DOI: 10.58254/viti.9.2026.07.88

УДК 004.9:355.4

канд. техн. наук, доцент Данилюк І. А. ORCID: 0000-0003-0955-0108 (ВІТІ ім. Героїв Крут)
канд. техн. наук Мочалюк В. В. ORCID: 0009-0002-8389-7986 (ВІТІ ім. Героїв Крут)

РОЗРОБКА МЕТОДИКИ ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ ОБРОБКИ НЕСТРУКТУРОВАНОЇ ТЕКСТОВОЇ ІНФОРМАЦІЇ В ІНФОРМАЦІЙНИХ, ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ СИСТЕМАХ ОБОРОННОГО ПРИЗНАЧЕННЯ

У статті проведено комплексний аналіз сучасних методів обробки неструктурованої текстової інформації, зокрема: лінгвістичних, статистичних, векторних моделей представлення тексту, машинного навчання та глибокого навчання, послідовного аналізу, і визначено їхні переваги й обмеження в умовах функціонування інформаційно-аналітичних систем оборонного призначення. Показано, що в умовах зростання обсягів інформаційних потоків і необхідності їх обробки у реальному масштабі часу більшість традиційних підходів характеризуються значними вимогами до обчислювальних ресурсів та затримками прийняття рішень, що зумовлено необхідністю повного аналізу неструктурованої текстової інформації. Встановлено, що існуючі методи не забезпечують належного балансу між точністю класифікації та швидкістю обробки на етапі первинної обробки та забезпеченням спроможності до подальшого глибинного аналізу на подальших етапах, що є критичним для систем підтримки прийняття рішень у військовій сфері.

Запропоновано методику, яка базується на гібридному підході і забезпечує підвищення ефективності обробки текстових даних завдяки використанню багатоальтернативних послідовних вирішальних правил на етапі первинної обробки і трансформерних моделей для подальшого глибинного аналізу, що передбачає можливість раннього прийняття рішень на попередньому етапі та на неповній вибірці. Такий підхід дозволяє скорочувати обсяг оброблюваних даних та зменшувати час аналізу без суттєвого зниження достовірності результатів. На відміну від класичних методів класифікації, запропоноване рішення забезпечує гнучке управління процесом обробки неструктурованої вхідної текстової інформації.

Запропоновано застосування сховищ типу Data Lake для інтеграції різнорідних джерел інформації та забезпечення масштабованості системи.

Обґрунтовано, що впровадження запропонованої методики дозволяє підвищити оперативність обробки текстової інформації, знизити обчислювальні витрати та забезпечити необхідний рівень точності класифікації в умовах невизначеності та інформаційного перевантаження. Отримані результати можуть бути використані при створенні та модернізації інформаційно-аналітичних систем оборонного призначення, а також при інтеграції з існуючими платформами обробки даних у рамках сучасних військових стандартів.

Ключові слова: *неструктуровані дані, обробка тексту, класифікація, послідовний аналіз, інформаційні системи, оборонні системи, машинне навчання.*

I. Danyliuk, V. Mochaliuk. Development of a methodology for increasing the efficiency of processing unstructured text information in defense information, information and communication systems

The article provides a comprehensive analysis of modern methods for processing unstructured text information, including: linguistic, statistical, vector models of text representation, machine learning and deep learning, sequential analysis, and identifies their advantages and limitations in the conditions of functioning of information and analytical systems for defense purposes. It is shown that in the conditions of increasing volumes of information flows and the need for their processing in real time, most traditional approaches are characterized by significant requirements for computing resources and delays in decision-making, which is due to the need for a complete analysis of unstructured text information. It is established that existing methods do not provide a proper balance between classification accuracy and processing speed at the stage of primary processing and ensuring the ability to further in-depth analysis at subsequent stages, which is critical for decision-making support systems in the military sphere.

A method based on a hybrid approach is proposed that provides increased efficiency in text data processing through the use of multi-alternative sequential decision rules at the initial processing stage and transformer models for subsequent in-depth analysis, which provides the possibility of early decision-making at the preliminary stage and on an incomplete sample. This approach allows you to reduce the amount of processed data and reduce the analysis time without significantly reducing the reliability of the results. Unlike classical classification methods, the proposed solution provides flexible management of the process of processing unstructured input text information.

The use of Data Lake type storages is proposed to integrate heterogeneous information sources and ensure system scalability.

It is substantiated that the implementation of the proposed method allows you to increase the efficiency of text information processing, reduce computational costs and ensure the required level of classification accuracy in conditions of uncertainty and information overload. The results obtained can be used in the creation and modernization of

information and analytical systems for defense purposes, as well as in integration with existing data processing platforms within the framework of modern military standards.

Keywords: *unstructured data, text processing, classification, sequential analysis, information systems, defense systems, machine learning.*

Вступ

Сучасні інформаційно-аналітичні системи оборонного призначення функціонують в умовах стрімкого зростання обсягів даних, значна частина яких має неструктурований характер. Основними джерелами такої інформації є відкриті інформаційні ресурси, включаючи мережу Інтернет, соціальні мережі, новинні платформи та спеціалізовані канали комунікації.

В умовах впровадження принципів взаємосумісності з країнами-членами НАТО, зокрема в рамках концепцій Federated Mission Networking (FMN), Intelligence, Surveillance and Reconnaissance (ISR) та Command, Control, Communications, Computers, Intelligence, Surveillance and Reconnaissance (C4ISR), особливого значення набуває забезпечення швидкої первинної обробки інформації в режимі реального часу та подальший глибинний аналіз класифікованої інформації.

Проведений аналіз свідчить, що переважна частина інформаційних потоків представлена у вигляді неструктурованих та некласифікованих текстових даних. Більшість існуючих методів глибинного аналізу за допомогою трансформерних методів нейронних мереж великих мовних моделей (Large Language Model, LLM) передбачають необхідність врахування контексту, тобто визначення області знань (фокусу на темі) для прийняття коректного рішення, що у свою чергу потребує повної обробки тексту перед прийняттям рішення, це зумовлює значні обчислювальні витрати та часові затримки й ускладнює обробку інформаційних потоків із застосуванням традиційних підходів.

Постановка проблеми

Більшість сучасних методів обробки неструктурованої текстової інформації потребують значних обчислювальних ресурсів і передбачають повну обробку тексту перед прийняттям рішення, крім того такі методи потребують попередньо визначеної області знань неструктурованої текстової інформації (фокусу на темі) для уникнення фантомних (не коректних) висновків під час глибинного аналізу, що не дозволяє їх використовувати для швидкої первинної обробки інформації в режимі близькому до реального часу.

Таким чином, виникає науково-практична задача підвищення ефективності обробки неструктурованої текстової інформації завдяки створенню методики, на основі гібридного підходу з використанням швидких методів, які не потребують значних обчислювальних ресурсів для попередньої оперативної обробки різнорідних джерел неструктурованої текстової інформації у масштабі часу, близькому до реального з подальшим глибинним аналізом відібраних і класифікованих текстів за допомогою трансформерних методів нейронних мереж великих мовних моделей (LLM).

Аналіз останніх публікацій на тему дослідження дозволяє зазначити, що обробка неструктурованої текстової інформації є складною задачею, яка потребує застосування різних підходів залежно від обсягу даних, вимог до точності та обчислювальних ресурсів. Авторами статті були проаналізовані сучасні методи обробки неструктурованої текстової інформації, які можна умовно поділити на лінгвістичні, статистичні та методи машинного навчання.

Аналіз публікацій щодо лінгвістичних методів дозволяє зазначити, що лінгвістичні методи базуються на формальному аналізі природної мови та включають лексичний, морфологічний, синтаксичний і семантичний аналізи. Ці підходи широко застосовуються у задачах глибинного розуміння тексту та побудови знаннєво-орієнтованих систем.

Зокрема, у роботі [13] запропоновано знаннєво-орієнтований підхід до аналізу природномовної інформації, який забезпечує формалізацію предметної області та підвищення точності інтерпретації текстових даних.

Водночас, лінгвістичні методи характеризуються високою складністю реалізації та значною залежністю від якості мовних ресурсів.

Аналіз публікацій щодо статистичних методів дозволяє зазначити, що статистичні методи базуються на ймовірнісних моделях та використовують частотні характеристики тексту. Одним із базових підходів є наївний байєсівський класифікатор, який застосовується для задач фільтрації текстових повідомлень.

У роботах [3; 4] розглянуто застосування байєсівських методів для задач фільтрації небажаної інформації, де показано їхню ефективність при обробці великих обсягів текстових даних.

Подальший розвиток статистичних методів представлений у роботі [5], де запропоновано ймовірнісну модель класифікації текстів, що враховує структуру документів.

Перевагами статистичних методів є їхня обчислювальна ефективність та простота реалізації, однак вони обмежені припущенням незалежності ознак та недостатнім урахуванням контексту.

Аналіз публікацій щодо векторних моделей представлення тексту дозволяє зазначити, що у векторній моделі подання документів текст представлено у вигляді вектора в багатовимірному просторі термів [6; 8].

Цей підхід широко використовується в задачах інформаційного пошуку та класифікації, зокрема при роботі з колекціями текстів, такими як Reuters-21578 [7].

Основним недоліком векторних моделей є втрата семантичного контексту та залежність від методів зважування термів.

Аналіз публікацій щодо методів машинного навчання та глибокого навчання дозволяє зазначити, що більшість сучасних підходів до обробки текстових даних базуються на використанні трансформерних моделей, які використовують механізми попереднього визначення тематики – самоуваги. У роботі [15] запропоновано архітектуру Transformer, яка забезпечує високу точність обробки і можливість глибинного аналізу.

Такі методи забезпечують можливість глибинного аналізу текстових даних, але заразом потребують значних обчислювальних ресурсів і передбачають повну обробку тексту перед прийняттям рішення.

Аналіз публікацій щодо методів послідовного аналізу дозволяє зазначити, що послідовний аналіз можливо виділити в окремий клас методів, які базуються на послідовному аналізі, запропонованому у роботі [9]. Такі методи дозволяють приймати рішення на основі часткової інформації шляхом послідовного накопичення статистичних характеристик, це дозволяє використовувати методи послідовного аналізу у задачах обробки сигналів та розпізнавання образів [11–14], де реалізується відбір правдивих гіпотез із відкиданням неефективних варіантів.

Перевагою послідовного аналізу є можливість зменшення обсягу оброблюваних даних і скорочення часу прийняття рішень.

Таким чином серед сучасних методів обробки неструктурованої текстової інформації доцільно виділити трансформерні моделі машинного навчання та глибокого навчання, які забезпечують високу точність обробки, можливість глибинного аналізу і дозволяють моделювати складні семантичні зв’язки між фрагментами тексту для прогнозування, генерації та інтелектуальної інтерпретації контексту, але при цьому потребують попередньо визначеного фокусу (теми або направлення) неструктурованої текстової інформації, значних обчислювальних ресурсів і передбачають повну обробку тексту перед прийняттям рішення.

Іншою групою є методи послідовного аналізу, які дозволяють класифікувати – визначати фокус (тему або направлення) неструктурованої текстової інформації використовуючи

незначні обчислювальні ресурси і приймати рішення на неповній вибірці, але при цьому не дозволяють проводити глибинний аналіз тексту.

Із вищезазначеного виникає протиріччя: використовувати швидкі методи з незначними вимогами до ресурсів, але які не можуть проводити глибинний аналіз чи методи, які потребують значних обчислювальних ресурсів і передбачають повну обробку тексту, але при цьому дозволяють проводити глибинний аналіз, моделювати складні семантичні зв'язки між фрагментами тексту для прогнозування, генерації та інтелектуальної інтерпретації контексту.

Метою статті є розроблення методики підвищення ефективності методів обробки текстових даних за допомогою гібридного підходу з використання швидкого і невимогливого до ресурсів послідовного вирішального правила на етапі попередньої класифікації з подальшим наданням класифікованих даних до моделей, які дозволяють проводити глибинний аналіз, моделювати складні семантичні зв'язки між фрагментами тексту для прогнозування, генерації та інтелектуальної інтерпретації контексту.

Виклад основного матеріалу

Багатоальтернативне послідовне вирішальне правило для класифікації текстових документів із використанням верхніх порогів

У роботах [11–14] запропоноване багатоальтернативне послідовне вирішальне правило для класифікації текстових документів із використанням верхніх порогів.

У дискретному часі завдання послідовної класифікації текстових документів формулюється наступним чином. Маємо L простих гіпотез $C_l,\ \ l = \overline{1,L}$, що відображають теми, за якими проводиться класифікація. Кожен документ, що надходить до системи *dj*, представлений набором з N слів $W = (w_1, w_2 \ldots w_N)$, які характеризують його тематичну спрямованість.

Аналіз тексту розглядається як обробка потоку подій $w_i(k) \in N$, де N – множина словникових одиниць, що представляють зміст документа. Умовні ймовірності спостережень $P(w_i(k)|C_l)$, відомі з точністю до L гіпотез. Гіпотези є несумісними та утворюють повну групу подій. Апріорні ймовірності гіпотез $p_i,\ \ i = \overline{1,L}$ передбачаються відомими.

При заданих умовних ймовірностях помилкового розпізнавання гіпотез $P(\hat{C}_i|C_l),\ \ i \neq l,\ \ i,l = \overline{1,L}$ та ймовірностях правильного розпізнавання, не нижче заданих $P(\hat{C}_l|C_l) \geq P_{\text{зад}}(\hat{C}_l|C_l),\ \ l = \overline{1,L}$, послідовне вирішальне правило дозволяє в результаті спостереження реалізації $W(k),\ \ k = 1,2,3, \ldots$ прийняти одне з рішень.

Завдання багатоальтернативного розпізнавання L простих гіпотез зводиться до L двоальтернативних завдань перевірки простої C_l гіпотези проти складної альтернативи Θ_l.

Складна гіпотеза Θ_l є поєднанням простих гіпотез $C_i,\ \ i = \overline{1,L},\ \ i \neq l$:

$$\Theta_l = \bigcup_{\substack{i=1 \\ i \neq l}}^{L} C_i.$$

Геометрична інтерпретація складної гіпотези представлена на рисунку 1.

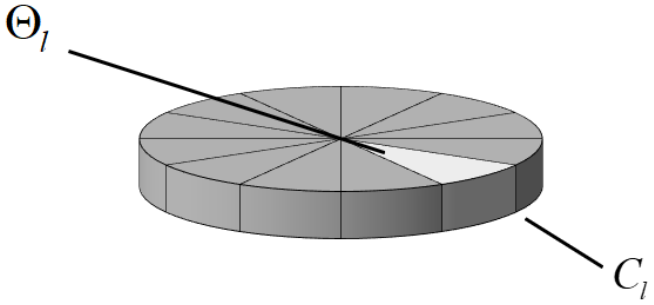

Рис. 1. Геометрична інтерпретація складної гіпотези

Прості гіпотези C_l є несумісними. Складні гіпотези Θ_l є сумісними.

Відповідно до послідовного вирішального правила на кожному *k*-му кроці визначаються L відносин правдоподібності перевірки простої гіпотези проти складної альтернативи у вигляді:

$$\Lambda_l(k) = \frac{P(W(k)|C_l)}{P(W(k)|\Theta_l)}, \quad l = \overline{1, L},$$

де $P(W(k)|C_l) = P(w_i(k)|C_l)P(W(k-1)|C_l)$ – функція правдоподібності простої гіпотези;

$P(W(k)|\Theta_l)$ – функція правдоподібності складної гіпотези Θ_l;

$W(k) = w_i(k), \ldots, w_i(k)$, $i = \overline{1, V}$, k = 1,2,3… – послідовність спостережень, що надходять.

Рішення на користь простої гіпотези C_l приймається при порівнянні відношення правдоподібності $\Lambda_l(k)$ з верхніми порогами $\Gamma_{1_l} l = \overline{1, l}$ (1):

$$\Lambda_l(k) = \frac{P(W(k)|C_l)}{\sum_{i=1, i\neq l}^{L} \gamma_{i|l} P(W(k)|C_i)} \geq \Gamma_{1_l}. \quad (1)$$

За умови, що на *k*-тому кроці відношення правдоподібності досягло значення верхнього порогу Γ_{1_l}, для гіпотези C_l, приймається гіпотеза C_l, і відхиляються інші $L-1$ гіпотези. Геометрична інтерпретація процедури ухвалення рішення при використанні верхніх порогів надана на рисунку 2.

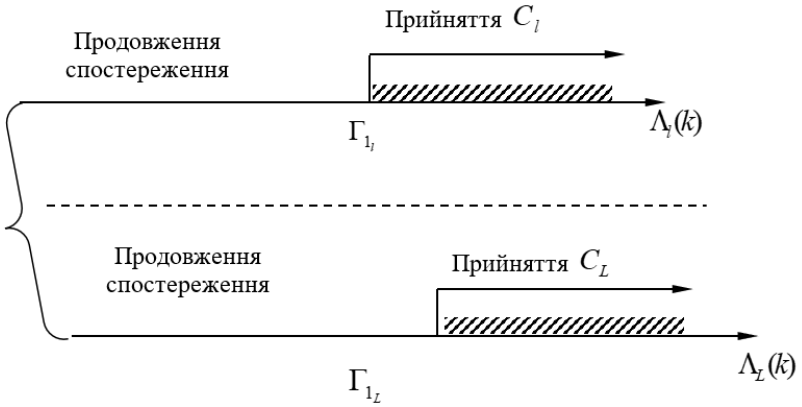

Рис. 2. Геометрична інтерпретація процедури прийняття рішення

Значення верхнього порога визначається за виразом (2):

$$\Gamma_{1_l} = \frac{P(\hat{C}_l|C_l)}{P(\hat{C}_l|\Theta_l)}. \quad (2)$$

У загальному випадку умовні ймовірності $P(\hat{C}_l|C_i)$ можливо представити у вигляді матриці:

$$P_{\text{реш}} = \begin{matrix} P(\hat{C}_1|C_1) & P(\hat{C}_1|C_2) & \ldots & P(\hat{C}_1|C_L) \\ P(\hat{C}_2|C_1) & P(\hat{C}_2|C_2) & \ldots & P(\hat{C}_2|C_L) \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ P(\hat{C}_L|C_1) & P(\hat{C}_L|C_2) & \ldots & P(\hat{C}_L|C_L) \end{matrix}$$

Номер рядка матриці відповідає рішенню, що приймається $\hat{C}_l,\ \ l=\overline{1,L}$, номер стовпця – істинній гіпотезі $C_i,\ i=\overline{1,L}$. Діагональні елементи матриці відповідають ймовірностям правильного прийняття рішень $P(\hat{C}_l|C_l), l=\overline{1,L}$. Недіагональні елементи кожного рядка використовуються для обчислення ймовірності хибного розпізнавання гіпотези $\hat{C}_l$ за умови, що має місце будь-яка інша проста гіпотеза. Недіагональні елементи кожного стовпця використовуються для визначення ймовірності нерозпізнавання гіпотези C_l:

$$P(\hat{\Theta}_l|C_l)=\sum_{i=1,i\neq l}^{L}P(\hat{C}_i|C_l)=1-P(\hat{C}_l|C_l), l=\overline{1,L}.$$

При дискретній вибірці може відбутися подія, коли достатня статистика перевищила значення верхнього порога. Однак якщо обсяг вибірки великий, перевищення порога досить малоймовірно, що призводить до незначної затримки у прийнятті рішення.

Аналіз правила було проведено на модельному прикладі у вигляді статистичного моделювання на ПЕОМ.

Дано: П'ять простих гіпотез $C_l,\ \ l=\overline{1,5}$, що відображають теми, за якими проводиться класифікація. Умовні ймовірності спостережень $P(w_i(k)|C_l),\ l=\overline{1,L}$ відомі з точністю до L гіпотез і наведені в таблиці 1.

Таблиця 1

Умовні ймовірності спостережень

$P(w_i \mid C_l)$	C_1	C_2	C_3	C_4	C_5
W_1	0,147	0,240	0,047	0,025	0,025
W_2	0,245	0,144	0,047	0,025	0,025
W_3	0,245	0,144	0,047	0,025	0,025
W_4	0,147	0,240	0,047	0,100	0,050
W_5	0,024	0,120	0,189	0,050	0,101
W_6	0,147	0,041	0,237	0,005	0,050
W_7	0,025	0,024	0,142	0,251	0,050
W_8	0,005	0,010	0,128	0,201	0,151
W_9	0,005	0,005	0,009	0,191	0,292
W_{10}	0,010	0,030	0,104	0,125	0,227

Гіпотези є несумісними та утворюють повну групу подій. Апріорні ймовірності гіпотез передбачалися рівноймовірними $p_i=0{,}2,\ i=\overline{1,5}$. Аналіз тексту розглядається як обробка потоку подій $w_i(k)\in N$, де N – множина словникових одиниць, що представляють зміст документа. Необхідно при заданих умовних ймовірностях помилкового розпізнавання гіпотез $P(\hat{C}_i|C_l),\ i\neq l,\ i,l=\overline{1,L}$ та ймовірності правильного розпізнавання, не менше заданих у результаті спостереження реалізації прийняти рішення на користь однієї з тематик $C_l,\ \ l=\overline{1,L}$.

Умовні ймовірності розпізнавання $P(\hat{C}_l|C_i),\ \ i,l=\overline{1,L}$ задані у вигляді матриці (3), отриманої з використанням алгоритму багатоальтернативного розпізнавання гіпотез на фіксованому інтервалі спостереження. Рішення на користь простої гіпотези приймалося наприкінці інтервалу спостереження за максимумом відношення правдоподібності. При моделюванні обсяг вибірки вважався рівним 25. Для кожної гіпотези проводилося 10 000 випробувань (3).

$$P(\hat{C}_l|C_i)=\begin{matrix}0{,}9640 & 0{,}0410 & 0{,}0001 & 0 & 0\\ 0{,}0360 & 0{,}9589 & 0{,}0002 & 0 & 0\\ 0 & 0{,}0001 & 0{,}9983 & 0{,}0002 & 0{,}0007\\ 0 & 0 & 0{,}0007 & 0{,}9723 & 0{,}0249\\ 0 & 0 & 0{,}0007 & 0{,}0275 & 0{,}9744\end{matrix} \quad (3)$$

Відповідно до матриці (3), за допомогою виразу (2), було визначено значення верхніх порогів послідовного вирішального правила (1).

$$\Gamma_{l_1}=93{,}82;\ \ \Gamma_{l_2}=105{,}96;\ \ \Gamma_{l_3}=3993{,}2;\ \ \Gamma_{l_4}=151{,}92;\ \ \Gamma_{l_5}=138{,}21\,.$$

На основі умовних імовірностей спостережень, представлених у таблиці 1, а також з урахуванням отриманих значень верхніх порогів, було проведено 10 000 випробувань для кожної з гіпотез. Довжина вибірки при кожному випробуванні дорівнювала 25.

Умовні ймовірності розпізнавання гіпотез на основі виразу (1), отримані шляхом статистичного моделювання, представлені у вигляді матриці (4):

$$P(\hat{C}_l|C_i)=\begin{matrix}0{,}9583 & 0{,}0481 & 0{,}0020 & 0 & 0\\ 0{,}0415 & 0{,}9517 & 0{,}0032 & 0{,}0002 & 0{,}0002\\ 0{,}0002 & 0{,}0002 & 0{,}9892 & 0{,}0003 & 0{,}0012\\ 0 & 0 & 0{,}0034 & 0{,}9691 & 0{,}0264\\ 0 & 0 & 0{,}0022 & 0{,}0304 & 0{,}9722\end{matrix}. \quad (4)$$

Графічна інтерпретація результатів моделювання представлена на рисунку 3.

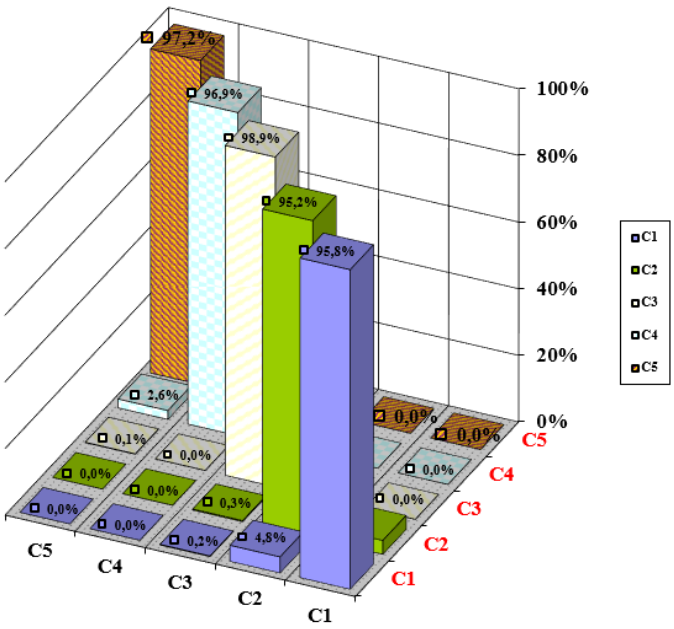

Рис. 3. Графічна інтерпретація результатів моделювання

Із рисунка видно, що показники розпізнавання гіпотез загалом не гірше ніж у алгоритму розпізнавання гіпотез на фіксованому інтервалі, але середня кількість кроків, необхідних для прийняття рішення, для усіх гіпотез дорівнює 13,6 (з вибірки на 25 слів).

Середня кількість кроків, необхідних для прийняття рішення $\bar{T}_l, l = \overline{1,5}$, для кожної гіпотези, і середньоквадратична помилка σ_l наведені у таблиці 2.

Таблиця 2

Результати моделювання з використанням верхніх порогів

Параметри	C_1	C_2	C_3	C_4	C_5
$\overline{T_l}$	13,395	13,634	14,046	13,166	13,759
σ_l	11,809	17,336	19,876	9,552	16,930

Із розглянутого прикладу видно, що алгоритм (1) забезпечує показники розпізнавання гіпотез (4) в цілому не гірше, ніж алгоритм розпізнавання гіпотез на фіксованому інтервалі (3). При цьому забезпечує зменшення часу спостереження на 45 відсотків.

Трансформери (Transformers)

Трансформери – це концепція архітектури нейронних мереж, таких як BERT і GPT, яка розроблена спеціально для роботи з послідовними даними. Трансформери використовуються для задач обробки природної мови Natural Language Processing (NLP). Трансформери забезпечили значний прорив у NLP завдяки своїй здатності ефективно враховувати контекст у послідовностях, проводити глибинний аналіз, моделювати складні семантичні зв'язки між фрагментами тексту для прогнозування, генерації та інтелектуальної інтерпретації контексту.

Ідея трансформерів вперше була запропонована у [14] дослідниками з Google. Головна ідея трансформерів полягає у використанні механізму самоуваги (Self-Attention).

Загальну архітектуру трансформерів можливо уявити у наступному вигляді:

1. Вхідні дані:

текст перетворюється у векторне представлення за допомогою вбудовувань (Embeddings);

додається позиційне кодування (Positional Encoding), щоб врахувати порядок слів.

2. Механізм самоуваги (Self-Attention):

дозволяє моделі "звертати увагу" на всі інші слова у тексті для кожного окремого слова;

враховує контекст кожного слова, незалежно від його позиції у послідовності.

3. Багатошарова архітектура:

трансформери складаються з N-енкодерів (encoder) та N-декодерів (decoder), що працюють спільно.

4. Головна увага (Multi-Head Attention):

забезпечує моделювання різних контекстів у тексті паралельно.

Методика підвищення ефективності обробки неструктурованої текстової інформації в інформаційних, інформаційно-комунікаційних системах оборонного призначення

Аналіз переваг і недоліків багатоальтернативного послідовного вирішального правила показав його високу точність, стійкість до шумових даних та здатність працювати з неповною вибіркою, але при цьому багатоальтернативне послідовне вирішальне правило не може забезпечити глибинний аналіз, з іншого боку трансформери можуть проводити глибинний аналіз, моделювати складні семантичні зв'язки між фрагментами тексту для прогнозування, генерації та інтелектуальної інтерпретації контексту, але для якісного глибинного аналізу їм необхідне визначення області знань (фокусу на темі), що потребує повного аналізу тексту і великих обчислювальних затрат.

Враховуючи, що в умовах реального часу для військових інформаційно-комунікаційних систем (C4ISR, ISR, FMN) критичними є швидкість обробки, масштабованість та інтеграція різнорідних джерел інформації, автори статті вважають доцільним сформулювати методику, яка забезпечить практичні рекомендації щодо підвищення ефективності систем обробки неструктурованої текстової інформації, які охоплюють архітектурні рішення, гібридну обробку, автоматизацію збору даних та оптимізацію алгоритмів.

Методика складається з наступних етапів:

Етап 1. Формування системи з розподіленими обчислюваннями на розподілених аналітичних вузлах.

Ефективна обробка текстових інформаційних потоків у військових системах потребує архітектури з розподіленими обчисленнями, що дозволить одночасно обробляти великі обсяги неструктурованої текстової інформації.

Враховуючи незначні вимоги багатоальтернативного послідовного вирішального правила до обчислювальних ресурсів автори пропонують розгорнути алгоритми попереднього аналізу і класифікації неструктурованої текстової інформації на розподілених аналітичних вузлах, через які проходять значні потоки відповідної інформації.

У такому разі час на прийняття рішення щодо класифікації і необхідності подальшого глибинного аналізу неструктурованого текстового потоку T_{decision} буде дорівнювати (5):

$$T_{\text{decision}} = \frac{T_{\text{seq}} + k_{\text{useful}} T_{\text{comm}}}{N_{\text{nodes}}}, \tag{5}$$

де T_{seq} – середній час послідовної обробки неструктурованого текстового потоку на окремому вузлі;

k_{useful} – коефіцієнт корисної інформації;

T_{comm} – середній час на передачу цільового класифікованого потоку до кластеру глибинного аналізу;

N_{nodes} – кількість вузлів попередньої обробки неструктурованого текстового потоку.

А коефіцієнт корисної інформації k_{useful} відповідно буде дорівнювати:

$$k_{\text{useful}} = \frac{V_{\text{useful}}}{V_{\text{seq}}},$$

де V_{seq} – середній об'єм загального неструктурованого текстового потоку;

V_{useful} – середній об'єм неструктурованого текстового потоку, визначеного для подальшого глибинного аналізу.

Враховуючи, що загальний неструктурований текстовий потік завжди буде більшим за текстовий потік, визначений для подальшого глибинного аналізу, то k_{useful} завжди буде меншим за 1, а відповідно вже на етапі попереднього аналізу на одному вузлі ми отримаємо виграш завдяки скороченню часу на передачу загального потоку і звільнення мережевого ресурсу для передачі відповідного потоку. Крім того, як видно з виразу (5), такий підхід скоротить середній час на прийняття рішення щодо класифікації і необхідності подальшого глибинного аналізу неструктурованого текстового потоку прямо пропорційно кількості вузлам, задіяним у попередньому аналізі.

Етап 2. Формування централізованого сховища типу Data Lake.

Надалі необхідно розгорнути централізоване сховище типу Data Lake, яке забезпечить зберігання неструктурованих даних, що надійшли від окремих вузлів і структурованих даних,

що з'явились в результаті глибинного аналізу. Крім того наявність централізованого сховища типу Data Lake є важливим для забезпечення масштабованості, адаптованості до зростання обсягів даних та кількості користувачів без втрати продуктивності, надійності та можливості інтеграції різнорідної інформації в рамках стандартів НАТО (FMN, ISR, C4ISR).

Етап 3. Формування розподіленого кластера глибокого аналізу.

Наступним етапом є формування розподіленого кластера глибокого аналізу, який забезпечить масштабованість, адаптованість та надійність функціонування алгоритмів трансформерів, для глибинного аналізу.

Час на проведення глибинного аналізу T_{deep} можливо розрахувати за виразом:

$$T_{\text{deep}} = \frac{T_{\text{class}}}{N_{\text{core}}},$$

де T_{class} – глибинний аналіз класифікованого потоку одним ядром;

N_{core} – кількість процесорних ядер, задіяних у глибинному аналізі.

Таким чином, на етапі глибинного аналізу ми отримаємо скорочення часу на загальний аналіз зворотно кількості процесорних ядер, задіяних у відповідному аналізі.

Таким чином, підвищення ефективності обробки текстових інформаційних потоків у військових системах можливо досягнути завдяки використанню методики, який поєднує масштабовану архітектуру з розподіленими обчисленнями, гібридну модель обробки даних, автоматизацію збору та адаптивне навчання алгоритмів.

Використання сховищ типу Data Lake та інтеграція з розподіленими обчислювальними ресурсами скорочують час прийняття рішень, а багаторівнева фільтрація і порогова оптимізація забезпечують високу точність аналізу навіть у присутності шумових або неповних даних.

Можливість адаптивного навчання, як алгоритмів послідовного вирішального правила, так і трансформерної моделі, дозволяє системі самостійно підлаштовуватися під змінний характер інформаційних потоків, а уніфікація форматів обміну та відповідність стандартам НАТО (FMN, ISR, C4ISR) забезпечить ефективну інтеграцію з союзними платформами та оперативну підтримку бойових рішень.

Запропонована методика формує стійку та надійну систему обробки неструктурованих текстових потоків, здатну діяти у реальному масштабі часу.

Наукова новизна отриманих результатів

На думку авторів, наукові результати проведених досліджень полягають у наступному:

удосконалено підхід до обробки неструктурованої текстової інформації шляхом попередньої класифікації текстового потоку за допомогою багатоальтернативних послідовних вирішальних правил із наступним наданням класифікованих документів до трансформерних моделей, що на відміну від існуючих методів, забезпечує можливість прийняття рішення на неповній вибірці даних;

набув подальшого розвитку метод класифікації текстових документів на основі ймовірнісної моделі подання даних завдяки застосуванню послідовного критерію відношень правдоподібності з верхніми порогами, що дозволяє скоротити середню кількість оброблюваних елементів тексту, а відповідно зменшити час прийняття рішення без погіршення точності класифікації;

вперше обґрунтовано доцільність використання гібридного підходу, який поєднує швидкі і невибагливі до обчислювальних ресурсів статистичні методи на етапі класифікації та фільтрації та трансформерні моделі на етапі глибокого аналізу, що забезпечує підвищення ефективності обробки великих потоків неструктурованої текстової інформації на

попередньому етапі в умовах обмежених обчислювальних ресурсів та дозволяє класифікувати інформаційні потоки для подальшого глибинного аналізу;

удосконалено модель організації процесу обробки даних в інформаційно-аналітичних системах завдяки впровадженню принципу раннього прийняття рішень, що дозволяє зменшити обчислювальне навантаження та підвищити оперативність обробки інформації.

Висновки й перспективи подальших досліджень

Таким чином, у статті проведено аналіз існуючих методів, їхніх переваг і обмежень з акцентом на якість і швидкості обробки неструктурованих текстових інформаційних потоків і вимог до обчислювальних ресурсів.

Розроблено формальний опис методики, підвищення ефективності обробки неструктурованих текстових інформаційних потоків у військових інформаційно-комунікаційних та інформаційних системах , зокрема в контексті стандартів НАТО (FMN, ISR, C4ISR).

Запропонована методика використовує сучасні підходи, формує гібридну модель обробки неструктурованих текстових інформаційних потоків і пропонує:

виконувати первинну класифікацію на розподілених аналітичних вузлах за допомогою багатоальтернативного послідовного вирішального правила, яке при незначних вимогах до обчислювальних ресурсів дозволить збільшити швидкості первинної класифікації зі збереженням якості прийняття рішень;

використовувати сховища типу Data Lake для подальшого зберігання інформаційно важливої, класифікованої складової неструктурованої текстової інформації, що дозволить інтегрувати різнорідні джерела інформації, забезпечити масштабованість системи та автоматизувати збір даних;

здійснювати глибинний аналіз класифікованої неструктурованої текстової інформації за допомогою кластеру трансформерних методів нейронних мереж великих мовних моделей (LLM), який забезпечить глибинний аналіз і високу точність обробки та дозволить моделювати складні семантичні зв'язки між фрагментами тексту для прогнозування, генерації та інтелектуальної інтерпретації контексту.

Підтримка балансу між швидкістю обробки та точністю прийняття рішення на первинному етапі обробки неструктурованого інформаційного потоку та можливістю подальшого глибинного аналізу відібраної цільової інформації є критично важливою для бойових умов.

Результати дослідження продемонстрували, що запропонований підхід дозволяє підвищити точність і стійкість системи до шумових і неповних даних, скоротити час обробки та забезпечити ефективний обмін інформацією між платформами відповідно до стандартів НАТО.

Автори вважають, що отримані результати можуть бути використані для розробки та вдосконалення інформаційно-комунікаційних та інформаційних систем оборонного призначення, зокрема:

1. Для систем ситуаційної обізнаності – забезпечення підвищення швидкості обробки інформаційних потоків, а відповідно зменшення часу доведення аналітичної інформації до керівництва.

2. Для підрозділів, що здійснюють OSINT-аналіз, – підвищення ефективності фільтрації інформації з відкритих джерел, а відповідно зменшення навантаження на аналітичний персонал.

3. Для інформаційно-комунікаційних, інформаційних систем ЗС України – забезпечення можливості обробки великих обсягів неструктурованих даних у реальному масштабі часу, а відповідно своєчасне отримання актуальної інформації;

4. Для розробників програмного забезпечення запропоновані підходи можуть бути використані при створенні модулів автоматичної класифікації та фільтрації текстових даних,

які забезпечать можливість побудови масштабованих рішень на основі Data Lake і розподілених обчислень.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Про рішення Ради національної безпеки і оборони України від 20 серпня 2021 року “Про Стратегічний оборонний бюлетень України”: Указ Президента України від 17.09.2021 № 473/2021. URL: https://www.president.gov.ua/documents/4732021-40121.
2. Жарков Я. М., Васильєв А. О. Наукові підходи щодо визначення суті розвідки з відкритих джерел // Вісник Київського національного університету імені Тараса Шевченка. Військово-спеціальні науки. 2013. Вип. 30. С. 38–41.
3. Graham P. Better Bayesian Filtering. 2003. URL: https://paulgraham.com/better.html.
4. Graham P. A Plan for Spam. 2002. URL: https://paulgraham.com/spam.html.
5. Kalt T. A New Probabilistic Model of Text Classification and Retrieval // Technical Report IR-78. University of Massachusetts Center for Intelligent Information Retrieval, 1996. URL: https://ciir.cs.umass.edu/pubfiles/ir-78.pdf.
6. Salton G., McGill M. J. Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1983. 448 p. URL: https://dl.acm.org/doi/book/10.5555/576628.
7. Reuters-21578 Text Categorization Collection. URL: https://daviddlewis.com/resources/testcollections/reuters21578/.
8. Salton G., Wong A., Yang C. S. A vector space model for automatic indexing. Communications of the ACM. 1975. № 18 (11). P. 613–620. URL: https://doi.org/10.1145/361219.361220.
9. Wald A. Sequential Analysis. Mineola, New York: Dover Publications, 1947. URL: https://books.google.com.ua/books?id=oVYDHHzZtdIC.
10. Васильєв В. И. Распознающие системы. Киев: Наукова думка, 1983. 422 с.
11. Жук С. Я., Ковальов В. В. Алгоритм совместной фильтрации речевого сигнала и оценки ошибки синхронизации в двухканальной измерительной системе // Известия высших учебных заведений. Радиоэлектроника. 2000. № 6. С. 16–21.
12. Жук С. Я., Ковальов В. В. Совместная фильтрация состояния и распознавание типа структуры динамической системы с отбрасыванием неудачных гипотез // Известия высших учебных заведений. Радиоэлектроника. 2001. № 7. С. 53–57.
13. Шемаев В. Н., Замаруева И. В., Приймак М. В., Дубровский Е. Н. Знание-ориентированный подход к анализу естественно-языковой текстовой информации в интересах мониторинга и оценки ситуаций // Інтелектуальний аналіз інформації (IAI-2003): зб. праць Третього наукового семінару (Київ, 21–23 травня 2003 р.). К.: Просвіта, 2004.
14. Дубровський Є. М. Метод сумісної послідовної класифікації і фільтрації текстових документів в автоматизованій системі пошуку розвідувальних відомостей // Робочі матеріали.
15. Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need // Advances in Neural Information Processing Systems. 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Надійшла до редколегії 19.03.2026.
Схвалена до друку 22.05.2026.
Дата публікації 29.05.2026.