (рис. 7), де кожному обробленому відгуку надається числова оцінка тональності в діапазоні від 1 до 5. Відгуки відтворюються в структурованому вигляді із зазначенням оцінки й тексту повідомлення. Для забезпечення зручності імпорту текстових даних у застосунок передбачено кнопку "Імпорт відгуків", що дає змогу автоматично додавати нові повідомлення для аналізу без необхідності ручного введення.

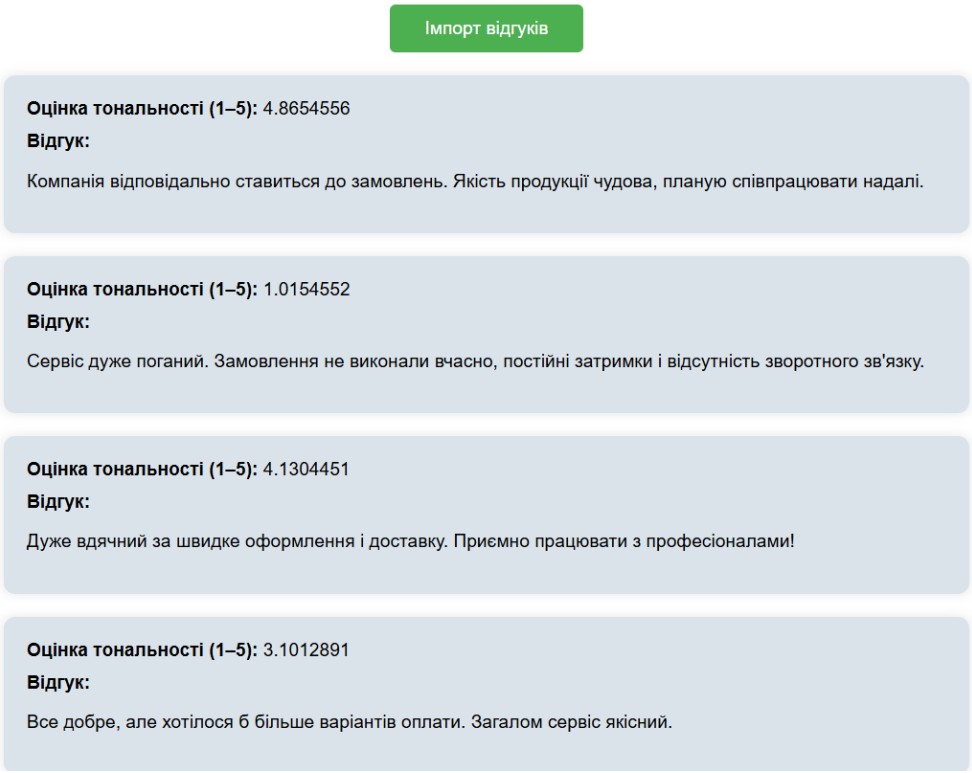

Рис. 7. Інтерфейс відтворення оцінювальних текстових відгуків у програмному застосунку

Особливістю практичної реалізації є орієнтація на мінімальні обчислювальні вимоги, що забезпечує можливість використання розробленої системи на стандартних персональних комп’ютерах середнього класу або у віртуалізованих середовищах із обмеженим доступом до ресурсів процесора й пам’яті.

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 2 (32)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

Для оцінювання якості запропонованого рішення були проведені експериментальні вимірювання ключових показників продуктивності. Модель тренувалася й тестувалася на апаратній платформі з обмеженими ресурсами: ноутбук з низькопотужним процесором Intel Core i3-8130U, 8 ГБ оперативної пам'яті та інтегрованою графікою. Основні результати навчання подано в табл. 2.

Таблиця 2. *Параметри тренування моделі машинного аналізу текстових даних*

Параметр	Значення
Кількість епох тренування	20
Кількість кроків на одну епоху	48
Середній час оброблення однієї епохи	≈73 с
Загальний час тренування	≈16 хв
Середнє завантаження процесора під час тренування	65–70 %
Максимальне використання оперативної пам'яті	≈3,2 ГБ
Середнє значення функції втрат (*Loss*) на тренувальній вибірці	4,82
Середнє значення функції втрат (*Loss*) на валідаційній вибірці	5,63

За результатами експериментів встановлено, що запропонована архітектура нейронної мережі забезпечує стабільне зниження функції втрат без суттєвого збільшення обчислювальних витрат. Отримані показники завантаження процесора та використання оперативної пам'яті свідчать про можливість впровадження розробленої системи навіть на апаратних платформах середнього класу без потреби в спеціалізованих графічних прискорювачах.

Висновки

Унаслідок проведеного дослідження розроблено та експериментально перевірено багаторівневу методику організації програмних і нейромережевих алгоритмів для машинного аналізу текстових даних, поданих природною мовою. Основну увагу приділено мінімізації обчислювальних витрат, що дало змогу забезпечити стабільну роботу системи на апаратних платформах з обмеженими ресурсами.

Запропонована система поєднує методи кластеризації, токенізації, ймовірнісної класифікації та глибокого аналізу за допомогою нейронних мереж типу *Bidirectional LSTM*. Проведене тестування показало здатність моделі зберігати високу точність визначення тональності стислих текстових повідомлень за мінімального навантаження на обчислювальні потужності.

Серед основних переваг розробленого підходу необхідно наголосити на можливості адаптації до різних типів текстових даних, зниженні витрат ресурсів без погіршення якості аналізу, а також на придатності для розгортання в середовищах з низькою обчислювальною спроможністю.

Надалі дослідження планується спрямувати на розширення функціональності системи за допомогою інтеграції оброблення мультимодальних даних, поглибленого використання зовнішніх баз знань і вдосконалення архітектур для роботи в режимі реального часу.

Список літератури

1. Petrov V.V., Zichun L., Kryuchyn A.A., Shanoylo S.M., Mingle F., Beliak I.V., Manko D.Y., Lapchuk A.S., Morozov E.M. Long-term storage of digital information. Akademperiodyka, Kyiv. 148 p. 2018. DOI: https://doi.org/10.15407/akademperiodyka. 360.148
2. Giordano V., Spada I., Chiarello F., Fantoni G. The impact of ChatGPT on human skills: A quantitative study on Twitter data. *Technological Forecasting and Social Change.* 2024. No. 203. 124 p. DOI: https://doi.org/10.1016/j.techfore.2024.123389
3. Dahri N. A., Extended Tam based acceptance of ai-powered ChatGPT for supporting metacognitive self-regulated learning in education: A mixed-methods study / Dahri N. A., Yahaya N., Al-Rahmi W. M., Aldraiweesh A., Alturki U., Almutairy S., Shutaleva A., Soomro R. B. *Heliyon.* 2024. No. 10(8). DOI: https://doi.org/10.1016/j.heliyon.2024.e29317
4. Malhotra A., Bajaj K. A hybrid pattern based text mining approach for malware detection using DBScan. *CSI Transactions on ICT, 4* (2-4), 2016. P. 141-149. DOI: https://doi.org/10. 1007/s40012-016-0095-y
5. The Trustees of Princeton University. What is WordNet? Princeton University. Retrieved January 12, 2022, URL: https://wordnet.princeton.edu.
6. Marcos T. Efficient Methods for Natural Language Processing: A Survey. / Marcos Treviso, Ji-Ung Lee, Tianchu Ji. et al. *Transactions of the Association for Computational Linguistics*, 11. 2023. P. 826–860. DOI: 10.1162/tacl_a_00577
7. Zhao W. X., A Survey of Large Language Models. / Zhao W. X., Zhou, K., Li, J. et al. *Computation and Language.* 144 p. 2023. DOI: https://doi.org/10.48550/arXiv.2303.18223
8. Lialin V., Deshpande V., Rumshisky A. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *Computation and Language.* 2023. DOI: https://doi.org/10.48550/arXiv.2303.15647

9. Yang J. A Survey of Knowledge Enhanced Pre-trained Models. / Yang J., Xiao G., Shen Y., Jiang W., Hu X., Zhang Y., Peng, J. et al. *Computation and Language.* 2021. 32 p. https://doi.org/10.48550/arXiv.2110.00269
10. Fournie Q., Caron G. M., Aloise D. A Practical Survey on Faster and Lighter Transformers. *ACM Computing Surveys, 55*(14s), P. 1-40. 2023. DOI: https://doi.org/10.1145/3586074
11. Zhang X. Edge intelligence optimization for large language model inference with batching and quantization. / Zhang X., Liu J., Xiong Z., Huang Y., Xie G., Zhang R. et al. *IEEE Wireless Communications and Networking Conference (WCNC).* 2024. DOI: https://doi.org/10. 1109/wcnc57260.2024.10571127
12. Wei X. Knowledge Enhanced Pretrained Language Models: A Compreshensive Survey. / Wei X., Wang S., Zhang D., Bhatia P., Arnold A.O. et al. *Computation and Language*. 2021. DOI: https://doi.org/10.48550/arXiv.2110.08455
13. Nagamatsu N., Hara-Azumi Y. Dynamic split computing-aware mixed-precision quantization for efficient deep edge intelligence. *IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom).* 2023. DOI: https://doi.org/10.1109/trustcom60117.2023.00355
14. Bao Y., Xu Y., Xiong H. Feature map alignment: Towards efficient design of mixed-precision quantization scheme. *IEEE Visual Communications and Image Processing (VCIP).* 2019. DOI: https://doi.org/10.1109/vcip47243.2019.8965724
15. Shamaeva I., Galley D. Simple and advanced Google Search. *Custom Search – Discover More*: P. 7–27. 2021. DOI: https://doi.org/10.1201/9781003 100133-2
16. Vo N.P., Popescu O. A multi-layer system for semantic textual similarity. *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management.* P. 56-67. 2016. DOI: https://doi.org/10.5220/0006045800560067
17. Yıldız E., Findik Y. Question similarity detection in Turkish using semantic textual similarity methods. *2019 27th Signal Processing and Communications Applications Conference (SIU)*. 2019. DOI: https://doi.org/10.1109/siu. 2019.8806308
18. Nel W., de Wet L., Schall R. Randomised controlled trial of the usability of major search engines (Google, Yahoo! and Bing) when using ambiguous search queries. *Proceedings of the 4th International Conference on Computer-Human Interaction Research and Applications.* P. 152-161. 2020. DOI: https://doi.org/10. 5220/0010133601520161
19. Oladipo F. O., Ohiani A. B.A. Text summarization system: An extractive approach using hierarchical text clustering. *International Journal of Computer Applications, 174* (23), P. 15–19. 2021. DOI: https://doi.org/10.5120/ijca202192 1015
20. Bindal A. Pathak A. A survey on K-means clustering and web-text mining. *International Journal of Science and Research (IJSR), 5*(4), P. 1049–1052. 2016. DOI: https://doi.org/10.21275/v5i4.nov162776
21. Shaposhnikov A. I. Feature-vector for the meanshift. *Proceedings of Tomsk State University of Control Systems and Radioelectronics, 24*(2), P. 34–38. 2021. DOI: https://doi.org/10.21293/ 1818-0442-2021-24-2-34-38
22. Tingting S. Application and research of DBSCAN optimization algorithm in big data analysis of experimental text. *Computer Science and Application, 10* (05), P. 906-913. 2020. DOI: https://doi.org/10.12677/csa.2020.105093
23. Otani N. Pre-tokenization of multi-word expressions in cross-lingual word embeddings / Otani N., Ozaki S., Zhao X., Li Y., St Johns M., Levin L. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* P. 4451–4464. 2020. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.360
24. Ribeiro E., Ribeiro R., de Matos D. A multilingual and Multidomain Study on Dialog Act recognition using character-level tokenization. *Information, 10*(3), 94 p. 2019. DOI: https://doi.org/10.3390/info10030094
25. Slimane F., Margner V. A new text-independent GMM writer identification system applied to Arabic handwriting. *International Conference on Frontiers in Handwriting Recognition.* 13 p. 2014. DOI: https://doi.org/10.1109/ icfhr.2014.124
26. Belogorskaya D. V. Summarizing news texts using quantitative methods (TF-IDF). *Proceedings of the VII (XXI) International Scientific and Practical Conference of Young Scientists.* 2020. DOI: https://doi.org/10.17223/978-5-94621-901-3-2020-30
27. Choi E.A., Han Y.E., Lee S., Oh M. A comparison of TF and TF-IDF analysis for trends of Blockchain in health and Welfare. *Journal of the Korean Data And Information Science Society, 30*(5), P. 1025–1036. 2019. DOI: https://doi.org/10.7465/jkdi.2019.30.5.1025
28. Ghawi R., Pfeffer J. Efficient hyperparameter tuning with grid search for text categorization using KNN approach with BM25 similarity. *Open Computer Science, 9*(1), P. 160–180. 2019. DOI: https://doi.org/10.1515/comp-2019-0011
29. Tinega G. A., Mwangi P. W., Rimiru D. R. Text mining in digital libraries using okapi BM25 model. *International Journal of Computer Applications Technology and Research, 7*(10), P. 398–406. 2018. DOI: https://doi.org/10.7753/ijcatr0710.1003
30. Ma X., Hovy E. End-to-end sequence labeling via bi-directional LSTM-cnns-CRF. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* P. 1064–1074. 2016. DOI: https://doi.org/10.18653/v1/p16-1101
31. Khalil F., Pipa P. D. Transforming the generative pretrained transformer into augmented business text writer. *CC BY 4.0.* 2021. DOI: https://doi.org/10.21203/rs.3.rs-1170589/v1
32. Soyalp G., Alar A., Ozkanli K., Yildiz B. Improving text classification with Transformer. *2021 6th International Conference on Computer Science and Engineering (UBMK)*. 12 p. 2021. DOI: https://doi.org/10.1109/ ubmk52708.2021.9558906

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 2 (32)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

33. Shen Y., Liu J. Comparison of text sentiment analysis based on Bert and word2vec. *2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC).* 17 p. 2021. DOI: https://doi.org/10.1109/icftic54370.2021.9647258
34. Pylkkönen J., Ukkonen A., Kilpikoski J., Tamminen S., Heikinheimo H. Fast text-only domain adaptation of RNN-Transducer Prediction Network. *Interspeech.* 2021. DOI: https://doi.org/10.21437/interspeech.2021-1191
35. Nismi Mol E. A., Santosh Kumar M. B. Study on impact of RNN, CNN and Han in text classification. *2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA).* 2020. DOI: https://doi.org/10.1109/accthpa49271.2020. 9213231
36. Zouzou A., Azami I. E. Text sentiment analysis with CNN & GRU model using glove. *2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS).* 2021. DOI: https://doi.org/10.1109/icds53782. 2021.9626715
37. Park P.W. Text-CNN based intent classification method for automatic input of intent sentences in chatbot. *The Journal of Korean Institute of Information Technology, 18* (1), P. 19–25. 2020. DOI: https://doi.org/10.14801/jkiit.2020. 18.1.19
38. Sun S., Gao Z., Huang C., Yu H. Glove-FRCNN: Comprehensive network algorithm for Vespa Mandarinia image-text extraction and classification. *2021 International Conference on Communications, Information System and Computer Engineering (CISCE).* 2021. DOI: https://doi.org/10.1109/ cisce52179.2021.9445902
39. Pan N., Yao W., Li X. Friends recommendation based on KBERT-CNN Text Classification Model. *International Joint Conference on Neural Networks (IJCNN).* 2021. DOI: https://doi.org/10.1109/ijcnn52387.2021. 9533618

References

1. Petrov, V. V., Zichun, L., Kryuchyn, A. A., Shanoylo, S. M., Mingle, F., Beliak, I. V., Manko, D. Y., Lapchuk, A. S., Morozov, E. M. (2018), *Long-term storage of digital information.* Akademperiodyka, Kyiv. 148 p. DOI: https://doi.org/10.15407/akademperiodyka. 360.148
2. Giordano, V., Spada, I., Chiarello, F., Fantoni, G. (2024), "The impact of ChatGPT on human skills: A quantitative study on Twitter data". *Technological Forecasting and Social Change.* No. 203. 124 p. DOI: https://doi.org/10.1016/j.techfore.2024.123389
3. Dahri, N. A., (2024), "Extended Tam based acceptance of ai-powered ChatGPT for supporting metacognitive self-regulated learning in education: A mixed-methods study" / Dahri, N. A., Yahaya, N., Al-Rahmi, W. M., Aldraiweesh, A., Alturki, U., Almutairy, S., Shutaleva, A., Soomro, R. B. *Heliyon.* 2024. No. 10(8). DOI: https://doi.org/10.1016/j.heliyon.2024.e29317
4. Malhotra, A., Bajaj, K. (2016), "A hybrid pattern based text mining approach for malware detection using DBScan". *CSI Transactions on ICT, 4* (2-4), P. 141-149. DOI: https://doi.org/10. 1007/s40012-016-0095-y
5. "The Trustees of Princeton University. What is WordNet? Princeton University". Retrieved January 12, 2022, available at: https://wordnet.princeton.edu.
6. Marcos, T. (2023), "Efficient Methods for Natural Language Processing: A Survey". / Marcos Treviso, Ji-Ung Lee, Tianchu Ji. et al. *Transactions of the Association for Computational Linguistics*, 11. P. 826–860. DOI: 10.1162/tacl_a_00577
7. Zhao, W. X.,(2023), "A Survey of Large Language Models". / Zhao W. X., Zhou, K., Li, J. et al. *Computation and Language.* 144 p. DOI: https://doi.org/10.48550/arXiv.2303.18223
8. Lialin, V., Deshpande, V., Rumshisky, A. (2023), "Scaling down to scale up: A guide to parameter-efficient fine-tuning. *Computation and Language.* DOI: https://doi.org/10.48550/arXiv.2303.15647
9. Yang, J. (2021), "A Survey of Knowledge Enhanced Pre-trained Models". / Yang J., Xiao G., Shen Y., Jiang W., Hu X., Zhang Y., Peng, J. et al. *Computation and Language.* 32 p. https://doi.org/10.48550/arXiv.2110.00269
10. Fournie, Q., Caron, G. M., Aloise, D. (2023), "A Practical Survey on Faster and Lighter Transformers. *ACM Computing Surveys, 55*(14s), P. 1-40. DOI: https://doi.org/10.1145/3586074
11. Zhang, X. (2024), "Edge intelligence optimization for large language model inference with batching and quantization". / Zhang X., Liu J., Xiong Z., Huang Y., Xie G., Zhang R. et al. *IEEE Wireless Communications and Networking Conference (WCNC).* DOI: https://doi.org/10. 1109/wcnc57260.2024.10571127
12. Wei, X. (2021), "Knowledge Enhanced Pretrained Language Models: A Compreshensive Survey". / Wei X., Wang S., Zhang D., Bhatia P., Arnold A.O. et al. *Computation and Language.* DOI: https://doi.org/10.48550/arXiv.2110.08455
13. Nagamatsu, N., Hara-Azumi, Y. (2023), "Dynamic split computing-aware mixed-precision quantization for efficient deep edge intelligence". *IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom).* DOI: https://doi.org/10.1109/trustcom60117.2023.00355
14. Bao, Y., Xu, Y., Xiong, H. (2019), "Feature map alignment: Towards efficient design of mixed-precision quantization scheme". *IEEE Visual Communications and Image Processing (VCIP).* DOI: https://doi.org/10.1109/vcip47243.2019.8965724
15. Shamaeva, I., Galley, D. (2021), "Simple and advanced Google Search". *Custom Search – Discover More*: P. 7–27. DOI: https://doi.org/10.1201/9781003 100133-2
16. Vo, N.P., Popescu, O. (2016), "A multi-layer system for semantic textual similarity". *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. P. 56-67. DOI: https://doi.org/10.5220/0006045800560067

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2025. No. 2 (32)*

17. Yıldız, E., Findik, Y. (2019), "Question similarity detection in Turkish using semantic textual similarity methods". *2019 27th Signal Processing and Communications Applications Conference (SIU)*. DOI: https://doi.org/10.1109/siu. 2019.8806308
18. Nel, W., de Wet, L., Schall, R. (2020), "Randomised controlled trial of the usability of major search engines (Google, Yahoo! and Bing) when using ambiguous search queries". *Proceedings of the 4th International Conference on Computer-Human Interaction Research and Applications*. P. 152-161. DOI: https://doi.org/10. 5220/0010133601520161
19. Oladipo, F. O., Ohiani, A. B.A. (2021), "Text summarization system: An extractive approach using hierarchical text clustering". *International Journal of Computer Applications, 174* (23), P. 15–19. DOI: https://doi.org/10.5120/ijca202192 1015
20. Bindal, A. Pathak, A. (2016), "A survey on K-means clustering and web-text mining". *International Journal of Science and Research (IJSR), 5*(4), P. 1049–1052. DOI: https://doi.org/10.21275/v5i4.nov162776
21. Shaposhnikov, A. I. (2021), "Feature-vector for the meanshift". *Proceedings of Tomsk State University of Control Systems and Radioelectronics, 24*(2), P. 34–38. DOI: https://doi.org/10.21293/ 1818-0442-2021-24-2-34-38
22. Tingting, S. (2020), "Application and research of DBSCAN optimization algorithm in big data analysis of experimental text". *Computer Science and Application, 10* (05), P. 906-913. DOI: https://doi.org/10.12677/csa.2020.105093
23. Otani, N. (2020), "Pre-tokenization of multi-word expressions in cross-lingual word embeddings" / Otani N., Ozaki S., Zhao X., Li Y., St Johns M., Levin L. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* P. 4451–4464. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.360
24. Ribeiro, E., Ribeiro, R., de Matos, D. (2019), "A multilingual and Multidomain Study on Dialog Act recognition using character-level tokenization". *Information, 10*(3), 94 p. DOI: https://doi.org/10.3390/info10030094
25. Slimane, F., Margner, V.(2014), "A new text-independent GMM writer identification system applied to Arabic handwriting". *International Conference on Frontiers in Handwriting Recognition.* 13 p. DOI: https://doi.org/10.1109/ icfhr.2014.124
26. Belogorskaya, D. V. (2020), "Summarizing news texts using quantitative methods (TF-IDF). *Proceedings of the VII (XXI) International Scientific and Practical Conference of Young Scientists.* DOI: https://doi.org/10.17223/978-5-94621-901-3-2020-30
27. Choi, E.A., Han, Y.E., Lee, S., Oh, M. (2019), "A comparison of TF and TF-IDF analysis for trends of Blockchain in health and Welfare". *Journal of the Korean Data And Information Science Society, 30*(5), P. 1025–1036. DOI: https://doi.org/10.7465/jkdi.2019.30.5.1025
28. Ghawi, R., Pfeffer, J. (2019), "Efficient hyperparameter tuning with grid search for text categorization using KNN approach with BM25 similarity". *Open Computer Science, 9*(1), P. 160–180. DOI: https://doi.org/10.1515/comp-2019-0011
29. Tinega, G. A., Mwangi, P. W., Rimiru, D. R. (2018), "Text mining in digital libraries using okapi BM25 model". *International Journal of Computer Applications Technology and Research, 7*(10), P. 398–406. DOI: https://doi.org/10.7753/ijcatr0710.1003
30. Ma, X., Hovy, E. (2016), "End-to-end sequence labeling via bi-directional LSTM-cnns-CRF". *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* P. 1064–1074. DOI: https://doi.org/10.18653/v1/p16-1101
31. Khalil, F., Pipa, P. D. (2021), "Transforming the generative pretrained transformer into augmented business text writer". *CC BY 4.0.* DOI: https://doi.org/10.21203/rs.3.rs-1170589/v1
32. Soyalp, G., Alar, A., Ozkanli, K., Yildiz, B. (2021), "Improving text classification with Transformer". *2021 6th International Conference on Computer Science and Engineering (UBMK).* 12 p. DOI: https://doi.org/10.1109/ ubmk52708.2021.9558906
33. Shen, Y., Liu J. (2021), "Comparison of text sentiment analysis based on Bert and word2vec". *2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC).* 17 p. DOI: https://doi.org/10.1109/icftic54370.2021.9647258
34. Pylkkönen, J., Ukkonen, A., Kilpikoski, J., Tamminen, S., Heikinheimo, H. (2021), "Fast text-only domain adaptation of RNN-Transducer Prediction Network". *Interspeech.* DOI: https://doi.org/10.21437/interspeech.2021-1191
35. Nismi, Mol, E. A., Santosh, Kumar M. B. (2020), "Study on impact of RNN, CNN and Han in text classification". *2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA).* DOI: https://doi.org/10.1109/accthpa49271.2020. 9213231
36. Zouzou, A., Azami, I. E. (2021), "Text sentiment analysis with CNN & GRU model using glove". *2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS).* DOI: https://doi.org/10.1109/icds53782. 2021.9626715
37. Park, P.W. (2020), "Text-CNN based intent classification method for automatic input of intent sentences in chatbot". *The Journal of Korean Institute of Information Technology, 18* (1), P. 19–25. DOI: https://doi.org/10.14801/jkiit.2020. 18.1.19
38. Sun, S., Gao, Z., Huang, C., Yu, H. (2021), "Glove-FRCNN: Comprehensive network algorithm for Vespa Mandarinia image-text extraction and classification". *2021 International Conference on Communications, Information System and Computer Engineering (CISCE).* DOI: https://doi.org/10.1109/ cisce52179.2021.9445902
39. Pan, N., Yao, W., Li, X. (2021), "Friends recommendation based on KBERT-CNN Text Classification Model". *International Joint Conference on Neural Networks (IJCNN).* 2021. DOI: https://doi.org/10.1109/ijcnn52387.2021. 9533618

Надійшла (Received) 05.04.2025

Сучасний стан наукових досліджень та технологій в промисловості. 2025. № 2 (32)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

Відомості про авторів / About the Authors

Шкурко Вячеслав Вікторович – Харківський національний університет радіоелектроніки, аспірант кафедри прикладної математики , Харків, Україна; e-mail: viacheslav.shkurko@nure.ua; ORCID ID: https://orcid.org/0009-0003-3274-3941

Поляков Андрій Олександрович – кандидат технічних наук, доцент, Харківський національний економічний університет імені Семена Кузнеця, доцент кафедри інформаційних систем, Харків, Україна; Харківський національний університет радіоелектроніки, доцент кафедри прикладної математики, Харків, Україна; e-mail: Andrii.Poliakov@nure.ua; ORCID ID: https://orcid.org/0000-0003-1805-9011

Shkurko Viacheslav – Kharkiv National University of Radio Electronics, Postgraduate Student Department of Applied Mathematics (AM), Kharkiv, Ukraine.

Poliakov Andrii – PhD (Engineering Sciences), Associate Professor, Simon Kuznets Kharkiv National University of Economics, Associated Professor at the Department of Information Systems, Kharkiv, Ukraine; Kharkiv National University of Radio Electronics, Associated Professor at the Department of Applied Mathematics, Kharkiv, Ukraine.

ORGANIZATION OF SOFTWARE AND NEURAL NETWORK ALGORITHMS FOR MACHINE ANALYSIS OF TEXTUAL DATA PRESENTED IN NATURAL LANGUAGE

The article addresses the problem of organizing software and neural network algorithms for natural language processing. Given the rapid growth of information flows and the limitations of computational resources, optimizing methods for processing short, unstructured messages and complex structured documents has become especially urgent. This study aims to develop a comprehensive method for organizing computer-aided analysis of text data, ensuring a balance between the accuracy of results and the efficiency of computational resource use. The research systematically examines modern approaches to tokenization, clustering, semantic-relevant search, and deep learning architectures, with particular attention to their adaptability under resource-constrained conditions. A multilevel methodology is suggested, combining the preliminary classification of text arrays, semantic clustering, and the use of a Bidirectional LSTM neural network model. Practical implementation of the method was tested through an automated text analysis application, demonstrating stable reduction in the loss function and acceptable resource consumption. The ability to adapt to different types of text data, reduced resource consumption while maintaining high-quality analysis, and suitability for deployment in environments with low computing capacity should be considered the main advantages of the developed approach. The scientific novelty of the article is substantiated by the integration of semantic-relevant clustering and lightweight deep learning techniques into a single optimized framework. The practical value of the study is the possibility of applying the proposed methodology in real-world information systems where limited hardware capabilities require efficient and adaptive text processing solutions.

Keywords: computer-aided analysis; text data; natural language; software algorithms; neural network architecture.

Бібліографічні описи / Bibliographic descriptions

Шкурко В. В., Поляков А. О. Організація програмних і нейромережевих алгоритмів машинного аналізу текстових повідомлень, поданих природною мовою. *Сучасний стан наукових досліджень та технологій в промисловості*. 2025. № 2 (32). С. 151–167. DOI: https://doi.org/10.30837/2522-9818.2025.2.151

Shkurko, V., Poliakov, A. (2025), "Organization of software and neural network algorithms for machine analysis of textual data presented in natural language", *Innovative Technologies and Scientific Solutions for Industries*, No. 2 (32), P. 151–167. DOI: https://doi.org/10.30837/2522-9818.2025.2.151