

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Міністерство освіти і науки України

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Міністерство освіти і науки України

Кваліфікаційна наукова
праця на правах рукопису

БАРАН ДАНИЛО РОМАНОВИЧ

УДК 004.75 (004.62)

ДИСЕРТАЦІЯ

**МАТЕМАТИЧНІ МОДЕЛІ БАГАТОФАКТОРНОГО ПРОГНОЗУВАННЯ
АКТИВНОСТІ КОНТЕНТУ В ГЛОБАЛЬНИХ ІНФОРМАЦІЙНИХ
МЕРЕЖАХ**

123 Комп'ютерна інженерія

12 Інформаційні технології

Подається на здобуття наукового ступеня доктора філософії

Дисертація містить результати власних досліджень. Використання ідей,
результатів і текстів інших авторів мають посилання на відповідне джерело

Д.Р. Баран

Науковий керівник: ПИСАРЧУК Олексій Олександрович, д.т.н., професор

Київ – 2026

АННОТАЦІЯ

Баран Д.Р. Математичні моделі багатфакторного прогнозування активності контенту в глобальних інформаційних мережах. – Кваліфікаційна наукова робота на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 123 – Комп'ютерна інженерія з галузі знань 12 – Інформаційні технології. – Національний Технічний Університет України «Київський Політехнічний Інститут імені Ігоря Сікорського», Київ, 2025.

Дисертаційна робота присвячена розв'язанню *науково-прикладної задачі* з розробки і удосконалення математичних моделей багатфакторного аналізу та машинного навчання для обробки часових рядів.

Головне завдання та ідея дисертаційної роботи полягає у розробці та удосконаленні математичних моделей, що забезпечують послідовність етапів, спрямованих на виявлення критичної активності інформаційного контенту в глобальних медіа, здатних змінити настрої та призвести до небезпечних дій соціуму. Інформаційний контент, його обговорення, що присутні в глобальних інформаційних мережах як сайти новин, інформаційні канали месенджерів, соціальні мережі, підлягає постійному цілодобовому моніторингу. Результатом є Big Data масив повідомлень – даних, що зберігаються у відповідній базі даних. Первинні масиви даних підлягають кластеризації на соціальні сфери (економіка, політика, безпекова, духовна тощо). Для кожного повідомлення з відповідної сфери встановлюється його тональність: негатив, позитив, нейтральне. В подальшому розраховується частота повідомлень за тональністю для кожної сфери, що формує відповідні часові ряди. Надалі первинні часові ряди підлягають: OLAP (online analytical processing) візуалізації результатів кластеризації, попередній обробці, вторинній обробці з використанням статистичного та глибинного навчання і прогнозування. Поточні та прогнозовані результати приводяться до

інтегрованого показника за нелінійною схемою компромісів. Інтегрований показник аналізується на наявність біфуркаційного інтервалу (асимптотичне зростання), що може спричинити біфуркаційний перехід у настроях та діях соціуму під впливом негативних інформаційних повідомлень в різних сферах. Сумісний сигнатурний та окремий аналіз поведінки часових рядів з OLAP візуалізацією дозволяє оцінити причини появи біфуркаційних інтервалів – наслідків. Це може бути використано для своєчасної реакції у протидії негативним інформаційним повідомленням. Зазначені процеси реалізуються з використанням запропонованих математичних моделей кластеризації інформаційних повідомлень багатофакторного прогнозування активності контенту. Технологічно це реалізується на програмно-апаратному комплексі багатофакторного прогнозування активності контенту в глобальних інформаційних мережах, побудованому за принципами розподіленої комп'ютерної системи з відкритою архітектурою та із впровадженням принципів ситуаційного управління.

Використання запропонованих рішень забезпечує підвищення достовірності та повноти визначення домінантного інформаційного контенту в глобальних інформаційних мережах.

Результати роботи доведені до демонстраційного прототипу програмної системи, що доводить дієвість прийнятих рішень та отриманих науково-прикладних результатів.

Запропоновані рішення є доцільними для оперативного визначення соціальних настроїв різних сфер соціуму, фокус-групах, цільових аудиторій, що реалізуються: комерційними структурами в сфері надання послуг, реалізації продажів, тощо; в органах державного управління різної ланки – місцевого самоврядування, міністерств та відомств; в сфері безпеки і оборони.

Дисертаційні дослідження виконані відповідно до плану наукової роботи, затвердженому на кафедрі обчислювальної техніки КПІ ім. Ігоря Сікорського та у відповідності до положень Закону України «Про внесення

змін до деяких законів України щодо пріоритетних напрямів розвитку науки і техніки та інноваційної діяльності» – розвиток сучасних інформаційних, комунікаційних технологій, робототехніки, в межах науково-дослідної роботи «Нелінійні та багатокритеріальні математичні моделі для технологій Data Science та Embedded Systems», державний реєстраційний номер: 0124U003323. Практична цінність роботи підтверджена актами впровадження (див. Додаток А).

Мета дисертаційних досліджень. Підвищення достовірності (точності оцінювання і прогнозу) та повноти (множини врахованих факторів) визначення домінантного інформаційного контенту в глобальних інформаційних мережах.

Для досягнення сформованої мети в дисертаційній роботі було вирішено комплекс взаємопов'язаних задач досліджень:

- аналіз існуючих підходів до прогнозування інформаційного контенту в глобальних інформаційних мережах;
- розробка математичної моделі кластеризації інформаційних повідомлень з індикаторами активності інформаційного контенту за тональністю і сферами діяльності соціуму та ієрархічною компресією вхідних Big Data масивів;
- розробка математичної моделі багатфакторного прогнозування активності контенту в глобальних інформаційних мережах з обробкою та аналізом часових рядів з використанням методів статистичного та глибинного машинного навчання;
- розробка технології багатфакторного прогнозування активності контенту в глобальних інформаційних мережах, яка базується на реалізації математичних моделей кластеризації інформаційних повідомлень та багатфакторного прогнозування активності контенту на спеціалізованому програмно-апаратному комплексі, побудованому

за принципами розподіленої комп'ютерної системи з відкритою архітектурою;

- розробка макету програмно-апаратного комплексу прогнозування активності контенту в глобальних інформаційних мережах та оцінювання ефективності запропонованих рішень.

Наукова новизна отриманих результатів дисертаційних досліджень.

1. Отримала подальший розвиток математична модель кластеризації інформаційних повідомлень, яка базується на частотному аналізі їх тональності з використанням методологій обробки природньої мови (Natural Language Processing (NLP)) з підтримкою великих мовних, моделей; OLAP візуалізації результатів кластеризації та вирізняється встановленою системою індикаторів активності інформаційного контенту за сферами діяльності соціуму з ієрархічною компресією вхідних Big Data масивів. Це забезпечує підвищення повноти аналізу інформаційних повідомлень в глобальних інформаційних мережах шляхом врахування множини факторів за сферами діяльності соціуму.

2. Вперше запропоновано математичну модель багатofакторного прогнозування активності контенту в глобальних інформаційних мережах, яка базується на обробці та аналізі часових рядів з використанням методів статистичного та глибинного машинного навчання і відрізняється від відомих рішень:

виявленням і очищенням вхідних даних від аномальних вимірів з оперативним аналізом динамічних та статистичних властивостей вимірів в ковзному вікні;

визначенням параметрів нелінійних моделей з балансом диференціальних спектрів для методів статистичного навчання;

модифікацією штучних нейронних мереж для обробки часових рядів із впровадженням нелінійної схеми компромісів у вхідному прошарку та формування навчальної вибірки для екстраполяції за методами статистичного навчання;

визначенням інтегрованої оцінки активності інформаційного контенту за нелінійною схемою компромісів як результат багатфакторного аналізу вектору факторів за сферами діяльності соціуму;

пошуком прогнозованих біфуркаційних інтервалів як критичної активності інформаційного контенту;

Це забезпечує підвищення достовірності (ймовірності виявлення і точності прогнозу) домінантної активності контенту в глобальних інформаційних мережах.

3. Вперше розроблено технологію багатфакторного прогнозування активності контенту в глобальних інформаційних мережах, яка базується на реалізації запропонованих математичних моделей кластеризації інформаційних повідомлень та багатфакторного прогнозування активності контенту на спеціалізованому програмно-апаратному комплексі, побудованому за принципами розподіленої комп'ютерної системи з відкритою архітектурою та відрізняється впровадженням архітектурних рішень програмної компоненти з імплементацією розроблених математичних моделей кластеризації та багатфакторного прогнозування інформаційних повідомлень. Використання розробленої технології забезпечує підвищення достовірності та повноти визначення домінантного інформаційного контенту в глобальних інформаційних мережах.

Ключові слова: математична модель, машинне навчання, статистичне навчання, багатокритеріальна оптимізація, кластеризація, штучні нейронні мережі, комп'ютерна система, дослідження даних / Data Science, часові ряди / Time Series, обробка природньої мови / (NLP), великі дані / Big Data.

ABSTRACT

Baran D.R. Mathematical models of multifactorial forecasting of content activity in global information networks. – Qualification scientific work as a manuscript.

Dissertation for the degree of Doctor of Philosophy in the specialty 123 – Computer Engineering in the field of knowledge 12 – Information Technologies. – National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, 2025.

The dissertation is devoted to the solution *of a scientific and applied problem* for the development and improvement of mathematical models of multivariate analysis and machine learning for processing time series.

The main task and idea of the dissertation is to develop and improve mathematical models that provide a sequence of stages aimed at identifying the critical activity of information content in global media that can change moods and lead to dangerous actions of society. Information content, its discussions, which are present in global information networks such as news sites, information channels of messengers, social networks, are subject to constant round-the-clock monitoring. The result is a Big Data array of messages – data stored in the corresponding database. Primary data sets are subject to clustering into social spheres (economy, politics, security, spiritual, etc.). For each message from the corresponding sphere, its tone is established: negative, positive, neutral. Subsequently, the frequency of messages by tone for each sphere is calculated, which forms the corresponding time series. In the future, primary time series are subject to: OLAP (online analytical processing) visualization of clustering results, pre-processing, secondary processing using statistical and deep learning and forecasting. Current and projected results are brought to an integrated indicator according to a nonlinear scheme of compromises. The integrated indicator is analyzed for the presence of a bifurcation interval (asymptotic growth), which can

cause a bifurcation transition in the moods and actions of society under the influence of negative information messages in various areas. Compatible signature and separate analysis of the behavior of time series with OLAP visualization allows you to assess the causes of bifurcation intervals - consequences. This can be used for a timely response in countering negative information messages. These processes are implemented using the proposed mathematical models of clustering information messages, multifactorial forecasting of content activity. Technologically, this is implemented on a software and hardware complex for multifactorial forecasting of content activity in global information networks, built on the principles of a distributed computer system with an open architecture and with the introduction of the principles of situational management.

The use of the proposed solutions ensures an increase in the reliability and completeness of the definition of dominant information content in global information networks.

The results of the work were brought to a demonstration prototype of the software system, which proves the effectiveness of the decisions made and the obtained scientific and applied results.

The proposed solutions are expedient for prompt determination of social moods of various spheres of society, focus groups, target audiences, which are implemented by: commercial structures in the field of providing services, sales, etc.; in public administration bodies of various levels – local self-government, ministries and departments; in the field of security and defense.

The dissertation research was carried out in accordance with the plan of scientific work approved at the Department of Computer Engineering of Igor Sikorsky Kyiv Polytechnic Institute. Igor Sikorsky and in accordance with the provisions of the Law of Ukraine "On Amendments to Certain Laws of Ukraine on Priority Areas of Development of Science and Technology and Innovative Activity" – the development of modern information, communication technologies, robotics, within the framework of research work "Nonlinear and Multi-Criteria Mathematical Models for Data Science and Embedded Systems Technologies",

state registration number: 0124U003323. The practical value of the work is confirmed by the implementation acts (see Appendix A).

The purpose of dissertation research. Increasing the reliability (accuracy of assessment and forecast) and completeness (a set of factors taken into account) of determining the dominant information content in global information networks.

To achieve the formed goal in the dissertation, a set of interrelated research tasks was solved:

- analysis of existing approaches to forecasting information content in global information networks;
- development of a mathematical model for clustering information messages with indicators of information content activity by tone and areas of activity of society and hierarchical compression of incoming Big Data arrays;
- development of a mathematical model for multifactorial forecasting of content activity in global information networks with processing and analysis of time series using statistical and deep machine learning methods;
- development of a technology for multifactorial forecasting of content activity in global information networks, which is based on the implementation of mathematical models for clustering information messages and multifactor forecasting of content activity on a specialized software and hardware complex built on the principles of a distributed computer system with an open architecture;
- development of a layout of a software and hardware complex for forecasting content activity in global information networks and evaluating the effectiveness of the proposed solutions.

Scientific novelty of the results of dissertation research.

1. The mathematical model of clustering of information messages has been further developed, which is based on the frequency analysis of their tonality using Natural Language Processing (NLP) methodologies with the support of large language models; OLAP visualization of clustering results and is distinguished by

an established system of indicators of information content activity by areas of activity of society with hierarchical compression of incoming Big Data arrays. This provides an increase in the completeness of the analysis of information messages in global information networks by taking into account many factors in the areas of activity of society.

2. For the first time, a mathematical model of multifactorial forecasting of content activity in global information networks is proposed, which is based on the processing and analysis of time series using statistical and deep machine learning methods and differs from the well-known solutions:

- detection and clearance of input data from abnormal measurements with operational analysis of dynamic and statistical properties of measurements in the sliding window;

- determination of parameters of nonlinear models with a balance of differential spectra for statistical training methods;

- modification of artificial neural networks for processing time series with the introduction of a nonlinear scheme of trade-offs in the input layer and the formation of a training sample for extrapolation by statistical learning methods;

- determination of an integrated assessment of the activity of information content according to a non-linear scheme of compromises as a result of a multifactorial analysis of the vector of factors by areas of activity of society;

- search for predicted bifurcation intervals as a critical activity of information content;

This provides an increase in the reliability (probability of detection and accuracy of forecasting) of the dominant activity of content in global information networks.

3. For the first time, the technology of multifactorial forecasting of content activity in global information networks has been developed, which is based on the implementation of the proposed mathematical models of clustering of information messages and multifactor forecasting of content activity on a specialized software and hardware complex built on the principles of a distributed computer system with

an open architecture and is distinguished by the implementation of architectural solutions of the software component with implementation of the developed mathematical models of clustering and multifactor forecasting of information messages. The use of the developed technology provides an increase in the reliability and completeness of the definition of dominant information content in global information networks.

Keywords: mathematical model, machine learning, statistical learning, multi-criteria optimization, clustering, artificial neural networks, computer system, data science, time series, natural language processing, big data.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА

Наукові праці в яких опубліковано основні наукові результати дисертації:

1. Algorithms of Statistical Anomalies Clearing for Data Science

Applications - Oleksii Pysarchuk, Yurii Mironov, Illya Pysarchuk, Danylo Baran. // System research and information technologies. – No. 1 (2023), P.78-84. <https://doi.org/10.20535/SRIT.2308-8893.2023.1.06>, <http://journal.iasa.kpi.ua/article/view/260175>. (Стаття Scopus, категорії A).

Здобувач запропонував критерії виявлення аномалій та їх реалізації в ковзному вікні; Міронов Ю. – провів систематизацію отриманих результатів; Писарчук І. – провів експериментальні дослідження запропонованих підходів. Писарчук О. здійснив загальне керівництво та формування стратегії досліджень.

2. O. Pysarchuk, T. Andreieva, O. Grinenko, D. Baran. Multi-factor forecasting of statistical trends for data science problems // System Research & Information Technologies, 2024, No 2, P.21-34.

<https://doi.org/10.20535/SRIT.2308-8893.2024.2.02> <http://journal.iasa.kpi.ua/article/view/287739>. (Стаття Scopus, категорії A).

Здобувач розробив математичну модель і технологію багатфакторного прогнозування поведінки часових рядів; Грінченко О. – провела систематизацію отриманих результатів; Андреєва Т. – здійснила реалізацію та експериментальне дослідження сформованих рішень, провела збір вхідних даних для економічної сфери. Писарчук О. здійснив загальне керівництво та формування стратегії досліджень стосовно багатокритіального оцінювання часових рядів.

3. O. O Pysarchuk, O. A. Tuhanskykh, D. R. Baran (2025). Method for determining the structure of nonlinear models for time series processing. Radio Electronics, Computer Science, Control, 1, pp. 53-62 DOI: 10.15588/1607-3274-2025-1-5, <https://ric.zp.edu.ua/article/view/324253>, ISSN: 2313-688X, 1607-3274 | <https://ric.zp.edu.ua/article/view/324253>; (Стаття Web of Science, категорії A).

Здобувачу належить розвиток підходу нелінійного оцінювання часових рядів в розширені типів нелінійних моделей; Туганських О. – запропонував ознаки для вибору типу моделі, провів експериментальні дослідження; Писарчук О. здійснив загальне керівництво та формування стратегії розвитку методів нелінійного статистичного навчання.

4. Baran D., Pysarchuk O., Pavlova S (2025). The method of adaptation of the parameters of algorithms for the detection and cleaning of a statistical sample from anomalies for data science problems. Radio Electronics, Computer Science, Control, 3, pp. 37-44 DOI: 10.15588/1607-3274-2025-3-4, <https://ric.zp.edu.ua/article/view/339011>, ISSN: 2313-688X, 1607-3274 | <https://ric.zp.edu.ua/article/view/339011> (Стаття Web of Science, категорії A).

Здобувач запропонував реалізувати самонавчання алгоритму виявлення аномалій, відповідно до властивостей часових рядів; Павлова О. – сформувала ознаки часового ряду для самонавчання алгоритму виявлення аномалій, провела експериментальні дослідження; Писарчук О. здійснив загальне керівництво розробкою, систематизація результатів досліджень.

5. Baran D. Pysarchuk O., Pysarchuk I., Vasylieva M. (2025). Multi-criteria mathematical model of credit scoring in data science problems. System research and information technologies, No. 3 (2025), pp. 99-112. <http://journal.iasa.kpi.ua/article/view/308786>. (Стаття Scopus, категорії A).

Здобувач розробив математичну модель багатofакторного оцінювання; Васильєва М. – провела систематизацію результатів досліджень. Писарчук І. – здійснив практичну реалізацію моделі та її апробацію на реальних даних; Писарчук О. здійснив загальне керівництво розробкою, систематизація результатів досліджень.

6. Баран Д.Р., Писарчук О.О., Корочкін О.В., Визначення порядку поліноміальної моделі для побудови лінії тренду в задачах Data Science // Проблеми інформатизації та управління, 3(71), 2022/ – С.35-40. <https://doi.org/10.18372/2073-4751.71.17001> <https://jrnl.nau.edu.ua/index.php/PIU/article/view/17001> (Стаття фахова, категорії Б).

Здобувачу належить алгоритм визначення порядку поліноміальної моделі; Корочкін О. – провів аналіз існуючих підходів, систематизацію результатів досліджень.; Писарчук О. здійснив загальне керівництво розробкою.

7. Писарчук, О., Баран, Д., & Туганських, О. (2023). Методика статистичного навчання параметрів експоненційної математичної моделі. Measuring and computing devices in technological processes, (3), 224–227. <https://doi.org/10.31891/2219-9365-2023-75-26> <https://vottp.khmnu.edu.ua/index.php/vottp/article/view/177> (Стаття фахова, категорії Б).

Здобувачу належить методика статистичного навчання нелінійних моделей; Туганських О. – експериментальні розрахунки; Писарчук О. здійснив загальне керівництво розробкою, систематизація результатів досліджень.

8. Baran D., Pysarchuk O. (2025). Mathematical Model of Clustering of Informational Messages with Indicators of Activity for the Information Content by Tone and Areas of Society Activity. Information, Computing and Intelligent systems, 7, pp 39-48 DOI: <https://doi.org/10.20535/2786-8729.6.2025.339127>, <https://itvisnyk.kpi.ua/article/view/339127> (Стаття

фахова, категорії Б).

Здобувачу належить математична модель кластеризації інформаційних повідомлень; Писарчук О. здійснив загальне керівництво розробкою, систематизація результатів досліджень.

9. Pysarchuk O., Baran D. (2025). Optimization neural network for time series processing. Information, Computing and Intelligent systems, 6, pp 230-241 DOI: <https://doi.org/10.20535/2786-8729.7.2025.341480>, <https://itvisnyk.kpi.ua/article/view/341480> (Стаття фахова, категорії Б).
Здобувачу належить оптимізаційна нейронна мережа, технологія прогнозування часових рядів; Писарчук О. здійснив загальне керівництво розробкою, систематизація результатів досліджень.

Апробація наукових результатів дисертації:

10. Konrad, T., Pysarchuk, O., Noga, H., Gajewska, A., Volkogon, V., Baran, D. (2024). Method of Multifactor Analysis of the Alternatives Using the Example of Multi-criterial Selection of the Optimal Cargo Transportation Route. In: Yang, XS., Sherratt, S., Dey, N., Joshi, A. (eds) Proceedings of Ninth International Congress on Information and Communication Technology. ICICT 2024 2024. Lecture Notes in Networks and Systems, vol 1002. Springer, Singapore. https://doi.org/10.1007/978-981-97-3299-9_6 https://link.springer.com/chapter/10.1007/978-981-97-3299-9_6#citeas
Здобувачу належить математичну модель багатofакторного оцінювання; Нога Х, Гаєвська А. – провели систематизацію результатів досліджень; Конрад Т. – здійснила практичну реалізацію моделі та її апробацію на реальних даних; Писарчук О. здійснив загальне керівництво розробкою, систематизація результатів досліджень.
11. Tuhanskyk, O., Baran, D., Pysarchuk, O. (2024). Method for Statistical Evaluation of Nonlinear Model Parameters in Statistical Learning Algorithms. In: Yang, XS., Sherratt, S., Dey, N., Joshi, A. (eds)

Proceedings of Ninth International Congress on Information and Communication Technology. ICICT 2024 2024. Lecture Notes in Networks and Systems, vol 1013. Springer, Singapore. https://doi.org/10.1007/978-981-97-3559-4_21 https://link.springer.com/chapter/10.1007/978-981-97-3559-4_21

Здобувачу належать ідея та реалізація нелінійного оцінювання часових рядів в розширені типів нелінійних моделей; Туганських О. – запропонував конкретний тип нелінійних моделей, провів експериментальні дослідження; Писарчук О. здійснив загальне керівництво та формування стратегії розвитку методів нелінійного статистичного навчання.

12. Pysarchuk, O., Tuhanskykh, O., Baran, D. (2025). Structural Adaptation of Nonlinear Models in Time Series Processing Problems. In: Zgurovsky, M., Pankratova, N. (eds) System Analysis and Data Mining. Studies in Systems, Decision and Control, vol 609. Springer, Cham. ISBN 978-3-031-97529-5 https://doi.org/10.1007/978-3-031-97529-5_3/ (Розділ монографії Scopus).

Здобувачу належить розвиток методу нелінійного моделювання для статистичного навчання в напрямку обумовленого синтезу нелінійних моделей; Туганських О. – запропонував структуру композитних моделей та реалізував їх дослідження; Писарчук О. здійснив загальне керівництво та формування стратегії розвитку методів нелінійного статистичного навчання.

13. Метод аналітичного визначення параметрів нелінійних моделей та статистичною вибіркою вимірів – Баран Д., Туганських О., Писарчук О. // Міжнародна науково-практична конференція молодих вчених та студентів «Інженерія програмного забезпечення і передові інформаційні технології» (SoftTech-2023) Кафедра ІІІ ФІОТ, КПІ ім. Ігоря Сікорського, 2023, <https://ela.kpi.ua/handle/123456789/59987> <https://ela.kpi.ua/items/f02db5cb-3489-4c68-941e-179fdca13cbe/full>

Здобувачу належить розвиток методу нелінійного оцінювання часових

рядів; Туганських О. здійснив синтез конкретних моделей, провів експериментальні дослідження; Писарчук О. здійснив загальне керівництво.

14. Algorithms for purification of statistical data samples from anomalies for Data Science tasks – Baran D., Pysarchuk O. // The International Conference on Security, Fault Tolerance, Intelligence” (ICSFTI2022 Online).

Здобувачу належить розвиток методу нелінійного оцінювання часових рядів в класі експоненційних та трансцендентних моделей; Писарчук О. здійснив загальне керівництво.

15. Oleksii Pysarchuk, Oleksandr Tuhanskykh, and Danylo Baran Structural Adaptation of Nonlinear Models in Time Series Processing Problems // SAIC'2024 (IEEE International Conference on System Analysis & Intelligent Computing), http://saic.ieee.org.ua/?page_id=96

Здобувач реалізував конкретний клас нелінійних моделей в методі статистичного навчання; Туганських О. провів модельний експеримент, систематизував матеріали; Писарчук О. здійснив загальне керівництво.

16. Hrishchenko, K., Pysarchuk, O., Baran, D. (2026). Divide and Conquer Method for Constrained Programming Applied to Large Scale Production Scheduling Problems. In: Choudrie, J., Tuba, E., Perumal, T., Joshi, A. (eds) ICT for Intelligent Systems. ICTIS 2025. Smart Innovation, Systems and Technologies, vol 124. Springer, Singapore.
https://link.springer.com/chapter/10.1007/978-981-95-1353-6_9

Здобувач реалізував моделювання в класі трансцендентних унітарних функцій; Грищенко К. провів модельний експеримент, із трансцендентними комбінаторними функціями; Писарчук О. здійснив загальне керівництво, систематизацію результатів досліджень.

17. Pysarchuk O., Tuhanskykh O., and Baran D. Methodology for statistical learning of nonlinear model parameters with the integral form of ols in the

scheme of differential-nontaylor transformations // Security, Fault Tolerance, Intelligence: proceedings of the VIII International Conference ICSFTI2025, Kyiv, Ukraine, may 14, june 30, 2025. – Kyiv : Igor Sikorsky Kyiv Polytechnic Institute, publishing house “Polytechnica”, 2025.

<https://icsfti-proc.kpi.ua/article/view/338732>

Здобувачу належить розвиток методу нелінійного оцінювання часових рядів в класі експоненціальних та трансцендентних моделей;

Туганських О. провів модельний експеримент; Писарчук О. здійснив загальне керівництво.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ	22
ВСТУП	23
 РОЗДІЛ 1. АНАЛІЗ ІСНУЮЧИХ ПІДХОДІВ ДО ПРОГНОЗУВАННЯ ІНФОРМАЦІЙНОГО КОНТЕНТУ В ГЛОБАЛЬНИХ ІНФОРМАЦІЙНИХ МЕРЕЖАХ	 30
1.1. Аналіз особливостей розповсюдження контенту в глобальних інформаційних мережах та моніторингу його активності	 30
1.2. Методи кластеризації інформаційних повідомлень за принципами обробки природньої мови (NLP-процеси)	40
1.3. Аналіз існуючих методологічних підходів до прогнозування активності контенту в глобальних інформаційних мережах	45
1.3.1. Методи статистичного навчання	46
1.3.2. Методи апроксимації	51
1.3.3. Методи глибинного навчання	55
1.4. Постановка завдань досліджень	58
Висновки до 1 розділу	60
 РОЗДІЛ 2. МАТЕМАТИЧНА МОДЕЛЬ КЛАСТЕРИЗАЦІЇ ІНФОРМАЦІЙНИХ ПОВІДОМЛЕНЬ З ІНДИКАТОРАМИ АКТИВНОСТІ ІНФОРМАЦІЙНОГО КОНТЕНТУ ЗА ТОНАЛЬНІСТЮ І СФЕРАМИ ДІЯЛЬНОСТІ СОЦІУМУ	 62

2.1. Інфологічна модель джерел, факторів та показників активності контенту в глобальних інформаційних мережах	63
2.2.1. Математичний опис об'єктів моніторингу – інформаційних джерел	64
2.1.2. Індикатори активності інформаційного контенту	70
2.1.3. Формалізований опис інфологічної моделі джерел, факторів та показників активності контенту в глобальних інформаційних мережах	76
2.2. Математична моделі кластеризації інформаційних повідомлень	81
2.3. Верифікація та оцінювання ефективності застосування математичної моделі кластеризації інформаційних повідомлень	84
Висновки до 2 розділу	98

РОЗДІЛ 3. МАТЕМАТИЧНА МОДЕЛЬ БАГАТОФАКТОРНОГО ПРОГНОЗУВАННЯ АКТИВНОСТІ КОНТЕНТУ В ГЛОБАЛЬНИХ ІНФОРМАЦІЙНИХ МЕРЕЖАХ	100
3.1. Процеси багатофакторного прогнозування активного інформаційного контенту. Підготовчий етап	103
3.1.1. Первинний аналіз властивостей часових рядів та визначення їх статистичних характеристик	103
3.1.2. Очищення часового ряду від аномальних вимірів	107
3.1.3. Структурна оптимізація моделі даних	113
3.2. Статистичне навчання з нелінійними моделями	120
3.3. Обробки часових рядів з використанням оптимізаційних нейронних мереж	135
3.4. Інтегрована оцінка для багатофакторного прогнозування активності контенту в глобальних інформаційних мережах ...	151

3.5. Формалізований опис та індикатори ефективності математичної моделі багатфакторного прогнозування активності контенту в глобальних інформаційних мережах ...	165
Висновки до 3 розділу	174
РОЗДІЛ 4. ТЕХНОЛОГІЯ БАГАТОФАКТОРНОГО ПРОГНОЗУВАННЯ АКТИВНОСТІ КОНТЕНТУ В ГЛОБАЛЬНИХ ІНФОРМАЦІЙНИХ МЕРЕЖАХ	176
4.1. Технологія багатфакторного прогнозування активності контенту в глобальних інформаційних мережах	176
4.1.1. Апаратна реалізація кластеру обробки та прогнозування	181
4.2. Розробка макету програмно-апаратного комплексу прогнозування активності контенту в глобальних інформаційних мережах та оцінювання ефективності запропонованих рішень	185
Висновки до 4 розділом	201
ВИСНОВКИ ПО РОБОТІ	203
ЛІТЕРАТУРА	207
 Додаток А Акти впровадження	 218
Додаток Б Програмна компонента реалізації запропонованих рішень	223
Додаток В Список публікацій здобувача	253

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

АРМ – автоматизоване робоче місце;

БД – база даних;

ІВ – інформаційний вплив;

ІД – інформаційні джерела;

ІЗ – інформаційна загроза;

ОМ – об’єкти моніторингу;

ІП – інформаційне повідомлення;

МНК – метод найменших квадратів;

OLAP – online analytical processing.

NLP – Natural Language Processing.

LLM – large language model;

MAE – Mean Absolute Error;

RMSE – Root Mean Squared Error;

MSE – Mean Squared Error;

CNN – convolutional neural network;

RNN – recursive neural networks;

LSTM – Long Short-Term Memory;

LSM – Least Square Method;

ВСТУП

Актуальність теми.

Досвід подій сьогодення переконливо доводить домінуючу роль глобального інформаційного простору, породжуваного глобальними інформаційними мережами, в розповсюдженні будь-якої інформації та реалізації інформаційних впливів (ІВ). Популярність глобальних інформаційних мереж, їх відносно низька контрольованість та динамічність розвитку, різноманіття форм, методів та жанрів подання інформації, здатне породжувати множину інформаційних загроз (ІЗ) в багатьох соціальних сферах, що можуть призводити до зміни думок, настроїв та дій різних соціальних груп і соціуму в цілому не завжди в позитивному напрямку. Цьому також сприяє постійний розвиток інформаційних технологій.

Протидія ІЗ потребує організації ефективного моніторингу глобальних інформаційних мереж та прогнозування активності його інформаційного контенту.

Результати прогнозування активності контенту в глобальних інформаційних мережах є корисними для органів державного управління, місцевого самоврядування різної ланки, зацікавлених у дослідженні соціальних настроїв, протидії негативним ІВ будь-якої природи, для комерційних установ з метою визначені соціальних реакцій на дії менеджменту, наданих послуг, виготовленої та розповсюдження продукції будь-якої категорії тощо.

Своєчасна та адекватна реакція на виявлені ІЗ і недопущення їх переростання в небезпечні ситуації вимагає оперативних, достовірних та точних даних про: рівень їх безпеки та пріоритетність у нейтралізації; тенденцій та динаміці розвитку; критичних часових точок та інтервали зміни динаміки розвитку. Особливо важливо це за умов високої щільності потоку виникнення, динаміки розвитку та трансформації інформаційного контенту, що притаманно сучасному глобальному інформаційному простору [1-12].

Для забезпечення високих показників виявлення ІЗ в дисертаційній роботі запропоновано реалізувати багатофакторне прогнозування активності контенту в глобальних інформаційних мережах. Це забезпечує аналіз не лише наслідків ІВ, а і факторів, що призводять до появи ІЗ в різних соціальних сферах. Зазначене потребує та актуалізує завдання розробки математичних моделей багатофакторного прогнозування активності контенту в глобальних інформаційних мережах.

Аналіз існуючих підходів до кластеризації інформаційних повідомлень поданих природньою мовою дозволив виявити їх недоліки, що актуалізують завдання – розробки математичної моделі кластеризації інформаційних повідомлень [13-23].

Аналіз відомих підходів до обробки часових рядів та специфіка прогнозування активності інформаційного контенту дозволив сформулювати напрямки їх розвитку шляхом розробки математичної моделі багатофакторного прогнозування активності контенту з використанням та модифікацією методів статистичного та глибинного навчання [24-66].

У зв'язку із зазначеним, **актуальною є науково-прикладна задачі** з розробки і удосконалення математичних моделей багатофакторного аналізу та машинного навчання для обробки часових рядів.

Зв'язок роботи з науковими програмами, планами, темами.

Дисертаційні дослідження виконані відповідно до плану наукової роботи, затверджену на кафедрі обчислювальної техніки КПІ ім. Ігоря Сікорського та у відповідності до положень Закону України «Про внесення змін до деяких законів України щодо пріоритетних напрямів розвитку науки і техніки та інноваційної діяльності» – розвиток сучасних інформаційних, комунікаційних технологій, робототехніки, в межах науково-дослідної роботи «Нелінійні та багатокритеріальні математичні моделі для технологій Data Science та Embedded Systems», державний реєстраційний номер: 0124U003323.

Мета і завдання дослідження.

Мета дисертаційних досліджень полягає у підвищенні достовірності

(точності оцінювання та прогнозу) та повноти (множини врахованих факторів) визначення домінантного інформаційного контенту в глобальних інформаційних мережах.

Предмет досліджень: математичні моделі багатфакторного аналізу та машинного навчання для обробки часових рядів.

Об'єкт досліджень: процеси багатфакторного аналізу та машинного навчання з обробки часових рядів.

Методи дослідження: системний аналіз та синтез; методи статистичного та глибинного навчання; статистичних випробувань; математичного моделювання; багатфакторного аналізу та багатокритеріальної / оптимізації; кластеризації – для обробки і прогнозування часових рядів.

Для досягнення сформованої мети в дисертаційній роботі було вирішено комплекс взаємопов'язаних часткових завдань досліджень:

- аналіз існуючих підходів до прогнозування інформаційного контенту в глобальних інформаційних мережах;
- розробка математичної моделі кластеризації інформаційних повідомлень з індикаторами активності інформаційного контенту за тональністю і сферами діяльності соціуму та ієрархічною компресією вхідних Big Data масивів;
- розробка математичної моделі багатфакторного прогнозування активності контенту в глобальних інформаційних мережах з обробкою та аналізом часових рядів з використанням методів статистичного та глибинного машинного навчання;
- розробка технології багатфакторного прогнозування активності контенту в глобальних інформаційних мережах, яка базується на реалізації математичних моделей кластеризації інформаційних повідомлень та багатфакторного прогнозування активності контенту на спеціалізованому програмно-апаратному комплексі, побудованому

за принципами розподіленої комп'ютерної системи з відкритою архітектурою;

- оцінювання ефективності запропонованих рішень.

Наукова новизна отриманих результатів.

1. Отримала подальший розвиток математична модель кластеризації інформаційних повідомлень, яка базується на частотному аналізі їх тональності з використанням методологій обробки природньої мови (Natural Language Processing (NLP)) з підтримкою великих мовних, моделей; OLAP візуалізації результатів кластеризації та вирізняється встановленою системою індикаторів активності інформаційного контенту за сферами діяльності соціуму з ієрархічною компресією вхідних Big Data масивів. Це забезпечує підвищення повноти аналізу інформаційних повідомлень в глобальних інформаційних мережах шляхом врахування множини факторів за сферами діяльності соціуму.

2. Вперше запропоновано математичну модель багатофакторного прогнозування активності контенту в глобальних інформаційних мережах, яка базується на обробці та аналізі часових рядів з використанням методів статистичного та глибинного машинного навчання і відрізняється від відомих рішень:

виявленням і очищенням вхідних даних від аномальних вимірів з оперативним аналізом динамічних та статистичних властивостей вимірів в ковзному вікні;

визначенням параметрів нелінійних моделей з балансом диференціальних спектрів для методів статистичного навчання;

модифікацією штучних нейронних мереж для обробки часових рядів із впровадженням нелінійної схеми компромісів у вхідному прошарку та формування навчальної вибірки для екстраполяції за методами статистичного навчання;

визначенням інтегрованої оцінки активності інформаційного контенту за нелінійною схемою компромісів як результат багатфакторного аналізу вектору факторів за сферами діяльності соціуму;

пошуком прогнозованих біфуркаційних інтервалів як критичної активності інформаційного контенту;

Це забезпечує підвищення достовірності (ймовірності виявлення і точності прогнозу) домінантної активності контенту в глобальних інформаційних мережах.

3. Вперше розроблено технологію багатфакторного прогнозування активності контенту в глобальних інформаційних мережах, яка базується на реалізації запропонованих математичних моделей кластеризації інформаційних повідомлень та багатфакторного прогнозування активності контенту на спеціалізованому програмно-апаратному комплексі, побудованому за принципами розподіленої комп'ютерної системи з відкритою архітектурою та відрізняється впровадженням архітектурних рішень програмної компоненти з імплементацією розроблених математичних моделей кластеризації та багатфакторного прогнозування інформаційних повідомлень. Використання розробленої технології забезпечує підвищення достовірності та повноти визначення домінантного інформаційного контенту в глобальних інформаційних мережах.

Практичне значення отриманих результатів.

Отримані в роботі результати спрямовані на розвиток і удосконалення математичних моделей багатфакторного аналізу та машинного навчання для обробки часових рядів.

Використання запропонованих рішень забезпечує підвищенні достовірності (точності оцінювання та прогнозу до 20%) та повноти (множини врахованих факторів до 30%) визначення домінантного інформаційного контенту в глобальних інформаційних мережах. Результати оцінювання ефективності рішень отримані на модельних та на реальних даних.

Результати роботи доведені до демонстраційного прототипу програмної системи, яка розгорнута на макеті апаратної компоненти, що доводить дієвість прийнятих рішень та отриманих науково-прикладних результатів.

Запропоновані рішення є доцільними для оперативного визначення соціальних настроїв, протидії негативним ІВ будь-якої природи в різних прошарках соціуму, фокус-групах, цільових аудиторій, що реалізуються: комерційними структурами в сфері надання послуг, реалізації продажів, тощо; в органах державного управління різної ланки – місцевого самоврядування, міністерств та відомств; в сфері безпеки і оборони.

Практична цінність результатів дисертаційних досліджень підтверджена актами впровадження (див. додатки)

Особистий внесок здобувача. Дисертаційна робота розроблена автором особисто. В процесі досліджень здобувачем здійснено:

1) аналіз існуючих підходів виявлення конфлікту теорії та практики, формалізація задачі досліджень та розробку схеми наукового пошуку;

2) розроблено математичну модель кластеризації інформаційних повідомлень, яка базується на частотному аналізі їх тональності з використанням методологій обробки природньої мови (NLP) з підтримкою великих мовних, моделей;

3) запропоновано математичну модель багатofакторного прогнозування активності контенту в глобальних інформаційних мережах;

4) розроблено технологію багатofакторного прогнозування активності контенту в глобальних інформаційних мережах, яка базується на реалізації запропонованих математичних моделей кластеризації інформаційних повідомлень та багатofакторного прогнозування активності контенту на спеціалізованому програмно-апаратному комплексі;

5) проведені дослідження ефективності запропонованих рішень та доведена їх достовірність.

Апробація результатів дисертації. Основні результати роботи опубліковано у 7 тезах та доповідались і обговорено на міжнародних та

всеукраїнських наукових конференціях: Ninth International Congress on Information and Communication Technology, ICICT 2024; Міжнародна науково-практична конференція молодих вчених та студентів «Інженерія програмного забезпечення і передові інформаційні технології» (SoftTech-2023); The International Conference on Security, Fault Tolerance, Intelligence” (ICSFTI 2022); SAIC'2024 (IEEE International Conference on System Analysis & Intelligent Computing); Security, Fault Tolerance, Intelligence: proceedings of the VIII International Conference ICSFTI2025

Публікації. За результатами досліджень опубліковано 10 наукових статей з них: 4 статті у наукових фахових виданнях України; 6 статей у періодичних наукових фахових виданнях проіндексованих у базах Scopus та/або Web of Science Core Collection. Також опубліковано 1 розділ у співавторстві у колективній монографії.

Структура та обсяг роботи. Дисертаційна робота складається зі вступу, чотирьох розділів, загальних висновків, списку використаних джерел із 87 найменувань. Загальний обсяг дисертації становить 254 сторінок, з яких 161 сторінки основного тексту, 3 додатки на 39 сторінках, та містить 55 рисунків (21 сторінка повного тексту), 78 формул, 26 таблиць (17 сторінок повного тексту).

РОЗДІЛ 1

АНАЛІЗ ІСНУЮЧИХ ПІДХОДІВ ДО ПРОГНОЗУВАННЯ КОНТЕНТУ В ГЛОБАЛЬНИХ ІНФОРМАЦІЙНИХ МЕРЕЖАХ

Дисертаційні дослідження спрямовані на реалізацію процесів моніторингу інформаційного простору з метою прогнозування і виявлення критичної активності інформаційного контенту, якій обумовлює появу інформаційних загроз що є наслідками негативного інформаційного впливу.

Зазначене потребує комплексного та критичного аналізу існуючих методологічних і технологічних підходів до реалізації процесів:

розповсюдження контенту в глобальних інформаційних мережах та моніторингу його активності;

визначення і вимірювання індикаторів активності інформаційного контенту та загроз і створення їх причинно-наслідкової (інфологічної) моделі;

кластеризації інформаційних повідомлень за принципами обробки природньої мови;

прогнозування розвитку у часі індикаторів активності інформаційного контенту.

Результати аналізу передбачається формалізувати у наукове завдання досліджень та переліку часткових задач дисертаційних досліджень.

1.1. Аналіз особливостей розповсюдження контенту в глобальних інформаційних мережах та моніторингу його активності

Аналіз особливостей розповсюдження контенту в глобальних інформаційних мережах має на меті, насамперед, визначення і дослідження процесів, що підлягають моделюванню з метою ефективного прогнозування і виявлення критичної активності інформаційного контенту. Важливим також є аналіз існуючих підходів та специфіки задач моніторингу інформаційного

простору для прогнозування і виявлення критичної активності інформаційного контенту.

Питанням моніторингу і аналізу інформаційного простору для виявлення ІЗ присвячено доволі багато робіт, найвідомішими з яких є перелік [1-12]. Класичні методи моніторингу інформаційного простору і виявлення критичної активності в домінанті орієнтовані на відстеження розвитку тематики критичних повідомлень або домінантного контенту у часі за зміною частоти їх згадування. Це, по суті, однофакторний аналіз даних формату часових рядів (Time Series). Для цього використовують різні методи обробки часових рядів з переліку класів: статистичне навчання; методи апроксимації; глибинне навчання із штучними нейронними мережами. При цьому використовуються моделі монотонного тренду: поліноміальна з різним ступенем складності для відтворення сезонності; апроксимаційна крива Гауса; експоненціальна модель та інші. Такі підходи мають за недоліки однофакторне прогнозування наслідків, а не факторів, що до них призводить. Обмеженням відомих підходів також є модельне спрощення тренду досліджуваного процесу до монотонної функції, що не завжди відповідає дійсності. Зазначене призводить до зниження повноти (множини врахованих факторів) і достовірності визначення домінантного інформаційного контенту в глобальних інформаційних мережах.

В дисертаційній роботі категорії і поняття із сфери прогнозування активності контенту в глобальних інформаційних мережах вживаються у такому загальноприйнятому тлумачення [1-12].

Інформаційний вплив – організоване цілеспрямоване втручання у свідомість (підсвідомість) або фізичний стан цільової аудиторії та/або в процес функціонування технічних об'єктів інформаційної інфраструктури за допомогою використання інформаційних засобів та технологій.

Інформаційна загроза – це наміри, дії чи явища, які за допомогою інформаційного впливу на соціальні об'єкти, інформаційну інфраструктуру та

інформаційні ресурси можуть ускладнити (унеможливити) реалізацію національних інтересів держави (функцій її структурних органів).

Інформаційний простір – це інформаційне середовище, в якому відбуваються інформаційні процеси та інформаційні відносини щодо створення, збору, отримання, зберігання, використання, поширення та захисту інформації, інформаційних продуктів та інформаційних ресурсів.

Засобами інформаційного впливу є: засоби масової інформації, програмні, технічні та інші спеціальні засоби, які призначені для здійснення інформаційного впливу.

Будь-яке розповсюдження інформації в глобальному інформаційному просторі несе в собі позитивний чи негативний ефект – ІВ, залежно від контенту інформаційного повідомлення. У випадку негативного ІВ мова йде про інформаційні загрози (див. рис. 1.1).

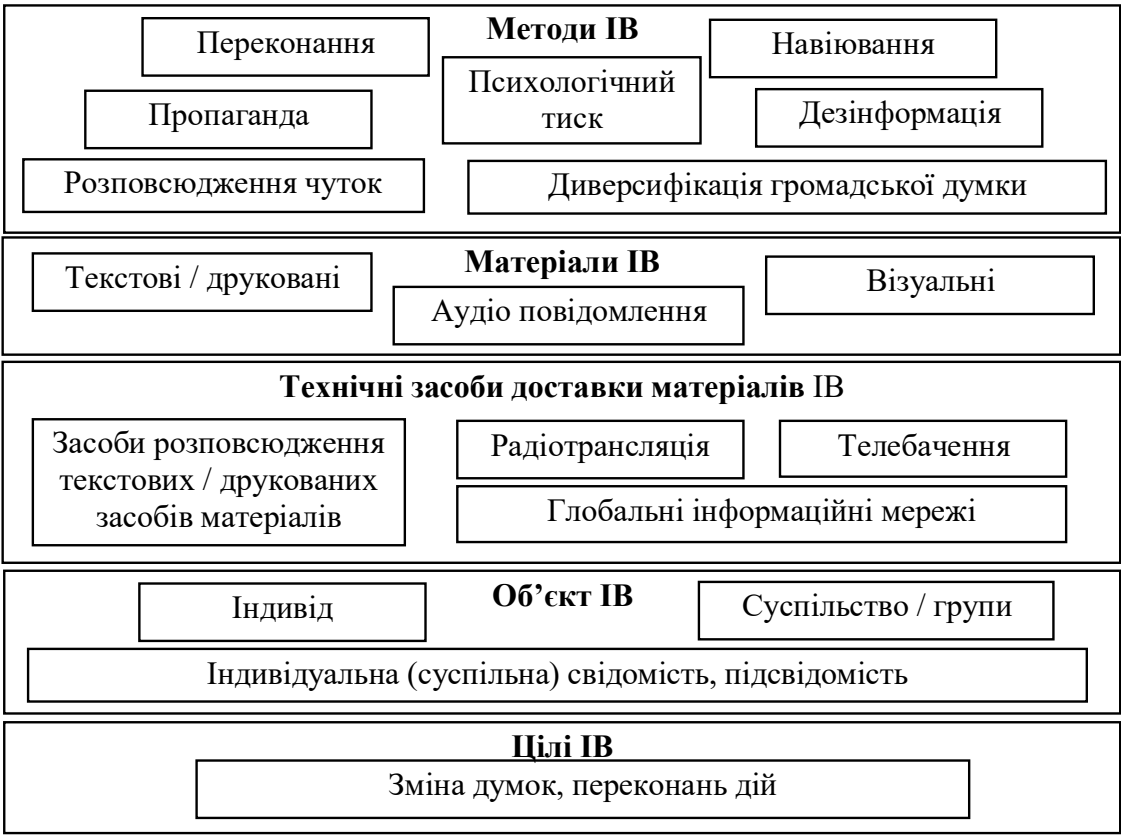


Рисунок 1.1. Методи і засоби досягнення цілей ІВ.

Досягнення цілей ІВ здійснюється за допомогою конкретних методів і засобів, що визначаються кінцевими цілями інформаційного впливу [1-12].

Моніторинг інформаційного простору з метою прогнозування і виявлення критичної активності інформаційного контенту, якій обумовлює появу ІЗ, здійснюється з реалізацією взаємопов'язаних етапів (див. рис.1.2): моніторинг інформаційного простору; зберігання результатів моніторингу; обробка результатів моніторингу.



Рисунок 1.2. Процеси моніторингу інформаційного простору.

Результатом моніторингу є індикатори активності інформаційного контенту та динаміка їх зміни.

Зазвичай ІВ реалізується в послідовності відомих етапів, що відображають динаміку зміни індикаторів активності інформаційного контенту як частоти появи повідомлень в інформаційному середовищі [1, 2]:

1. *Інформаційний етап* – створення актуальної або резонансної новини;
2. *Етап активації* – популяризація резонансної новини;
3. Досягнення цілей ІВ шляхом приведення до пікової активності резонансної новини – *цільовий етап*;
4. *Завершальний етап* – закріплення цілей ІВ.

Традиційне уявлення про динаміку зміни активності інформаційного контенту за етапами ІВ представлена графіком рис.1.3 (апроксимаційна крива Гауса). Тобто, маємо динаміку зростання до пікової активності з подальшим її спадом через втрату зацікавленості об’єкта ІВ новиною – втрата актуальності.

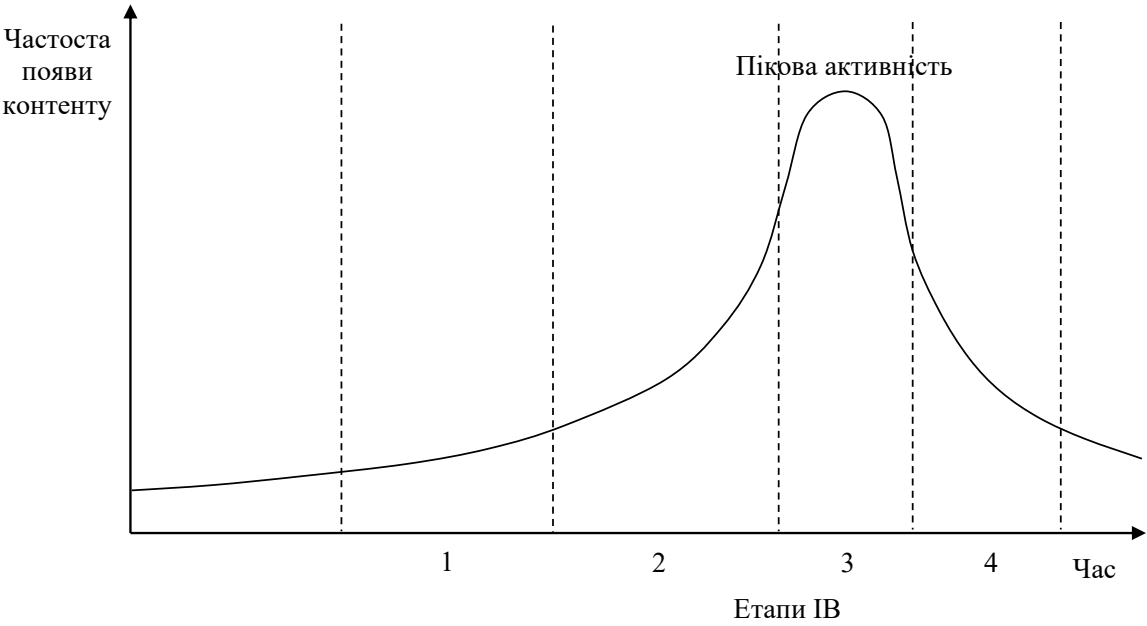


Рисунок 1.3. Динаміка зміни активності інформаційного контенту за етапами ІВ – традиційне уявлення

З точки зору масштабів впровадження та розвитку – ІВ слід розглядати на глобальному, першому, другому, третьому, середньому та локальному рівнях. Порівняти ці термінологічні одиниці з практичними термінами ІВ можна, керуючись даними табл.1.1 та графіком, рис.1.4. [2].

Часові масштаби ІВ

Таблиця 1.1

Форми інформаційної війни	Рівень цілей	Терміни реалізації	Об'єкти впливу	Завдання
1. Інформаційна кампанія	Стратегічні (глобальні)	Від одного до декількох місяців	Група та/або окремий об'єкт – населення держави, окремі соціальні групи, індивіди, які приймають державні рішення	Здійснюється в декількох напрямках ІВ в одній зі сфер життєдіяльності об'єкта впливу
2. Інформаційна операція	Оперативно-стратегічний (перший)	Від одного до декількох місяців	Група та/або окремий об'єкт – населення держави, окремі соціальні групи, індивіди, які приймають державні рішення	Здійснюється в декількох напрямках ІВ в одній зі сфер життєдіяльності об'єкта впливу
3. Інформаційний удар	Оперативний (другий)	Від одного до декількох місяців	Груповий та/або окремий об'єкт – для одного або декількох об'єктів державної системи управління	Здійснюється в декількох напрямках ІВ в одній зі сфер життєдіяльності об'єкта впливу
4. Інформаційна акція	Оперативно-тактичні (третій)	Від кількох тижнів до кількох місяців	Група та/або окремий об'єкт – конкретні соціальні групи, індивіди, які приймають державні рішення, громадські чи політичні діячі	Здійснюється в одному інформаційному напрямку
5. Інформаційна атака	Тактичний (середній)	Від кількох днів до тижня	Окремий об'єкт - конкретна людина, яка приймає рішення в конкретній сфері діяльності або група людей критичної структури	Здійснюється в одному інформаційному напрямку
6. Інформаційний захід	Тактичний (локальний)	Короткостроково, кілька днів	Окремий об'єкт – свідомість, підсвідомість індивіда	Здійснюється в одному інформаційному напрямку

Інформація табл.1.1 та рис.1.4 дає можливість оцінити часовий інтервал реалізацію процесів моніторингу інформаційного простору з метою прогнозування і виявлення критичної активності інформаційного контенту, якій обумовлює появу ІЗ, що є наслідками негативного ІВ. Ці дані можливо також використовуватися в якості вимог до прогностичних властивостей математичних моделей досліджуваного процесу.

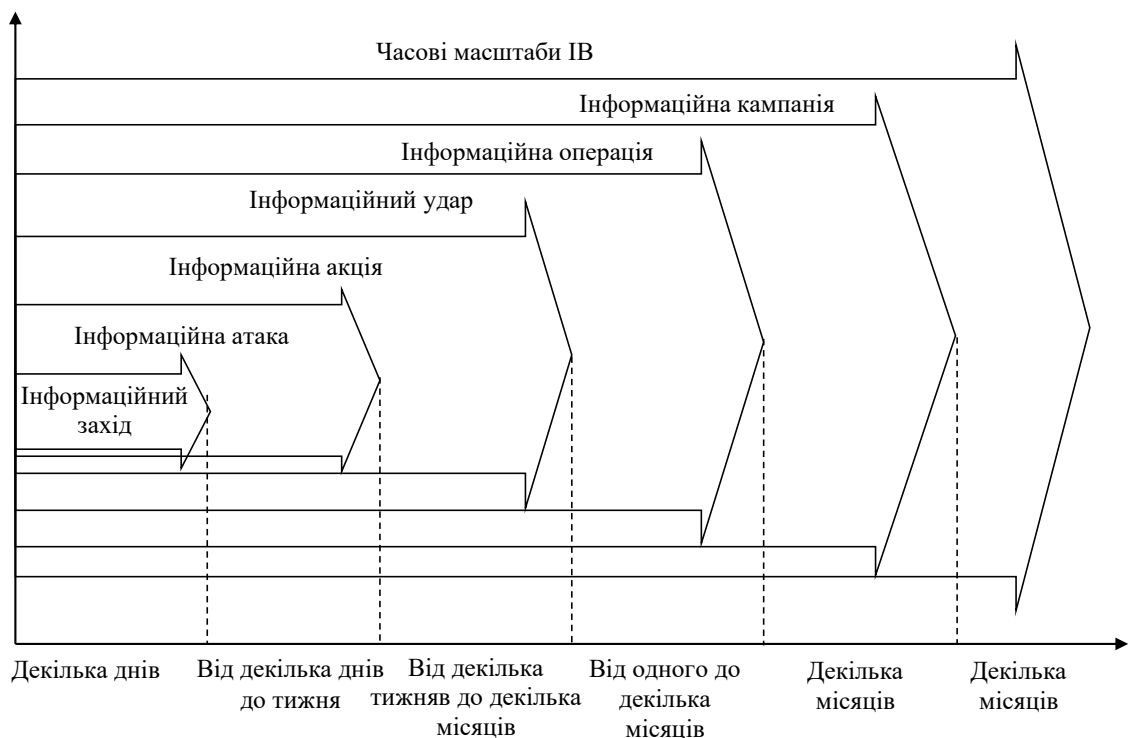


Рисунок 1.4. Часові масштаби ІВ.

На теперішній час домінантна кількість інформації розповсюджується глобальними інформаційними мережами. Цьому сприяють особливості, закономірності та динаміка розвитку глобального інформаційного простору як середовища розповсюдження інформаційного контенту, що полягає у наступному [1-12].

1. Розвиток глобальних інформаційно-комунікаційних технологій, поява і використання гіпермедійних методів в поєднанні з їх відносно низькою вартістю та доступністю для широкого кола користувачів;

2. Зниження рівня впливу, контролю або повна втрата монополії держав на інформацію, процеси її поширення, модифікації, формування змісту та емоційного забарвлення у т.ч. в питаннях інформаційного впливу, що, власне, і призводить до виникнення активності, збільшення динаміки зміни, кількості та класів ІЗ;

3. Високий потенціал і значна динаміка зміни форм, методів і способів соціальної комунікації, їх неповторюваність;

4. Нівелювання понять приватного і публічного, територіальних, часових і державних кордонів, моральних норм, мовних, соціальних, культурних і релігійних відмінностей між суб'єктами спілкування;

5. Можливості анонімної комунікації та висока захищеність джерел інформаційних повідомлень, що призводять до безкарності, безвідповідальності та апостеріорної (можливо, запізнитої) реакції на ІЗ;

6. Підміна реальності в свідомості учасників віртуального спілкування взаємним перенесенням і плутаниною моделей поведінки, цінностей, обмежень;

7. Виникнення психологічної та фізичної залежності індивіда від необхідності тривалої «присутності» в глобальному інформаційному просторі.

Практика доводить, що інформаційні впливи сьогодення є комплексними і охоплюють одночасно декілька сфер цінностей і діяльності соціуму. [1-12]. Наприклад: економічна; політична; соціальна; духовна; правова; безпекова; свобод і прав особистості тощо. *Реалізацією ІВ з комплексними ІЗ підтримується синергетичний ефекти у досягненні когнітивного дисонансу об'єкту впливу, що може призводити до досягнення цілей ІВ – зміна думок, настроїв та у діях.*

Тому, залежно від масштабу ІВ, способів його реалізації, інформаційного контенту з позицій сфер цінностей об'єкта впливу, суті інформаційного повідомлення (актуальної новини) та множини випадкових факторів з розповсюдження інформаційного контенту в глобальному інформаційному,

криві тренду рис. 1.3 можуть мати різні часові кореляції в локалізації піків активності інформаційного контенту, ступеня вираженості максимумів і їх співвідношень за абсолютною величиною.

Отже, розвиток активності інформаційного контенту *залежить від багатьох факторів*: характер розподілу по сферам цінностей, тональності, кількості; поточна соціальна ситуація; конкретика аудиторії з геопросторовою локацією; активність соціальних механізмів розповсюдження інформаційного контенту в соціумі, рівень комунікації та соціалізації; технічна доступність об'єктів впливу до технічних засобів розповсюдження інформації та інше. Тобто маємо багатофакторність та складні механізми розвитку процесу розповсюдження інформаційного контенту.

Тому дисертаційна робота спирається на альтернативний класичному уявленню (див. рис.1.3) погляд щодо динаміка зміни активності інформаційного контенту за етапами ІВ [1]. У цьому разі процеси ІВ розглядаються такими, що відбуваються на соціотехнічній системі. Соціотехнічну систему утворює соціум (індивід, соціальні групи, суспільство), що отримує інформацію від глобальних інформаційних мереж, утворених із використанням комп'ютерних систем та мереж різного класу, рівня та призначення. Процес розповсюдження інформаційного контенту підпорядкований модельним уявленням відповідно до базових положень синергетики та теорії хаосу [1, 24] – рис.1.5.

Тобто, *динаміка зміни активності контенту є фазовою траєкторією поведінки складної соціотехнічної системи з множинними аттракторами (підтримка / заперечення контенту) та точками біфуркації (зміна думок і настроїв)* – див. рис.1.5.

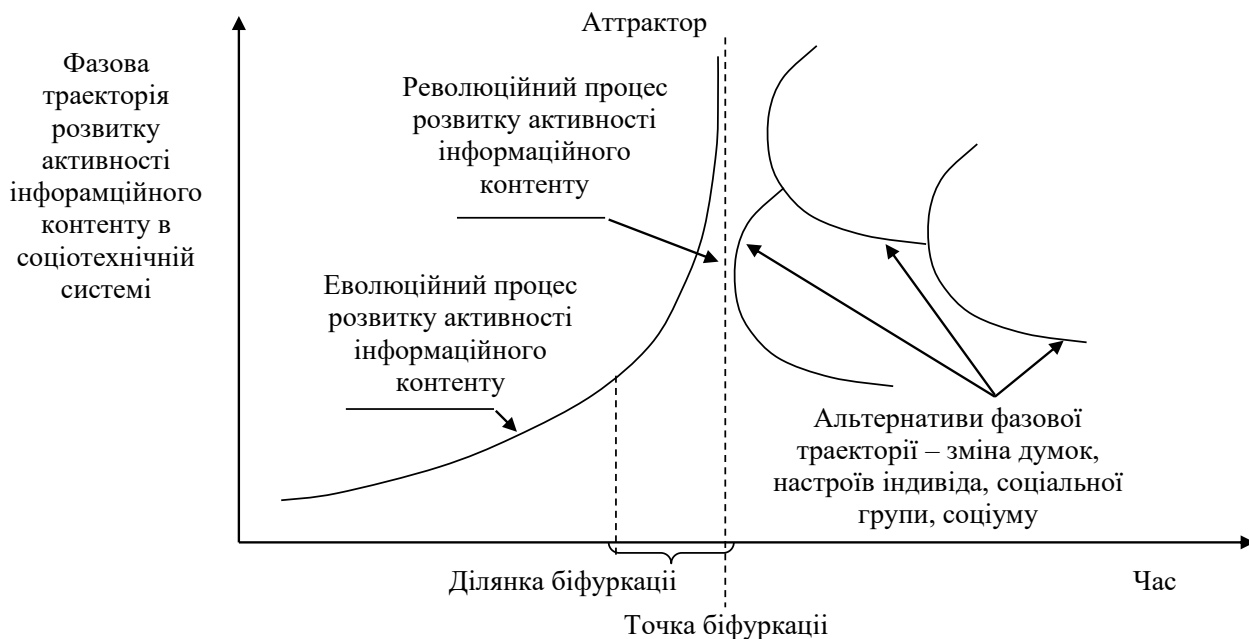


Рисунок 1.5. Фазова траєкторія зміни активності інформаційного контенту на соціотехнічній системі

Фазова траєкторія буде розглядатись як суперпозиція динаміки зміни інформаційного контенту в різних сферах, що визначатиметься як наслідки впливу багатьох факторів через статистику появи інформаційних повідомлень в мережі. *Прогнозування активності має на меті контроль зміни по сферах (економічна; політична; соціальна; духовна; правова; безпекова; свобод і прав особистості тощо), виявлення аттракторів та небезпечних ділянок фазової траєкторії із наближенням до потенційних точок біфуркації.*

Таким чином для прогнозування розвитку активності інформаційного контенту пропонується реалізувати комплекс взаємопов'язаних етапів:

1. Моніторинг глобального інформаційного простору на рівні новин та збереження результатів моніторингу в форматі баз даних. Моніторинг матиме встановлену періодичність у часі із зверненням до встановленого переліку джерел;

2. Встановлення належності новин до сфер цінностей переліку (економічна; політична; соціальна; духовна; правова; безпекова; свобод і прав особистості тощо);
3. Встановлення тональності інформаційних повідомлень за визначеними сферами цінностей соціуму;
4. Підрахунок кількості повідомлень за тональністю у кожній із сфер цінностей та каталогізацію отриманих вимірів в форматі часового ряду;
5. Багатофакторне оцінювання та прогнозування отриманих часових рядів з метою виявлення критичної активності інформаційного контенту, якій обумовлює появу інформаційних загроз що є наслідками негативного інформаційного впливу.

Зазначену послідовність етапів пропонується реалізувати на програмно-апаратному комплексі, побудованому за принципами ситуаційного центру, що забезпечує високі показники достовірності (ймовірності виявлення і точності прогнозу) та повноти (множини врахованих факторів) визначення домінантного інформаційного контенту в глобальних інформаційних мережах. Функціонування комплексу має враховувати інтенсивність, значну динаміку зміни і щільність потоку інформаційних повідомлень та мати здатність до нарощування функціональності і масштабованості.

1.2. Методи кластеризації інформаційних повідомлень за принципами обробки природньої мови (NLP-процеси)

Накопичений за результатами моніторингу масив інформаційних повідомлень підлягає кластеризації на сфери цінностей (економічна; політична; соціальна; духовна; правова; безпекова; свобод і прав особистості тощо) та по тональності (негатив / позитив / нейтральні). Інформаційні повідомлення за суттю подані природньою мовою. Передбачається моніторинг україномовних та англомовних джерел із можливостями

застосування запропонованих рішень на інші мови. Потребує аналізу наявні методологічні підходи і технології обробки природньої мови (Natural Language Processing – NLP) для розв’язку сформованої у такий спосіб задачі прикладної задачі.

Кластеризація інформаційних повідомлень за суттю відноситься до задач розпізнавання образів із специфікою обробки природньої мови. В узагальненому та спрощеному розумінні це передбачає створення образу (тематичної фрази, словнику-каркасу) та пошук міри його близькості з об’єктом кластеризації – поточним текстом. Зазначена схема характерна для задач кластеризації, класифікації та ідентифікації. Для обробки інформаційних повідомлень природньої мови різниця між їх переліком умовна і відповідає рівню докладності результату. Оскільки для задач дисертаційних досліджень передбачено поділ інформаційних повідомлень за тематикою і тональністю контенту, без докладної класифікації та ідентифікації логічного сенсу повідомлення – варто називати цей процес як поділ інформаційних повідомлень на інформаційні кластери, тобто кластеризацією.

Відомі підходи до кластеризації можна розділити на три групи (рис.1.6) [13-23]:

1. Методи машинного навчання;
2. Техніки Word2vec, GloVe, RNNLM та класичні нейронні мережі;
3. Методи глибинного навчання з використанням великих мовних моделей.



Рисунок 1.6. Відомі методи кластеризації

Методи машинного навчання представлені відомим переліком: метод k-Means / Hard-c-means / k-середніх; метод «найближчого сусіда» / k-nearest neighbors algorithm / k-NN; метод опорних векторів (support vector machine - SVM); метод Роше [13-23].

Методи k-середніх та найближчого сусіда є класичними і полягають у пошуку мінімальної евклідової відстані між ключовою фразою та текстом інформаційного повідомлення за виразом:

$$l = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + \dots}, \quad (1.1)$$

Де l – величина евклідової відстані між ознаками образу $x_1, y_1, \dots$, та об'єкту кластеризації $x_2, y_2, \dots$,

Для k-середніх властива апriorне формування кількості кластерів k , відносно яких і утворюються кластери. Метод k-NN не потребує апriorного значення k , натомість передбачає повний перебір об'єктів кластеризації з образом для утворення кластерів.

Метод опорних векторів орієнтовано на пошуку мінімуму кута між векторами ознак / образу та об'єкту ідентифікації з розрахунком скалярного добутку векторів

$$\cos \alpha = \frac{(x_1, y_1) + (x_2, y_2)}{\sqrt{(x_1^2 + y_1^2)} \sqrt{(x_2^2 + y_2^2)}}, \quad (1.2)$$

Метод Роше (класифікатор Роше) забезпечує рубрикацію документа (класифікацію), визначаючи його приналежність до певної рубрики (категорії), базуючись на його семантичній близькості до еталонного вектора цієї рубрики за векторним аналізом відстаней в просторі ознак [13-23].

Метод базується на етапах:

векторизація: документ перетворюється на вектор ознак (наприклад, частота ключових слів);

еталони рубрик: кожна рубрика утворює каркас / словник ознак;

обчислення відстані: класифікатор визначає відстань між еталоном та текстом, наприклад за (1.1), або (1.2);

кластеризація: документу присвоюється рубрика найближчого еталону.

Аналіз показав, що усі класичні методи кластеризації працюють на так званому статичному каркасі. Каркасом для природньої мови є словник з переліку слів / словосполучень. Каркас виступає в якості образу, тож для задачі досліджень поділ за тональністю і соціальними сферами потребуватиме свого словника мовою Інформаційних повідомлень.

Недоліками таких підходів є статика, тобто ефективність кластеризації залежить від обсягу словника. Статичність словника не дозволяє мати адаптацію до різнобарвності природньої мови у формування висловів. Тож використання класичних методів кластеризації для завдань дисертаційних досліджень є не продуктивним.

Техніки Word2vec, GloVe, RNNLM та класичні нейронні мережі передбачає: векторного уявлення слів; підходи первинного моделювання

штучних нейронних мереж; підходи з конструювання; підходи, засновані на існуючих та / або навчених нейронних мереж.

Перші три підходи вимагають значних витрат на визначення вдалої архітектури нейронної мережі та її навчання. Тому фокус досліджень спрямовано на існуючих навчених штучних нейронних мереж. В цьому напрямку варто виділити техніки та нейронні мережи класу Word2vec, GloVe RNNLM. Їх використання планується для встановлення тональності інформаційних повідомлень.

Методи глибинного навчання, засновані на великих мовних моделях (LLM – large language model) каркасу – із підтримкою генеративних штучних нейронних мереж. Цей напрям наразі є найбільш ефективним і передбачається для застосування з кластеризацію інформаційних повідомлень за соціальними сферами. Типовими архітектурними рішеннями LLM є генеративні трансформери класу: GPT, BERT, RAG, Falcon, OPT.

Виходячи із зазначеного можна виділити недоліки відомих підходів до кластеризації інформаційних повідомлень за принципами обробки природньої мови:

1. Препроцесінг – складність роботи в національному інформаційному просторі;
2. Суб'єктивізм в кластеризація – необхідність врахування суб'єктивної спрямованості в кластеризації за тональністю;
3. Конкретика цільового спрямування – потребує врахування сфер та об'єктів ІВ;
4. Великий обсяг та значна динаміка зміни інформаційного контенту – потребує оперативного аналізу, зберігання та компресії вхідних даних.

Для кластеризації інформаційних повідомлень за тональністю та соціальними сферами будуть використані методи глибинного навчання з використанням великих мовних моделей.

1.3. Аналіз існуючих методологічних підходів до прогнозування активності контенту в глобальних інформаційних мережах

Реалізація процесів моніторингу інформаційного простору базуються на даних формату часового ряду - time series. Їх обробка доволі показово представлена (див. рис.1.7): методами апроксимації (алгоритми ковзного середнього – Moving Average (MA) експоненційної апроксимації – Exponential Moving Average (EMA), автокореляційні алгоритми типу ARIMA); методами статистичного навчання (згладжування) – Statistical Learning (метод найменших квадратів (МНК) – Least Square Method (LSM), фільтрація Калмана тощо); методами глибинного навчання з використанням штучних нейронних мереж [24-62].



Рисунок 1.7. Методи обробки часових рядів

Надалі слід провести докладний аналіз існуючих підходів з обробки часових рядів відповідно до класифікаційним ознакам, наведеним на рис.1.7.

1.3.1. Методи статистичного навчання

Традиційно методи Statistical Learning використовують лінійні за параметрами моделі у формі степеневих поліномів. Але реальність полягає в тому, що більшість процесів, які підлягають аналізу мають нелінійну природу, тим самим апріорно обмежуючи можливість використання цього підходу. Саме тому іноді надають перевагу методам апроксимації та глибинного навчання, які, по-факту, відтворюють стохастичний процес. Відомо доволі багато підходів, які дозволяють частково вирішити проблему нелінійного згладжування (лінеаризація, чисельні методи) [41, 54]. Більшість з них полягає у спрощенні нелінійної моделі до набору лінійних компонент, або потребують початкових умов запуску ітераційних процесів «підгонки» параметрів нелінійних моделей до даних. Однак в практиці застосування такі моделі втрачають як більшу частину своєї нелінійної інформації, так і власну прогностичну цінність.

Статистичне навчання базується на обробці вимірів часового ряду (Time Series) – дискретної множини вимірів у досліджуваного процесу в часі x_i :

$$y = \{y(x_1), y(x_2), \dots, y(x_i)\}. \quad (1.3)$$

Виміри (1.3) містять тренд, випадкові домішки (залишки) та сезонну закономірність.

Статистичне навчання спрямовано на обробку вимірів (1.3) для виділення / навчання математичної моделі тренду із зменшенням випадкової похибки даних з урахуванням сезонних повторень / коливань.

Для застосування методів статистичного навчання висувається гіпотеза про загальний вигляд математичної моделі тренду, що за обраним критерієм оптимальності проходить повз дискретну множину вимірів (1.3).

В загальному випадку, математична модель формується як функціональна залежність:

$$y(x) = f(c), \quad c = \{c_1, c_2, \dots, c_i\}. \quad (1.4)$$

Параметри моделі (1.4) – конкретизують тренд та визначаються / навчаються за вибіркою вимірів (1.3) виходячи із критерію оптимальності / наближення тренду до оригіналу вимірів (1.3):

$$\delta(c) = \Phi(y(x) - y) \rightarrow \min. \quad (1.5)$$

Конкретика критерію Φ – визначає клас і тип відомих метод статистичного навчання.

До типу моделі (1.4) за класичними підходами висуваються вимоги: адекватної інтерпретації результатів спостереження (1.3); лінійності за параметрами.

Важливим аспектом вимоги лінійності за параметрами моделі (1.4) є отримання аналітичного розв'язку задачі статистичного навчання – алгоритму статистичного навчання. Це забезпечує формування лінійної системи рівнянь для пошуку параметрів моделі (1.4) з функціоналу (1.5):

$$\frac{\partial \delta(c_i)}{\partial c_i} = 0. \quad (1.5)$$

Адекватність інтерпретації моделлю $y(x)$ даних (1.3) вбачається в мінімізації випадкової похибки оцінок вхідних даних та мінімізації абсолютного відхилення моделі – динамічної похибки.

Найчастіше критерій Φ спрямовано на мінімізацію випадкової складової вхідних даних. Мінімізація динамічної похибки забезпечується формуванням адекватної структури моделі $y(x)$ – та її тюнінгом.

Залежно від сценарію отримання вибірки вимірів (1.3) розрізняють методи статистичного навчання для накопиченої вибірки – big data масиву та методи статистичного навчання за вибіркою об'єм якої зростає.

До алгоритмів на – big data масивах відносять метод найменших квадратів (МНК) – Least Square Method (LSM) [41, 54].

До алгоритмів, що працюють в динамічних умовах зростання кількості вимірів відносять рекурентні процедури, наприклад Калманівської фільтрації – рекурентні фільтри Калмана [41, 54].

Алгоритми МНК та фільтрації Калмана базуються на жорсткій вимозі лінійності за параметрами моделі (1.4). Нелінійність входних даних (1.3) в кластичних алгоритмах статистичного навчання відтворюється в класі поліноміальних моделей. При цьому спостерігається збільшення обсягів поліноміальних складових, що далеко не завжди дозволяє з високими ступеням точності відтворити тренди із складними сезонними залежностями. Більш перспективними в цьому сенсі є нелінійні за параметрами математичні моделі.

Після формування первинної гіпотези про вигляд моделі (1.4) відбувається її структурна оптимізація – тюнінг. Оптимальною вважається модель, що дає мінімум випадкової та динамічної похибки моделі. Однак ці вимоги є суперечними і потребують компромісних рішень.

Індикаторами (показниками ефективності) моделей даних вигляду (1.4) зазвичай є наступні [41, 54].

1. Глобальне лінійне відхилення оцінки – (mean absolute error (MAE)) – є відносною, оцінюється методом порівняння з вимогою мінімізацією:

$$\Delta = \frac{1}{n-1} \sum_{i=1}^n |y_i - \hat{y}_i|. \quad (1.6)$$

2. Достовірності апроксимації R^2 (coefficient of deternimation) – змінюється в межах $0 \dots 1$ та має бути мінімальним:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (1.7)$$

де n – об'єм вибірки $\bar{y}_l = \frac{1}{n} \sum_{i=1}^n y_i$;

y_i – виміряне значення;

$\hat{y}_i$ – оцінка виміряного значення.

3. Похідні вищих порядків – контроль нульових значення похідних:

$$y_j^{(p)} = \frac{y_{j+1}^{(p-1)} - y_j^{(p-1)}}{\Delta t}, \quad j = \overline{1 \dots m}, \quad p = \overline{1 \dots n}. \quad (1.8)$$

Для визначення оптимальної структури моделі даних реалізується тестування різних структурних рішень шляхом обробки вхідних даних з контролем індикаторів (1.6)-(1.8). Для використання обирається найкраща за контрольованими параметрами модель.

Зрозуміло, що такий процес є обтяжливий для великих даних та потребує удосконалення в напрямку оперативності.

Ефективне застосування класичних методів статистичного навчання можливо за умов виконання вимог, що висуваються до якості первинних даних – вимірів (1.3). Зокрема, вони мають бути рівноточні, рівнодискретні та однорідні. Реальна практика отримання та обробки часових рядів свідчить про порушення цих умов, зокрема за наявності аномальних вимірів типу пропуски та грубі значення. Це значно спотворює результати обробки [24-37].

Пропуски даних доволі ефективно відновлюються відомими методами екстраполяції.

Стосовно аномалій типу грубі виміри також застосовують підходи відновлення, але до того їх слід виявити.

У задачах очищення вибірки від аномалій існує чимало різновидів методів і алгоритмів, заснованих на різних підходах і принципах [24-37]. Усі відомі підходи базуються на унітарному та/або комбінаторному аналізі ознак аномалій. Ознаками аномалій типу грубі значення є:

- зміна динаміки часового ряду (динамічні властивості);

- різниця значення окремого виміру порівняно з іншими вимірами (властивості абсолютних значень вимірювань);
- зміни статистичних властивостей вибірки за наявності аномалій (статистичних властивостей).

Залежно від ознак, що використовуються для виявлення АМ, та принципів їх виявлення, слід розрізняти такі класи:

- методи кластеризації відповідно до принципів машинного навчання [30];
- методи аналізу динамічних властивостей часових рядів [32];
- методи аналізу статистичних властивостей часових рядів [33].

Незважаючи на універсальність і широке представлення цих підходів, їхній головний недолік – емпіричне (дослідницьке) коригування параметрів залежно від характеру властивостей часового ряду. Це може бути неприйнятно, оскільки пошук найкращих рішень може зайняти значний час на етапі розробки. Це також ускладнює їхню практичну реалізацію та масштабованість для часових рядів із широким спектром властивостей, які іноді змінюються під час роботи алгоритму детекції аномалій. Недоліки відомих підходів до навчання за параметрами алгоритмів для очищення вибірки з аномаліями також полягають у складності їх реалізації на масивах Big Data з значними нелінійностями та сезонністю.

Тому виявленню та компенсації впливу аномалій буде приділена увага в дослідженні і розробці ефективних, практично спрямованих підходів.

Таким чином, класичні підходи до статистичного навчання потребують модифікації в напрямку застосування нелінійних за параметрами моделей, підвищення ефективності формування структури моделі даних та очищення вхідних вимірів від аномалій.

1.3.2. Методи апроксимації

Методи апроксимації мають значне практичне розповсюдження і, відповідно широкую технологічну підтримку багатьма спеціалізованими бібліотеками. Вони доволі ґрунтовно опрацьовані теоретично та мають значний досвід практичного застосування. Саме це пояснює їх популярність.

Головні парадигма методів апроксимації полягає у визначенні параметрів певної моделі даних частіше з вимогою проведення лінії тренду через усі точки часового ряду в межах ковзного вікна – Moving Average (MA). Тобто на обмеженому інтервалі розгляду часового ряду. Це, зокрема забезпечує послаблення вимог до адекватності і складності моделі даних.

Прикладами алгоритмів апроксимації є [55-58]:

- алгоритм статистичного середнього;
- алгоритм середньозваженого;
- алгоритми MA із методами статистичного навчання.

Алгоритм статистичного середнього базується на обчисленні математичного очікування в ковзному вікні із заміною елемента часового ряду на математичне очікування:

$$m_y = \frac{1}{n} \sum_{i=1}^n y_i, \quad (1.9)$$

y_i – реалізація / вимір з часового ряду у вікні;

n – розмір ковзного вікні.

Такий підхід передбачає лінійність моделі даних у вікні, а прогнозування реалізується за умовною моделлю на відстань ковзного вікна.

Алгоритм середньозваженого передбачає врахування точності вимірів відповідно до виразу

$$m_y = \frac{1}{n} \sum_{i=1}^n \frac{1}{\sigma_y} y_i, \quad (1.10)$$

$\sigma_y = \sqrt{D_y}$ - середньоквадратичне відхилення похибки виміру у вікні;

$D_y = \frac{1}{n-1} \sum_{i=1}^n (y_i - m_y)^2$ – дисперсія похибки виміру у вікні.

Алгоритми МА із методами статистичного навчання передбачають можливість застосування моделей даних високого порядку нелінійності, але у ковзному вікні.

Переваги алгоритмів Moving Average включають наступне [58].

- Зменшення шуму: Моделі ковзних середніх ефективно згладжують шум, присутній у даних часових рядів, що знижує короткострокові коливання в даних часових рядів.
- Стаціонарність: Моделі ковзних середніх самі по собі припускають, що дані часових рядів є стаціонарними, що є основною вимогою в аналізі часових рядів.
- Комбінація моделей: Моделі ковзних середніх можна легко комбінувати з авторегресивними та інтегрованими моделями, і все моделювання ARIMA можна застосовувати до даних часових рядів.

Недоліки алгоритмів Moving Average.

- Припущення стаціонарності: У моделях ковзних середніх вже передбачається, що дані часових рядів є стаціонарними, але насправді більшість даних реального світу не є стаціонарними.
- Обмежений діапазон прогнозування: Моделі ковзних середніх підходять для короткострокового прогнозування, оскільки зі збільшенням часу накопичується помилка спрощеної моделі даних.
- Чутливість до викидів.
- Визначення порядку моделі: Якщо порядок моделі вибрано неправильно, модель може забезпечити адекватне оцінювання даних.

Наступним кроком в розвитку принципів ковзного вікна є алгоритми, побудовані на оцінюванні авторегресії та їх модифікації.

ARIMA (Auto Regressive Integrated Moving Average) авторегресивна інтегрована модель ковзних середніх.

Компоненти ARIMA включають наступне.

Авторегресія (AR): Використовує минулі значення ряду як предиктори, зв'язуючи поточне значення з минулими.

Інтегрована (I): Проводиться диференціювання часового ряду для перетворення його в стаціонарний (тобто, стабільного з часом), що є необхідним для коректної роботи моделі.

Ковзна середня (MA): Використовує минулі помилки прогнозу як предиктори, допомагаючи зрозуміти циклічні коливання.

Коефіцієнт автокореляції визначається за виразом:

$$r(t) = \frac{\frac{1}{n-t} \sum_{i=1}^{n-t} (y_i - m_y)(y_{i+t} - m_y)}{\frac{1}{n} \sum_{i=1}^n (y_i - m_y)^2}, \quad (1.11)$$

n – довжина часового ряду;

t – часове зрушення;

m_y – математичне очікування y_i

Типи моделей ARIMA включають [58]:

- ARIMA: Несезонні авторегресивні інтегровані ковзні середні
- SARIMA: Сезонна ARIMA
- SARIMAX: Сезонний ARIMA з екзогенними змінними

До алгоритмів апроксимації також слід відносити клас алгоритмів експоненціального згладжування - Exponential Moving Average (EMA). Такі алгоритми, по суті, є конверсією і модифікацією алгоритмів Калманівської фільтрації для ковзного вікна. Їх переваги та недоліки повністю вміщують аналіз алгоритмів Moving Average.

Отже, алгоритми апроксимації для задач прогнозування активності інформаційного контенту мають низку суттєвих недоліків. Зокрема до них слід віднести наступне:

проблематика оптимізації трендової кривої з вимогою мінімізації випадкових складових забезпечується лише опосередковано;

робота із стаціонарними проміжками часового ряду – є частковим випадком часового ряду;

значна залежність якості роботи алгоритмів від первинних параметрів їх налаштувань;

вимога апріорної визначеності характеристик вхідного часового ряду.

Останнє є домінантною у задачах прогнозування активності інформаційного контенту. Оскільки стаціонарність реальних даних – не гарантовано, а властивості результатів моніторингу за часовим рядом є апріорно невідомими. Тобто якщо алгоритми апроксимації ефективно налаштувати на ретроспективні дані, то їх використання для прогнозування хоча б на пів інтервалу спостереження є проблемним, а якість результату – не гарантовано. Тож використання алгоритмів апроксимації у даній роботі – не передбачається за можливе.

1.3.3. Методи глибинного навчання

Наразі стрімкого розвитку отримали методи глибинного навчання з використанням штучних нейронних мереж. Їх реалізація присутня в типах: CNN, RNN, LSTM, LLM.

Головна парадигма застосування штучних нейронних мереж для обробки часових рядів полягає у накопиченні значної кількості реалізацій випадкових процесів формату часового ряду. Такі технології відображають потужність методів статистичних випробувань. Результат обробки часового ряду з використанням штучних нейронних мереж – це власне повна тотожність реальній реалізації.

Наразі найбільшої популярності здобули згорткові та рекурентні штучні нейронні мережі.

Згорткові нейронні мережі (convolutional neural network – CNN). Традиційно розроблялись для аналізу зображень, але з успіхом застосовуються для часових рядів.

Головна ідея згорткової нейронної мережі показано на рис.1.8.

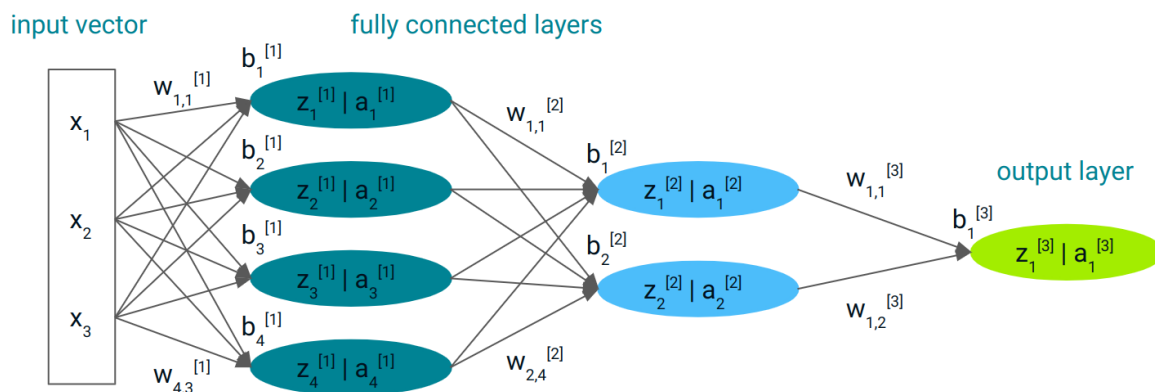


Рисунок 1.8. Згорткова нейронна мережа

<https://codilime.com/blog/convolutional-networks-for-time-series-classification/>

Згортковий шар служить основним будівельним блоком CNN. Цей процес реалізується як ковзне вікно по сітці даних, де фільтр переміщується горизонтально-вертикальним чином із множенням параметрів фільтра на відповідні частини даних. Це забезпечує видобуток корисної інформації з сітки даних.

Ключові переваги CNN для аналізу та обробки часових рядів:

1. Локальна зв'язність: CNN використовують згорткові шари, які зосереджені на локальних областях вхідних даних. Ця характеристика дозволяє їм ефективно фіксувати короткострокові закономірності, що є вирішальним для прогнозування часових рядів.
2. Ієрархічне представлення: CNN складаються з кількох згорткових шарів, що дозволяє їм вивчати ієрархічні представлення даних. Це робить їх здатними фіксувати як короткострокові, так і довгострокові залежності.

3. Паралельна обробка: CNN можуть обробляти кілька послідовностей одночасно, що робить їх ефективними для обробки великих наборів даних та багатовимірних часових рядів.

Рекурентні штучні нейронні мережі (recursive neural networks – RNN).

Ідея рекурентної нейронної мережі / RNN полягає в реалізації прихованого прошарку за рекурентною процедурою оцінювання наступних значень за попередніми значеннями часового ряду.

Алгоритм RNN (рис.1.9) [55, 61].

RNN приймає вхідний вектор x і генерує вихідний вектор y , скануючи дані послідовно зліва направо, з кожним кроком у часі оновлюючи прихований стан і створюючи вихід.

В якості вагових коефіцієнтів RNN використовують один і той же набір параметрів U, V, W .

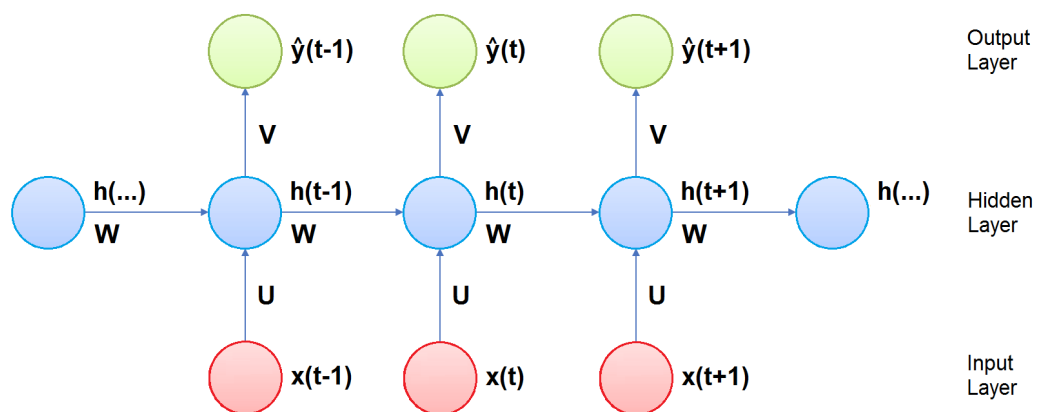


Рисунок 1.9. Рекурентна нейронна мережа

<https://towardsdatascience.com/time-series-forecasting-with-deep-learning-and-attention-mechanism-2d001fc871fc/>

U – ваговий параметр, що регулює з’єднання від вхідного шару x до прихованого шару h ; W – вага, пов’язана зі з’єднанням між прихованими шарами; V – вага для з’єднання від прихованого шару h до вихідного шару y .

Прихований стан – вектор $h(t)$ є пам'яттю мережі та обчислюється на основі поточних вхідних даних та прихованого стану попереднього кроку часу за рекурентним виразом:

$$h(t) = \sigma(Wh(t-1) + Ux(t)). \quad (1.12)$$

Вихідне значення обчислюється за рівнянням:

$$\hat{y}(t) = \max(Vs(t)), \quad (1.13)$$

$s(t)$ – помилка поширення.

Таким чином, наведений аналіз дозволяє сформулювати наступні недоліки існуючих підходів до обробки та прогнозуванню часових рядів:

1. Препроцесінг – (робота з АВ) є «повільним» та не враховує динаміки часових рядів;
2. Оптимізація структури / тюнінг – моделі орієнтовано на перебір з повторенням;
3. Статистичне навчання – орієнтовано на застосування лінійних за параметрами моделей;
4. Глибинне навчання – ускладнене випадковими похибками даних та має відносно невеликий інтервал прогнозу.

На їх подолання спрямовано дисертаційні дослідження.

1.4. Постановка завдань досліджень

Дисертаційна робота присвячена розв'язку актуальної є науково-прикладної задачі розробки і удосконалення математичних моделей багатofакторного аналізу та машинного навчання для обробки часових рядів.

Метою дисертаційних досліджень є підвищення достовірності (ймовірності виявлення і точності прогнозу) та повноти (множини врахованих факторів) визначення домінантного інформаційного контенту в глобальних інформаційних мережах.

Для досягнення сформованої мети в дисертаційній роботі було вирішено комплекс взаємопов'язаних часткових завдань досліджень:

- аналіз існуючих підходів до прогнозування інформаційного контенту в глобальних інформаційних мережах;
- розробка математичної моделі кластеризації інформаційних повідомлень з індикаторами активності інформаційного контенту за тональністю і сферами діяльності соціуму та ієрархічною компресією вхідних Big Data масивів;
- розробка математичної моделі багатфакторного прогнозування активності контенту в глобальних інформаційних мережах з обробкою та аналізом часових рядів з використанням методів статистичного та глибинного машинного навчання;
- розробка технології багатфакторного прогнозування активності контенту в глобальних інформаційних мережах, яка базується на реалізації математичних моделей кластеризації інформаційних повідомлень та багатфакторного прогнозування активності контенту на спеціалізованому програмно-апаратному комплексі, побудованому за принципами розподіленої комп'ютерної системи з відкритою архітектурою;
- розробка макету програмно-апаратного комплексу прогнозування активності контенту в глобальних інформаційних мережах та оцінювання ефективності запропонованих рішень.

Висновки до 1 розділу

1. Аналіз показав, що протидія інформаційним загрозам потребує організації ефективного моніторингу глобальних інформаційних мереж та прогнозування активності його інформаційного контенту.

Результати прогнозування активності контенту в глобальних інформаційних мережах є корисними для органів державного управління, місцевого самоврядування різної ланки, зацікавлених у дослідженні соціальних настроїв, протидії негативним ІВ будь-якої природи, для комерційних установ з метою визначені соціальних реакцій на дії менеджменту, наданих послуг, виготовленої та розповсюдження продукції будь-якої категорії тощо.

Своєчасна та адекватна реакція на виявлені ІЗ і недопущення їх переростання в небезпечні ситуації вимагає оперативних, достовірних та точних даних про: рівень їх небезпеки та пріоритетність у нейтралізації; тенденцій та динаміці розвитку; критичних часових точок та інтервали зміни динаміки розвитку.

2. Накопичений за результатами моніторингу масив інформаційних повідомлень підлягає кластеризації на сфери цінностей (економічна; політична; соціальна; духовна; правова; безпекова; свобод і прав особистості тощо) та по тональності (негатив / позитив / нейтральні). Кластеризація інформаційних повідомлень за суттю відноситься до задач розпізнавання образів із специфікою обробки природньої мови. Відомі підходи до кластеризації можна розділити на три групи: методи машинного навчання; техніки Word2vec, GloVe, RNNLM та класичні нейронні мережі; методи глибинного навчання з використанням великих мовних моделей.

Недоліки відомих підходів до кластеризації інформаційних повідомлень за принципами обробки природньої мови включають: препроцесінг – складність роботи в національному інформаційному просторі; суб'єктивізм в

кластеризація – необхідність врахування суб'єктивної спрямованості в кластеризації за тональністю; конкретика цільового спрямування – потребує врахування сфер та об'єктів ІВ; великий обсяг та значна динаміка зміни інформаційного контенту – потребує оперативного аналізу, зберігання та компресії вхідних даних.

Для кластеризації інформаційних повідомлень за тональністю та соціальними сферами варто використати методи глибинного навчання з використанням великих мовних моделей.

3. Реалізація процесів моніторингу інформаційного простору базуються на даних формату часового ряду – *time series*. Їх обробка доволі показово представлена: методами апроксимації (алгоритми ковзного середнього – *Moving Average (MA)* експоненційної апроксимації – *Exponential Moving Average (EMA)*, автокореляційні алгоритми типу *ARIMA*); методами статистичного навчання (згладжування) – *Statistical Learning* (метод найменших квадратів (МНК) – *Least Square Method (LSM)*, фільтрація Калмана тощо); методами глибинного навчання з використанням штучних нейронних мереж.

Аналіз показав, що класичні підходи до статистичного навчання потребують модифікації в напрямку застосування нелінійних за параметрами моделей. В той же час використання алгоритмів апроксимації не передбачається за можливе.

Стрімкий розвиток методів глибинного навчання з використанням штучних нейронних мереж для обробки часових рядів робить їх пріоритетними для застосування і розвитку.

4. Дисертаційна робота присвячена розв'язку актуальної є науково-прикладної задачі розробки і удосконалення математичних моделей багатofакторного аналізу та машинного навчання для обробки часових рядів.

Метою дисертаційних досліджень є підвищення достовірності (ймовірності виявлення і точності прогнозу) та повноти (множини

врахованих факторів) визначення домінантного інформаційного контенту в глобальних інформаційних мережах.

Для досягнення сформованої мети в дисертаційній роботі було вирішено комплекс взаємопов'язаних часткових завдань досліджень: аналіз існуючих підходів до прогнозування інформаційного контенту в глобальних інформаційних мережах; розробка математичної моделі кластеризації інформаційних повідомлень; розробка математичної моделі багатфакторного прогнозування активності контенту в глобальних інформаційних мережах; розробка технології багатфакторного прогнозування активності контенту в глобальних інформаційних мережах; розробка макету програмно-апаратного комплексу та оцінювання ефективності запропонованих рішень.

РОЗДІЛ 2

МАТЕМАТИЧНА МОДЕЛЬ КЛАСТЕРИЗАЦІЇ ІНФОРМАЦІЙНИХ ПОВІДОМЛЕНЬ З ІНДИКАТОРАМИ АКТИВНОСТІ ІНФОРМАЦІЙНОГО КОНТЕНТУ ЗА ТОНАЛЬНІСТЮ І СФЕРАМИ ДІЯЛЬНОСТІ СОЦІУМУ

Головна ідея та призначення математичної моделі кластеризації інформаційних повідомлень полягає у реалізації послідовності підготовчих етапів для виявлення критичної активності інформаційного контенту в глобальних медіа. По факту це встановлення переліку і визначення значень індикаторів, що є мірою активності контенту в реальних (первинних) повідомленнях і доведення їх до рівня часового ряду (Time Series) – систематизованого Data Set масиву. В умовах високої щільності потоку виникнення, динаміки розвитку та трансформації інформаційного контенту матимемо справу з Big Data структурою інформаційних повідомлень. Отже, модель кластеризації, окрім поділу за інформаційними ознаками, повинна забезпечувати компресією вхідних Big Data масивів.

Модель, що підлягає розробці має давати обумовлений опис та результати розрахунків для реалізації взаємопов'язаного переліку функціональних етапів:

6. Встановлення об'єктів моніторингу – інформаційних джерел (глобальних інформаційних мережах як сайти новин, інформаційні канали месенджерів, соціальні мережі), ієрархічно розподілених за рівнем довіри / достовірності інформації / популярністю. Перелік інформаційних джерел має давати репрезентативну вибірку соціальних медіа;
7. Встановлення системи індикаторів активності інформаційного простору – інфологічної моделі джерел, факторів та показників;
8. Моніторинг глобального інформаційного простору (щоденне / цілодобове) на рівні новин та збереження результатів моніторингу в

- форматі баз даних. Моніторинг матиме встановлену періодичність у часі із зверненням до встановленого переліку інформаційних джерел;
9. Встановлення належності новин до сфер цінностей соціуму;
 10. Встановлення тональності інформаційних повідомлень за визначеними сферами цінностей соціуму;
 11. Підрахунок кількості повідомлень за тональністю у кожній із сфер цінностей та каталогізацію отриманих вимірів в форматі часового ряду;
 12. OLAP (Online Analytical Processing) візуалізації результатів кластеризації для відображення первинної аналітики даних.

Математична модель формуватиметься як композитна складова описаних етапів із послідовною трансформацією інформаційних одиниць – міри активності інформаційного контенту до конкретного Data Set структури Time Series. Для отримання математичної моделі будуть застосовані методи системного аналізу та синтезу, індукції та дедукції, теорія множин та методи і технології обробки природньої мови – NLP.

Отримана математична модель орієнтована і передбачається для імплементації у конкретну технологію із розробкою програмної компоненти кластеризації інформаційних повідомлень з індикаторами активності інформаційного контенту за тональністю і сферами діяльності соціуму.

Верифікація моделі в цілому та її компонент, оцінювання адекватності здійснюватиметься емпіричними методами.

2.1. Інфологічна модель джерел, факторів та показників активності контенту в глобальних інформаційних мережах

Прогнозування активності інформаційного контенту в глобальних інформаційних мережах потребує вимірювання цього процесу через встановлення системи індикаторів (показників Key Performance Indicators KPI) активності. Перелік KPI має відображати контент інформаційних

повідомлень за різними сферами соціуму, бути об'єктивно обраховані (визначені) із прив'язкою до інформаційних повідомлень, що розповсюджуються засобами глобальних інформаційних мереж – через джерела первинних повідомлень / даних. Тобто, встановлювати взаємозв'язок об'єктивних мір (у числовому еквіваленті) реальних повідомлень у форматі природньої мови, що відображає їх контент з множини сфер цінностей соціуму. Зазначене можна описати інфологією предметної галузі – інфологічною моделлю джерел інформаційних повідомлень, факторів та показників активності контенту в глобальних інформаційних мережах.

Формування інфологічної моделі здійснюється в три етапи:

1. Математичний опис об'єктів моніторингу – інформаційних джерел;
2. Встановлення системи індикаторів активності інформаційного контенту;
3. Формалізований опис інфологічної моделі джерел, факторів та показників активності контенту в глобальних інформаційних мережах.

2.2.1. Математичний опис об'єктів моніторингу – інформаційних джерел

Першим етапом процесу багатофакторного прогнозування розвитку активності інформаційного контенту в глобальних інформаційних мережах є отримання даних – значень індикаторів активності. Індикатори, на цьому етапі ще не встановлені і є фокусом досліджень наступних етапів, однак їх перелік / опис, у т.ч. визначається джерелами первинних даних. Первинними даними вважатимуться *інформаційні повідомлення (ІП)*, що розповсюджуються інтернет виданнями (в глобальних інформаційних мережах) різної спрямованості, формату, форм власності та достовірності повідомлень. Головна вимога – спрямованість на новинний контент інформаційних повідомлень. Зазначені категорії видань називатимемо *інформаційними джерелами (ІД)*. Отримання первинних інформаційних

повідомлень – *первинних («сирих») даних* здійснюється в процесі моніторингу інформаційних джерел. Моніторинг ІД передбачає періодичний, але системний і монотонний (декілька разів, або раз на день – протягом заданого часу) перегляд вмісту інформаційних повідомлень на інформаційних джерелах – *об'єктах моніторингу (ОМ)* та збереження результатів моніторингу.

Моніторинг ІД в практичній імплементації – здійснюватиметься як технологія Web-скрапінгу (парсингу) за допомогою розробленої програмної компоненти, як складова програмного комплексу багатофакторного прогнозування активності контенту. Задачі моніторингу глобального інформаційного комплексу покладаються на компоненту програмно-апаратного комплексу – комп'ютерну систему як множина первинних автоматизованих робочих місць (АРМ) із доступом до глобальних інформаційних мереж (структура і функціонал розкрито у розділі 3).

Перелік джерел інформації має бути репрезентативним за кількість, складом, належністю і забезпечувати високі показники достовірності (ймовірності виявлення і точності прогнозу) та повноти (множини врахованих факторів) визначення домінантного інформаційного контенту в глобальних інформаційних мережах.

В роботі верифікація ІП та ІД не проводиться, що обумовлено необхідністю адекватного визначення домінантного контенту в глобальних медіа. Це пояснюється здебільшого відсутністю навичок володіння і застосування методів і технологій верифікації вмісту ІП у об'єктів ІВ – пересічних громадян соціуму.

З урахуванням зазначеного, встановлено ієрархічну структуру об'єктів моніторингу – глобальних медіа: сайти новин, інформаційні канали месенджерів, соціальні мережі тощо, з розподілом за рівнем довіри / достовірності інформації / популярністю тощо. Результати систематизовано у табл.1.1.

Множина ІД поділена за п'ятьма групами, по рівню верифікації та авторизації ІП відповідно до авторитету в медійному просторі, належності до форми власності та представництва у глобальному інформаційному просторі. Кожна група має різну за обсягом множину ОМ. Кількісний склад ОМ є варіабельним параметром. Множина ІД, які поставлені на спостереження для моніторингу утворюють множину об'єктів моніторингу вигляду (1.1):

$$\begin{aligned} OM &= [OM_{1i_1}, OM_{2i_2}, OM_{3i_3}, OM_{4i_4}, OM_{5i_5}], \\ i_1 &= 1 \dots N_{OM_1}, i_2 = 1 \dots N_{OM_2}, i_3 = 1 \dots N_{OM_3}, \\ i_4 &= 1 \dots N_{OM_4}, i_5 = 1 \dots N_{OM_5}. \end{aligned} \quad (2.1)$$

До кожної групи ОМ можливо встановити шкалу вагових коефіцієнтів, пропорційних мірі довіри до ІП на ІД:

$$\begin{aligned} G &= \begin{cases} OM_{1i_1} = 9 \dots 10, \\ OM_{2i_2} = 7 \dots 8, \\ OM_{3i_3} = 5 \dots 6, \\ OM_{4i_4} = 3 \dots 4, \\ OM_{5i_5} = 1 \dots 2, \end{cases} \\ i_1 &= 1 \dots N_{OM_1}, i_2 = 1 \dots N_{OM_2}, i_3 = 1 \dots N_{OM_3}, \\ i_4 &= 1 \dots N_{OM_4}, i_5 = 1 \dots N_{OM_5}. \end{aligned} \quad (2.2)$$

Вагові коефіцієнти визначається експертним шляхом та в найпростішому випадку приймають одиничні значення.

Таблиця 2.1

Ієрархічна структура об'єктів моніторингу

Категорія	Опис	Приклад	Довіра	Вагова шкала	Позначення
I.	Верифіковані та авторизовані ОМ державної / недержавної форм власності з «давньою» історією та що мають в якості ІД: ефірне мовлення; офіційні сайти; YouTube-представлення, можливо канал в месенджерах	https://tsn.ua/ https://novy.tv/ua/ https://ictv.ua/ua/ https://www.5.ua/ https://1plus1.ua/	1	9-10	OM_{1i_1} , $i_1 = 1 \dots N_{OM_1}$.
II.	Верифіковані та авторизовані ОМ державної / недержавної форм власності з «давньою» історією та що мають в якості ІД: офіційні сайти новин	https://www.ukrinform.ua/ https://www.unian.ua/ https://www.rbc.ua/ https://www.ukr.net/ https://www.pravda.com.ua/news/	2	7-8	OM_{2i_2} , $i_2 = 1 \dots N_{OM_2}$.

Продовження Таблиці 2.1

III.	Офіційні контент-месенджери державних установ	@DIUkraine @CenterCounteringDisinformation @ministry_of_defense_ua @AFUStratCom @verkhovnaradaukrainy	3	5-6	$OM_{3i_3},$ $i_3 = 1 \dots N_{OM_3}.$
IV.	Контент-месенджери новин приватного / комерційного спрямування	https://t.me/lachentyt https://t.me/torontotv https://t.me/newsgrpua https://t.me/oko_ua https://t.me/novinach	4	3-4	$OM_{4i_1},$ $i_4 = 1 \dots N_{OM_4}.$
V.	Соціальні мережі / спільноти	https://www.facebook.com/ https://www.instagram.com/	5	1-2	$OM_{1i_5},$ $i_5 = 1 \dots N_{OM_5}.$

Таким чином, множина ІД, категорованих у відповідності до ієрархії табл. 2.1 поставлених на спостереження ОМ (вир.(2.1)), до яких встановлені вагові коефіцієнти (2.2) – являє собою математичний опис об’єктів моніторингу – інформаційних джерел.

Зазначені рішення є основою для формування переліку і визначення значень індикаторів активності інформаційного контенту в глобальних інформаційних мережах.

Слід також зазначити, що наведений перелік ОМ в табл.2.1 синхронізовано із національним інформаційним простором. Однак, аналогічним чином може формуватися перелік ОМ для інформаційного простору будь-якої країни та мовного контенту. Отже, сформований математичний опис ОМ є кросфункціональним по відношенню до мови ІП та національної належності інформаційного простору.

Процес моніторингу ОМ реалізується ітераційно у часі і передбачає зберігання результатів моніторингу (див. рис.1.2). Результатом моніторингу є первинні масиви даних – інформаційних повідомлень (в лінгвістиці природньої мови) із прив’язкою до: ОМ, дати, часу. Первинні масиви даних – ІП із плином часу утворюватимуть Big Data структури. Зберігання результатів моніторингу здійснюватиметься в базі даних (БД), що за структурою матиме обов’язкові поля, подані табл.2.2.

Таблиця 2.2

Структура полів БД результатів моніторингу

Об’єкти моніторингу	Дата	Час	Інформаційне повідомлення
OM_{1i_1}	00.00.0000	t_{morning} ранок $t_{\text{afternoon}}$ пообідді t_{evening} вечір	<i>text</i>
OM_{2i_2}			
OM_{3i_3}			
OM_{4i_4}			
OM_{5i_5}			

Необхідність збереження первинних результатів моніторингу обумовлено:

для верифікації математичних моделей обробки Big Data структур інформаційних повідомлень;

формуванням Data Set масивів для використання методів глибинного навчання з метою виявлення критичної активності інформаційного контенту;

запровадження заходів ретроспективної аналітики з виявлення причинно-наслідкових зав'язків в критичній активності інформаційного контенту, що призводить до змін настроїв думок, дій в соціумі.

Отже, математичний опис об'єктів моніторингу (2.1), (2.2) масштабується в структуру полів БД, що дає Big Data структур інформаційних повідомлень – *text* (див. табл.2.2).

2.1.2. Індикатори активності інформаційного контенту

Багатогранність форм і методів ІВ, спрямованих на свідомість і підсвідомість індивіда / соціальних груп / спільноти, масованість і неперервність ІВ на сфери цінностей соціуму (див розділ 1.1) потребують дзеркального опису в множині індикаторів ІВ. Цей процес побудуємо із логічно обумовленого початку – із структурованого системного опису сфер цінностей соціуму. Це можливо розпочати із схеми рис.2.1 [1]. Будемо вважати, що критичні активності з позицій ІВ мають за фокус сім базових сфер цінностей: економічна, політична, безпека, соціальна, правова, духовна, права і свободи особистості. Зрозуміло, що такий поділ є умовний і всі ці категорії об'єднуються у свідомості / підсвідомості об'єктів впливу. Однак, такий поділ є природним з позицій механізмів реалізації ІВ, тому є доцільним для формування індикаторів активності.

В позначеннях (2.3) індекс i – позначає об’єкт моніторингу, j – час (день – доба, час доби) отримання інформаційного повідомлення від i -го джерела.

Текстові повідомлення (2.3) розгорнуті в таблицю БД результатів моніторингу (див. табл.2.2). Таким чином із плином часу утворюється Big Data масив первинних інформаційних текстових повідомлень, поданих природньою мовою.

Надалі процес формування індикаторів активності інформаційного контенту здійснюється через етапи:

первинної обробки інформаційних повідомлень;

кластеризації першого рівня – за сферами рис.2.1;

кластеризації другого рівня – за тональністю (негатив, нейтральне, позитив) повідомлень в кожній сфері цінностей і діяльності соціуму;

встановлення частоти (кількості) повідомлень відповідної тональності в кожній сфері переліку за сферами рис.2.1.

Перелічені етапи реалізуються рекурентно у часі для потоку інформаційних повідомлень (2.3). Зазначене дозволяє утворити структуру даних типу Time Serie частот тональності потоку інформаційних повідомлень для різних сфер цінностей соціуму.

Таким чином встановлюється причинно-наслідковий (інфологічний) зв'язок / модель джерел, факторів та показників активності контенту в глобальних інформаційних мережах з їх декомпозицією за сферами і тональністю. Це дозволяє встановити числову міру активності контенту на рівні кількісно-якісного відображення повідомлень в глобальному інформаційному просторі.

Вміст етапів розкриває суть причинно-наслідкових зав'язків, які в подальшому деталізуються методами обробки інформаційних повідомлень, поданих природньою мовою.

Первинна обробка інформаційних повідомлень. Первинні інформаційні повідомлення підлягають *конвеєру підготовчих етапів NLP*, зокрема реалізацію процесів: фільтрація; нормалізація; токенизація; видалення

стоп-слів; лематизація. Результатом є оброблені інформаційні масиви до рівня: фільтрованого повідомлення із збереженням контенту – $IP_{ij,filter}$ (NLP процеси містять лише етап фільтрації); лематизованого повідомлення – $IP_{ij,lemma}$ (NLP процеси містять усі перераховані раніше етапи до лематизації включно):

$$IP_{ij,filter} = f_{NLP,filter}(text_{ij}), \quad (2.4)$$

$$IP_{ij,lemma} = f_{NLP,lemma}(text_{ij}), \quad (2.5)$$

де $f_{NLP,filter}$ – позначає операції фільтрації текстових повідомлень, $f_{ij,lemma}$ – операції NLP конвеєру: фільтрація; нормалізація; токенизація; видалення стоп-слів; лематизація.

Сутність NLP процесів $f_{NLP,filter}$ та $f_{ij,lemma}$ будуть розглянуті нижче.

Інформаційні масиви $IP_{ij,lemma}$ – використовуються для оперативного частотного та ймовірнісного аналізу. Інформаційні повідомлення $IP_{ij,filter}$ підлягають подальшій поглибленій обробці шляхом їх кластеризації і формування індикаторів активності інформаційного контенту. В подальшому результати поглибленої обробки мають формат OLAP візуалізації та аналізу.

Кластеризація першого рівня – *сферична* передбачає встановлення приналежності інформаційних повідомлень до сфер цінностей і діяльності соціуму (рис.2.1). Вхідною інформацією є результати первинної фільтрації отриманих інформаційних повідомлень (2.4), які дають сферичні кластери за моделлю перетворень:

$$S_{area} = f_{NLP,S_{clustering}}(IP_{ij,filter}), area = 1 \dots 7, \quad (2.6)$$

$$S_{area} = [S_1, S_2, S_3, S_4, S_5, S_6, S_7],$$

де S_{area} – відображають множини кластеризованих за сферами цінностей і діяльності соціуму (рис.2.1) інформаційних повідомлень (2.4);

$f_{NLP, S_{clustering}}$ – операція кластеризації вхідних текстових повідомлень природною мовою за сферами рис.2.1 за технологіями NLP процесів (конкретика реалізації розкривається нижче).

Кластеризація другого рівня – *сферична тональність* передбачає встановлення тональності інформаційних повідомлень в кожній сфері цінностей і діяльності соціуму (рис.2.1) за категоріями: негатив (negative), нейтральне (neutral), позитив (positive). Вхідною інформацією для кластеризації другого рівня є множини S_{area} . Результати другого рівня кластеризації можна подати перетворенням:

$$S_{area, tonality} = f_{NLP, S_{tonality}}(S_{area}), area = 1 \dots 7,$$

$$S_{area, tonality} = \begin{bmatrix} S_{area, tonality} = \text{negative, or} \\ S_{area, tonality} = \text{neutral, or} \\ S_{area, tonality} = \text{positive.} \end{bmatrix}, \quad (2.7)$$

де $S_{area, tonality}$ – множини кластеризованих за тональністю інформаційних повідомлень S_{area} ;

$f_{NLP, S_{tonality}}$ – операція кластеризації вхідних текстових повідомлень природною мовою за тональністю з технологіями NLP процесів (також конкретика реалізації розкривається нижче).

Вторинна обробка кластеризованих інформаційних повідомлень – *частотний аналіз кластерів сферичної тональності* за вхідну інформацію приймають сформовані кластери (2.6). Задача етапу – встановити частоту / кількість повідомлень кожної із трьох тональностей із сфери цінностей рис.2.1 для кожного моменту часу j дискретних моментів часу моніторингу ОМ. Результатом етапу є частоти:

$$F_{j,area,tonality} = \begin{bmatrix} \hat{F}_{j,area,tonality} = \text{negative} \\ \hat{F}_{j,area,tonality} = \text{neutral} \\ \hat{F}_{j,area,tonality} = \text{positive} \end{bmatrix}, j = 1 \dots M. \quad (2.8)$$

де $\hat{F}$ – позначає середню за добу та за ОМ частоту появи негативних / нейтральних / позитивних інформаційних повідомлень зі сфери *area* – переліку рис.2.1.

Варто зазначити, що визначення середнього значення частоти (2.8) відбувається за тривіальним виразом:

$$\hat{F}_{j,area,tonality} = \sum_{l=1}^{l=3} \begin{bmatrix} F_{j,area,tonality} = \text{negative} \\ F_{j,area,tonality} = \text{neutral} \\ F_{j,area,tonality} = \text{positive} \end{bmatrix}, \quad (2.9)$$

$$l = 3 = t_{\text{morning}}, t_{\text{morning}}, t_{\text{evening}}.$$

Однак, середня за ОМ частоту визначається як середньозважена величина – пропорційна рівню довіри об'єкту моніторингу, наприклад за шкалою вагових коефіцієнтів G_l типу (2.2) у відповідності до виразу:

$$\hat{F}_{j,area,tonality} = \sum_{l=1}^{l=N_{OM}} G_l \begin{bmatrix} \hat{F}_{area,tonality} = \text{negative} \\ \hat{F}_{area,tonality} = \text{neutral} \\ \hat{F}_{area,tonality} = \text{positive} \end{bmatrix}, \quad (2.10)$$

$$l = N_{OM} = N_{OM_1} + N_{OM_2} + N_{OM_3} + N_{OM_4} + N_{OM_5}.$$

Таким чином, множини (2.8) відображають причинно-наслідковий (інфологічний) зв'язок / модель джерел, факторів та показників активності контенту в глобальних інформаційних мережах і являє собою структуру даних типу Time Serie – частот тональності потоку інформаційних повідомлень для різних сфер цінностей соціуму.

В процесі обробки реальних даних можливий перехід до відносних показників (2.10) шляхом формування їх відсоткових значень. Аналіз реальних даних буде проводитись як в абсолютних так і у відсоткових вимірах.

Відносні одиниці виміру дозволять обмінюватись інформацією із користувачами різної цільової спрямованості та дистанціюватись від повноти охоплення джерел моніторингу. Тоді як абсолютні значення дозволять аналізувати конкретику результатів моніторингу.

2.1.3. Формалізований опис інфологічної моделі джерел, факторів та показників активності контенту в глобальних інформаційних мережах.

Викладена послідовність етапів, сформований математичний опис об'єктів моніторингу – інформаційних джерел та індикатори активності інформаційного контенту дозволяють здійснити формалізований опис інфологічної моделі джерел, факторів та показників активності контенту в глобальних інформаційних мережах.

Інфологічна модель спрямована на забезпечення процесу виявлення критичної активності інформаційного контенту в глобальних медіа, здатних змінити настрої та призвести до небезпечних дій соціуму. Причинно-наслідкова модель дозволяє формувати багатофакторну систему індикаторів – вимірів – оцифрувань соціальних процесів, що протікають під впливом цільової інформації, яка поширюється в глобальному інформаційному просторі. Тобто, інфологічна модель дозволяє провести первинну формалізацію задачі з описом / вимірюванням реальної області до рівня адекватних абстракцій, придатних до подальшої обробки та аналізу.

Інфологічна модель формалізована у відповідності до уведених математичних описів і моделей (2.1) – (2.10) та подана у форматі табл.2.3 та схеми рис.2.2.

Причинно-наслідкові зв'язки інфологічної моделі мають такий лінгвістичний опис.

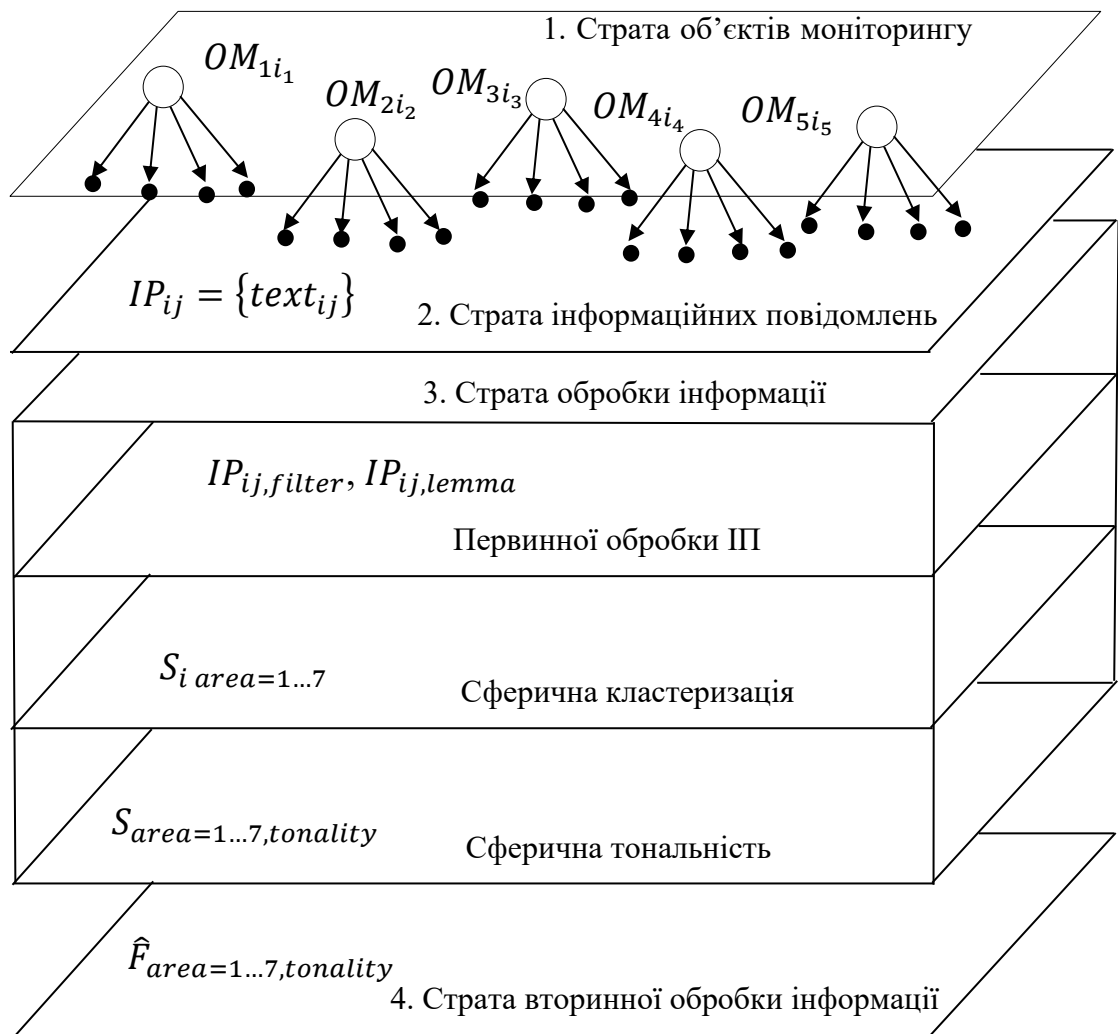


Рисунок 2.2. Страти (прошарки) інфологічної моделі джерел, факторів та показників активності контенту в глобальних інформаційних мережах

Інформаційний контент, його обговорення, що присутні в глобальних інформаційних мережах як сайти новин, інформаційні канали месенджерів, соціальні мережі, підлягає постійному цілодобовому моніторингу. Це реалізується через спостереження за репрезентативним переліком об'єктів моніторингу (2.1) із встановленими мірами довіри до них (2.2). Об'єкти моніторингу формують першу *страту (прошарок) об'єктів моніторингу*, що має безпосереднє віддзеркалення реального глобального інформаційного простору. Це імітує сприйняття інформаційного контенту об'єктом інформаційного впливу – людиною (індивідом, соціальною групою,

соціумом). Страта об'єктів моніторингу формують перший стовпчик табл.2.2 та перший рівень схеми рис.2.2.

Результатом моніторингу ОМ є Big Data масив інформаційних повідомлень (2.3) – первинних даних, поданих природньою мовою. Масиви ІП зберігаються у відповідній базі даних. Інформаційні повідомлення, здобуті як результат спостереження за ОМ, формують *другу страту (прошарок) інформаційних повідомлень* (другий стовпчик табл.2.2 та другий рівень схеми рис.2.2.).

Страту обробки інформації (третя) – інформаційних повідомлень природною мовою формують процеси (див. стовпчики 3-5 табл.2.3 та відповідні рівні схеми рис.2.2):

первинної обробки ІП, моделі (2.4), (2.5);

кластеризації першого рівня – сферична кластеризація (2.6);

кластеризація другого рівня – сферична тональність (2.7).

Ієрархічна компресія вхідних Big Data масивів ІП формує *четверту страту вторинної обробки інформації*. Результатом є структури даних типу Time Serie – частоти тональності потоку інформаційних повідомлень для різних сфер цінностей соціуму (рис.2.1).

Таким чином, реалізовано формалізований опис інфологічної моделі джерел, факторів та показників активності контенту в глобальних інформаційних мережах у відповідності до уведених математичних описів і моделей (2.1) – (2.10) та подана у форматі табл.2.3 та схеми рис.2.2. Інфологічна модель через кінцевий перелік індикаторів (2.8) відображає причинно-наслідковий (інфологічний) зв'язок / модель джерел, факторів та показників активності контенту в глобальних інформаційних мережах.

Для утворення системи індикаторів за множиною інформаційних джерел, сферами цінностей та у часі анонсовано застосування перелік етапів первинної обробки, кластеризації та вторинної обробки інформаційних повідомлень природньої мови за технологіями NLP.

Таблиця 2.3

Інфологічна модель джерел, факторів та показників активності контенту в глобальних інформаційних мережах

Об'єкти моніторингу	Інформаційні повідомлення (2.3)	Первинна обробка ПП (2.4), (2.5)	Кластеризація 1 рівня (2.6): сфера	Кластеризація 2 рівня: тональність (2.7)		Вторинна обробка – частота (2.8), (2.9), (2.10)
1	2	3	4	5	6	7
OM_{1i_1}	$IP_{i_1j} = \{text_{i_1j}\}$	$IP_{i_1j,filter}$ $IP_{i_1j,lemma}$	$S_{i_1area=1...7}$	$S_{area=1...7,tonality} = \text{negative},$ $S_{area=1...7,tonality} = \text{neutral},$ $S_{area=1...7,tonality} = \text{positive}.$	Ієрархічна компресія вхідних Big Data	$\hat{F}_{area,tonality} = \text{negative},$ $\hat{F}_{area,tonality} = \text{neutral},$ $\hat{F}_{area,tonality} = \text{positive},$ $area = 1 \dots 7.$
OM_{2i_2}	$IP_{i_2j} = \{text_{i_2j}\}$	$IP_{i_2j,filter}$ $IP_{i_2j,lemma}$	$S_{i_2area=1...7}$	$S_{area=1...7,tonality} = \text{negative},$ $S_{area=1...7,tonality} = \text{neutral},$ $S_{area=1...7,tonality} = \text{positive}.$		
OM_{3i_3}	$IP_{i_3j} = \{text_{i_3j}\}$	$IP_{i_3j,filter}$ $IP_{i_3j,lemma}$	$S_{i_3area=1...7}$	$S_{area=1...7,tonality} = \text{negative},$ $S_{area=1...7,tonality} = \text{neutral},$ $S_{area=1...7,tonality} = \text{positive}.$		
OM_{4i_4}	$IP_{i_4j} = \{text_{i_4j}\}$	$IP_{i_4j,filter}$ $IP_{i_4j,lemma}$	$S_{i_4area=1...7}$	$S_{area=1...7,tonality} = \text{negative},$ $S_{area=1...7,tonality} = \text{neutral},$ $S_{area=1...7,tonality} = \text{positive}.$		
OM_{1i_5}	$IP_{i_5j} = \{text_{i_5j}\}$	$IP_{i_5j,filter}$ $IP_{i_5j,lemma}$	$S_{i_5area=1...7}$	$S_{area=1...7,tonality} = \text{negative},$ $S_{area=1...7,tonality} = \text{neutral},$ $S_{area=1...7,tonality} = \text{positive}.$		

Їх композиція дає синергетичний ефект в компресії Big Data масивів до частоти повідомлень в множені факторів адекватного відображення інформаційних впливів, що реалізуються в глобальних інформаційних мережах.

Сформовані таким чином система індикаторів (2.8) вже є репрезентативною для первинного аналізу і може утворювати багатовимірний простір OLAP (Online Analytical Processing) візуалізації результатів обробки первинних інформаційних повідомлень. Структура OLAP куб результатів, первинної обробки, кластеризації та вторинної обробки подана на рис.2.3.

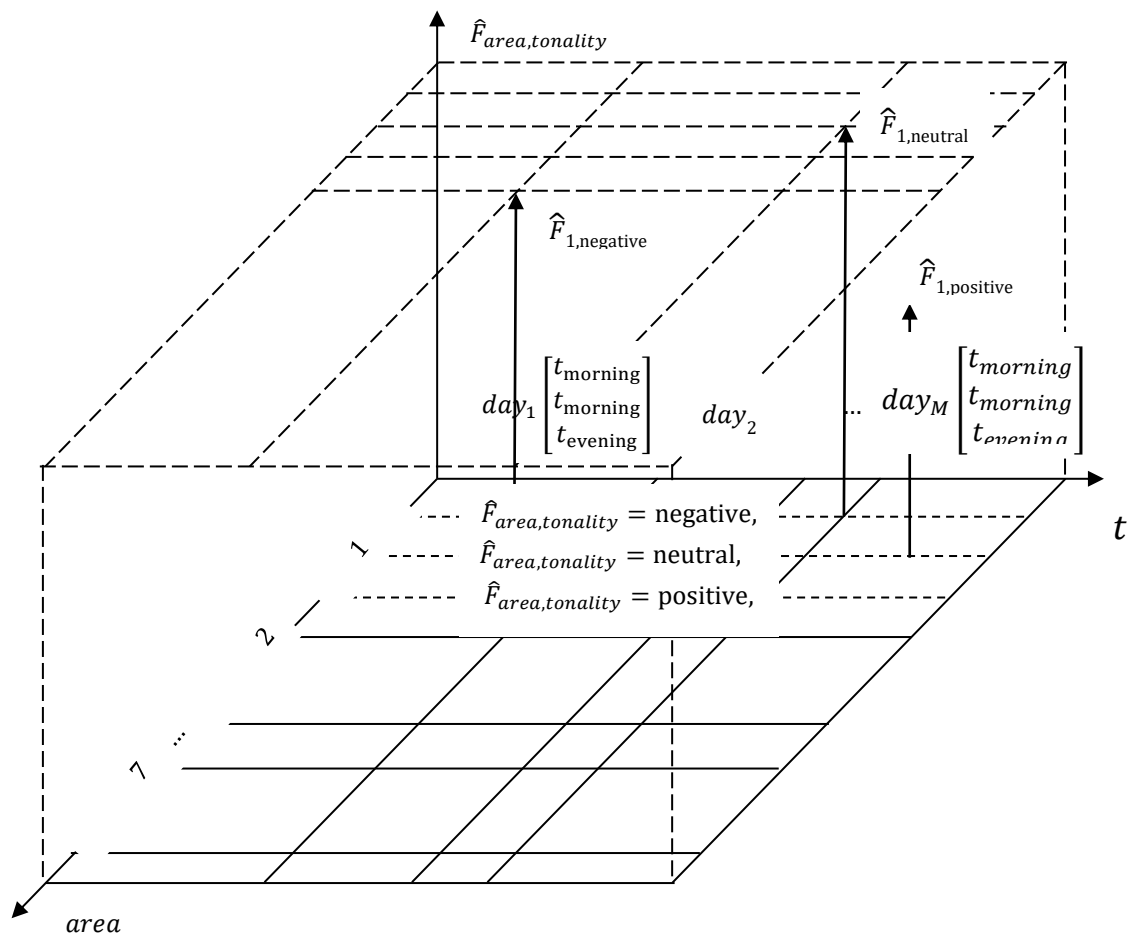


Рисунок 2.3. Структура OLAP куб результатів, первинної обробки, кластеризації та вторинної обробки інформаційних повідомлень

OLAP – куб візуалізації рис.2.3 має тривимірний простір зміни частоти тональності інформаційних повідомлень $\hat{F}_{area,tonality}$ у часті $t = day_M$ для сфери цінностей соціуму $area$. Це дозволяє проводити оперативний (в реальному часу) аналіз активності інформаційного контенту на час спостереження (в OnLine режимі) та в первинному вимірі факторів та сфери цінностей. Тобто «прослуховувати» контент трафік інформаційних повідомлень глобальних інформаційних мереж в близькому до реального часу. Оцінювання часової оперативності (продуктивності) буде здійснено в технологічному розділі роботи (див. розділ3).

2.2. Математична модель кластеризації інформаційних повідомлень

Для прогнозування розвитку активності інформаційного контенту пропонується реалізувати комплекс взаємопов'язаних етапів:

1. Моніторинг глобального інформаційного простору (щоденне / цілодобове) на рівні новин та збереження результатів моніторингу в форматі баз даних. Моніторинг матиме встановлену періодичність у часі із зверненням до встановленого переліку джерел;
2. Встановлення належності новин до сфер цінностей соціуму (2.1);
3. Встановлення тональності інформаційних повідомлень за визначеними сферами цінностей;
4. Підрахунок кількості інформаційних повідомлень за тональністю у кожній із сфер цінностей та здійснити їх усереднення із каталогізацію отриманих вимірів в форматі часового ряду;

Зазначений перелік задач формує інфологічну моделі джерел, факторів та показників активності контенту в глобальних інформаційних мережах.

Реалізації підлягають етапи обробки первинних інформаційних повідомлень природною мовою (2.3):

1. Первинна обробка інформаційних повідомлень. – *конвеєр підготовчих етапів NLP*: фільтрація; нормалізація; токенизація; видалення стоп-слів; лематизація;

2. Кластеризація першого рівня – *сферична* передбачає встановлення приналежності інформаційних повідомлень до сфер цінностей і діяльності соціуму (рис.2.1);

3. Кластеризація другого рівня – *сферична тональність* передбачає встановлення тональності інформаційних повідомлень в кожній сфері цінностей і діяльності соціуму (рис.2.1) за категоріями: негатив (negative), нейтральне (neutral), позитив (positive);

4. Вторинна обробка кластеризованих інформаційних повідомлень – *частотний аналіз кластерів сферичної тональності*. Задача етапу – встановити частоту / кількість повідомлень кожної із трьох тональностей із сфери цінностей рис.2.1 для кожного моменту часу j – дискретних моментів часу як частоти $\hat{F}$ –середня за добу та за ОМ частоту появи негативних / нейтральних / позитивних інформаційних повідомлень зі сфери цінностей переліку рис.2.1.

Реалізація зазначених етапів розкриває суть запропонованої *математичної моделі кластеризації інформаційних повідомлень*.

Суть математичної моделі відображає структурна схема рис.2.4 та реалізація наступних етапів.

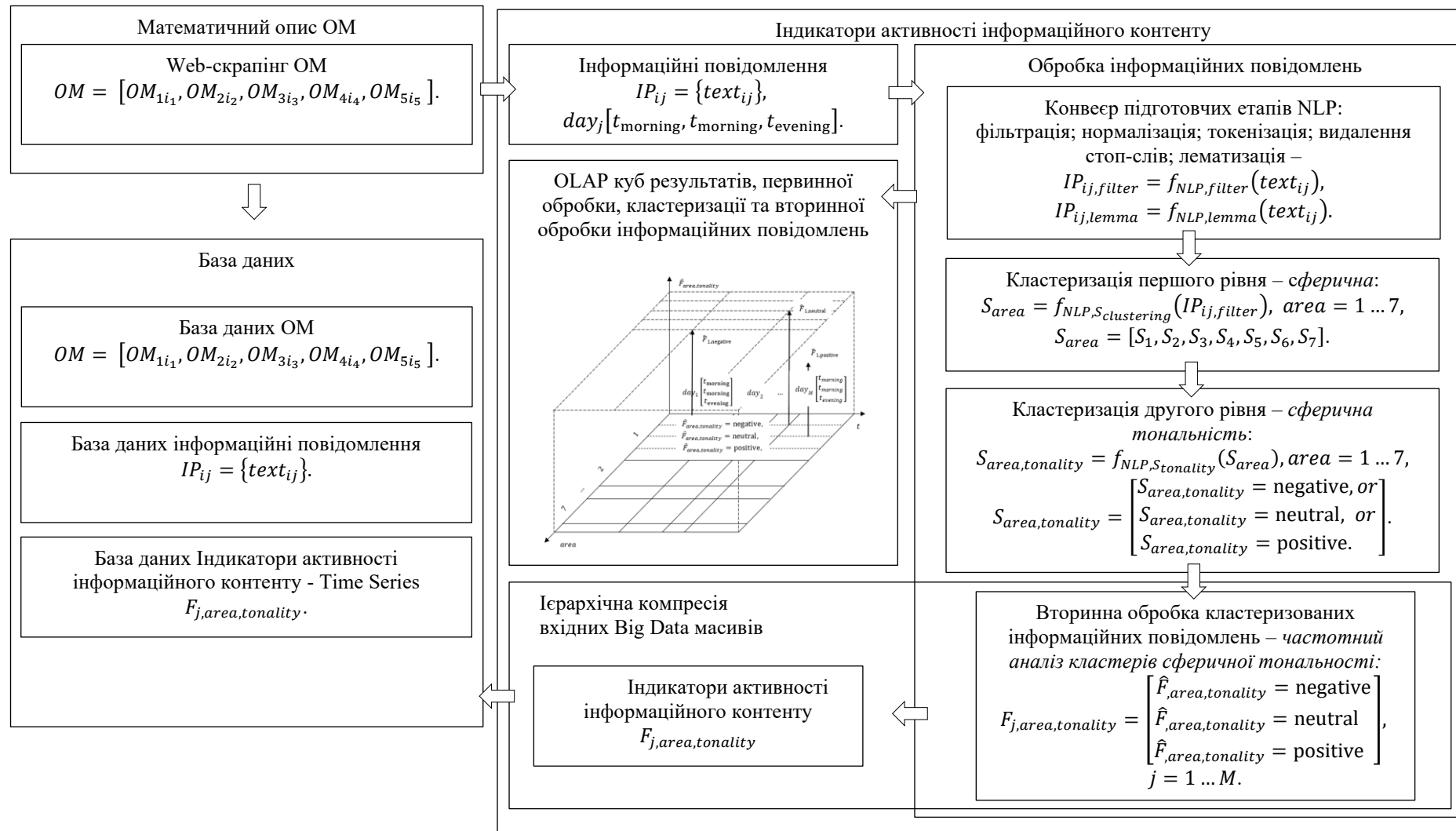


Рисунок 2.4. Структурна схема математичної моделі кластеризації інформаційних повідомлень

2.3. Верифікація та оцінювання ефективності застосування математичної моделі кластеризації інформаційних повідомлень

Для визначення функціональної придатності до практичного застосування, верифікації структури і вмісту та оцінювання ефективності запропонованої математичної моделі кластеризації інформаційних повідомлень розроблено програмний модуль моніторингу інформаційних джерел та кластеризації інформаційних повідомлень. В основу модуля покладено розроблену математичну модель кластеризації (див. структурну схему рис.2.4).

Архітектура програмного модуля подана на рис.2.5. Структура проекту, та робоче вікно програмного модуля подані на рис.2.6, 2.7. Текст програмних компонент наведено в Додатку Б. В основу програмної реалізації програмного модуля покладено технології python, підтримувані профільними для NLP та підтримуючими бібліотеками: bs4, spacy, nltk, textblob, spacytextblob, deep_translator, vaderSentiment.

Програмний модуль являє собою консольний застосунок як складова, що підтримує процес отримання та кластеризації вхідних даних для технології багатофакторного прогнозування активності контенту в глобальних інформаційних мережах.

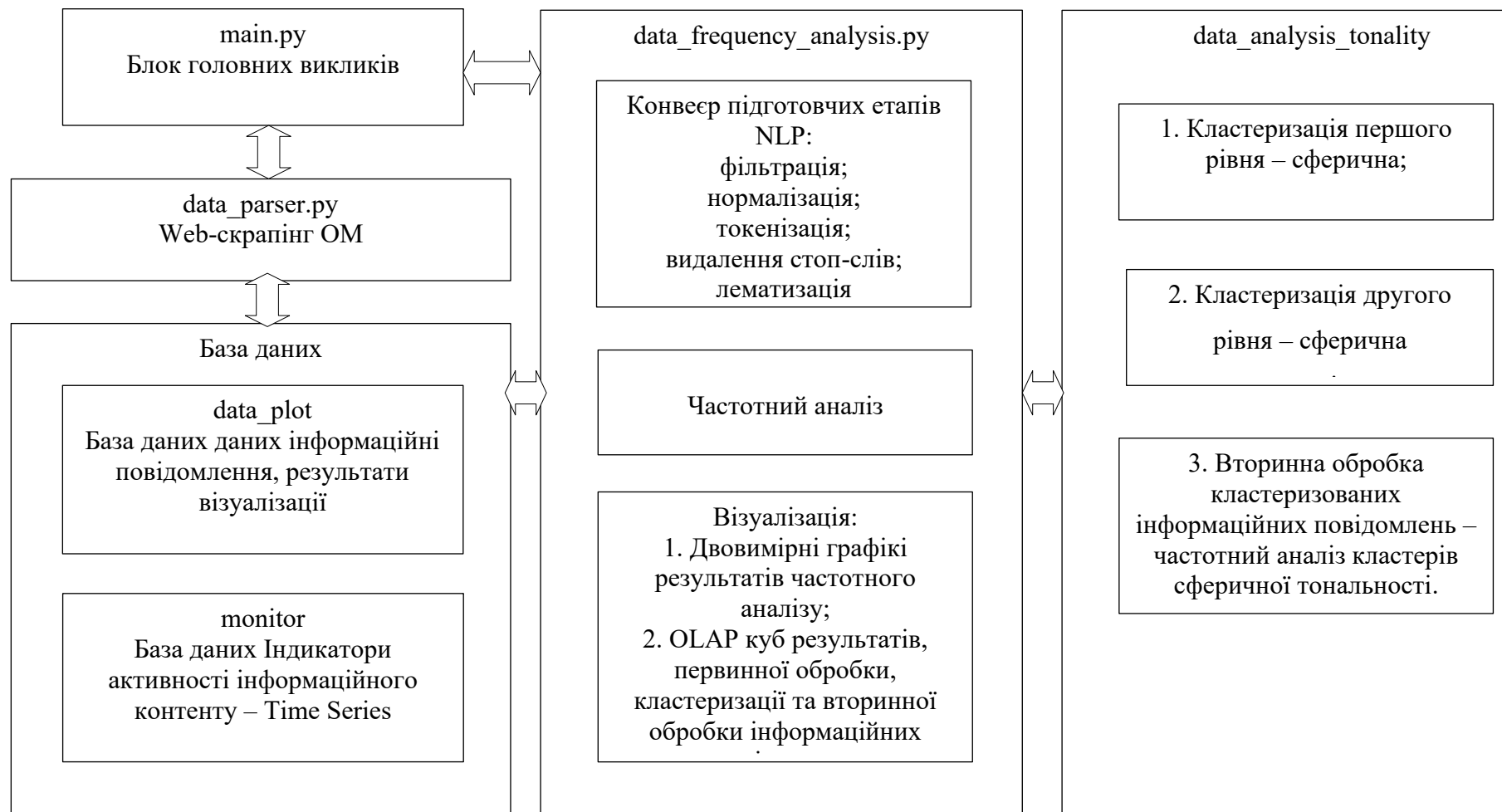


Рисунок 2.5. Архітектура програмного модуля моніторингу ІД та кластеризації ІП.

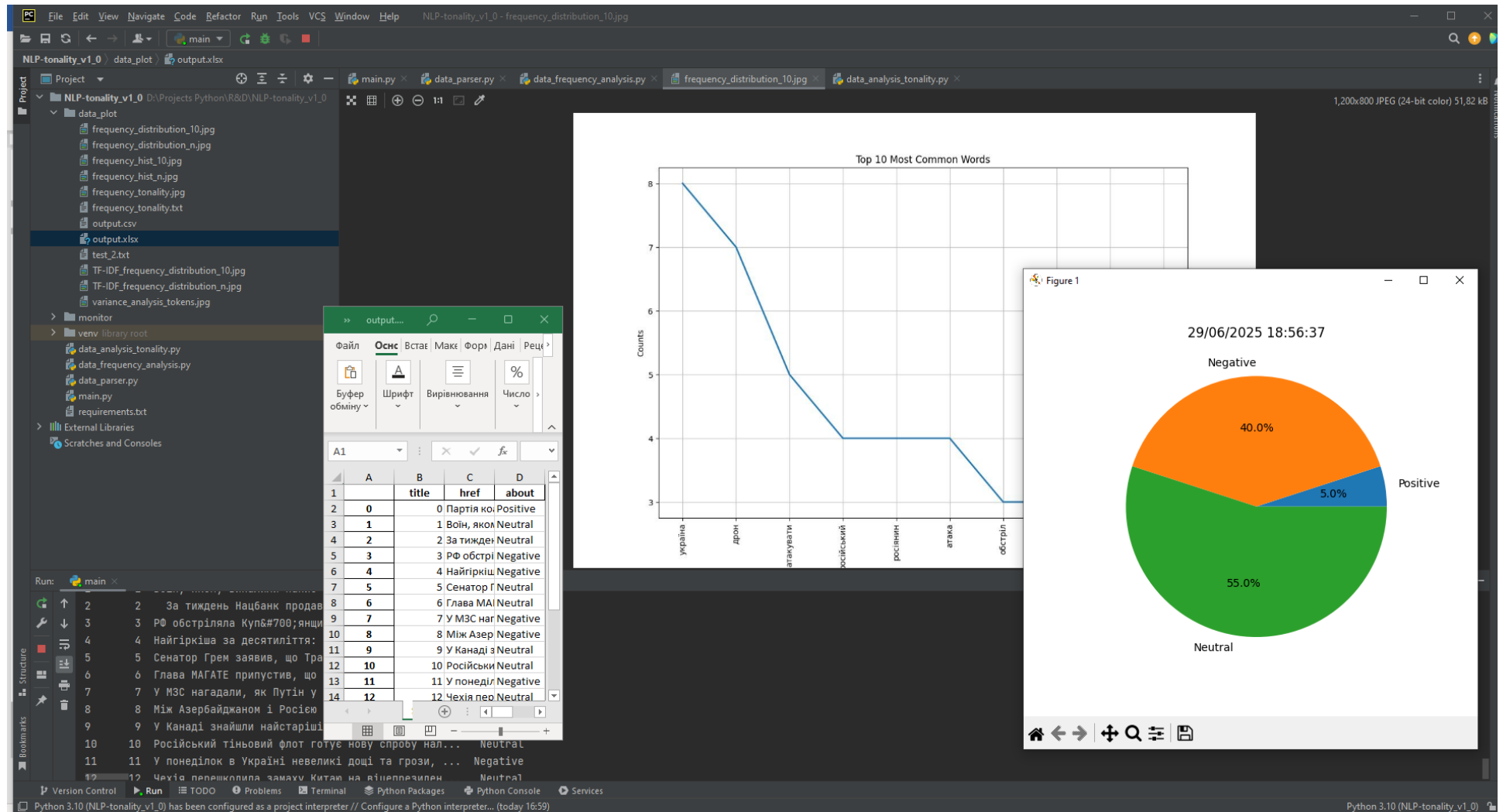


Рисунок 2.6. Робоче вікно програмного модуля моніторингу ІД та кластеризації ІІ

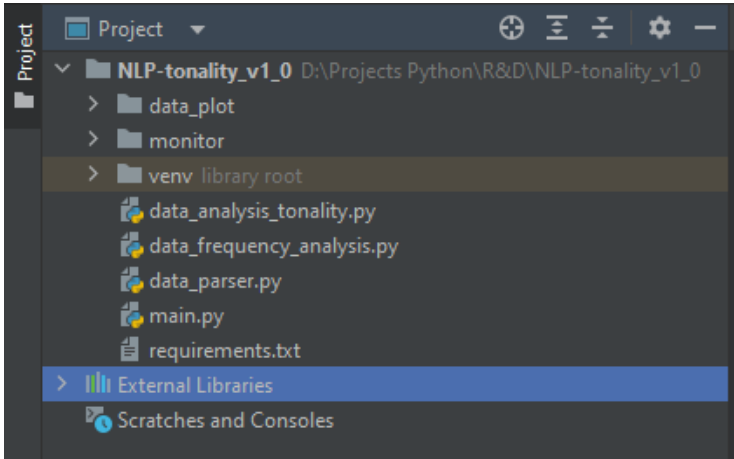


Рисунок 2.7. Структура проекту програмного модуля моніторингу ІД та кластеризації ПП

Функціонування модулю регламентується етапами, наведеними на структурній схемі математичної моделі кластеризації інформаційних повідомлень (рис.2.4) у відповідності до її структурної реалізації рис.2.5.

Вхідною є інформація про перелік об’єктів моніторингу, фіксовано заданих в модулі data_parser.py. В результаті Web-скрапінгу ОМ із заданого переліку формуються інформаційні масиви із проміжним збереженням у файлі, які потім потрапляють до бази даних первинних повідомлень – див., наприклад рис.2.8.

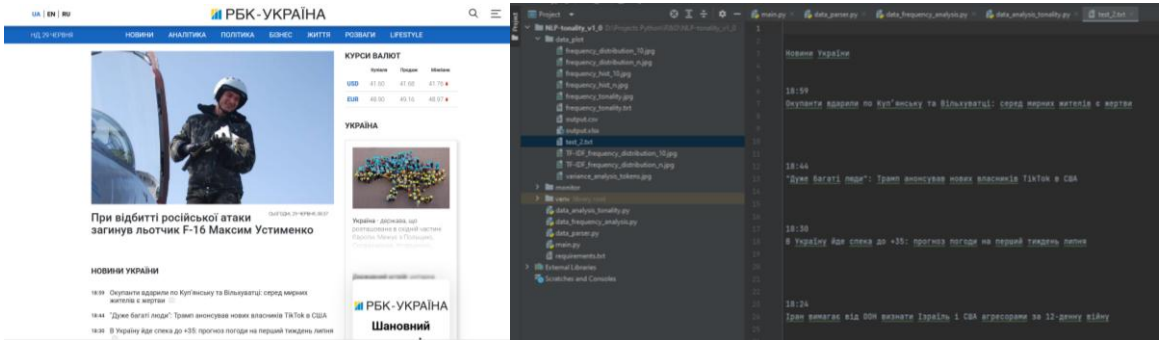


Рисунок 2.8. Інформаційні повідомлення

Web-скрапінг та первинний аналіз / фільтрація інформаційних повідомлень реалізовано з використанням технологій високорівневої мови програмування python: requests, bs4.

Моніторинг реалізовано за переліком ОМ табл.2.1 із створенням статичної множини парсерів. Моніторинг відбувався на трьох добових зрізах: ранок (08:00-10:00), обід (12:00-14:00), вечір (17:00-20:00). Верифікація та оцінювання ефективності застосування математичної моделі кластеризації інформаційних повідомлень здійснювалось за результатами моніторингу, отриманими протягом 2-х тижнів. Тобто отримана вибірка в 42 виміри / результатів парсингу інформаційного контенту, яка містить понад 1000 інформаційних повідомлень.

На отриманому сегментів вхідних даних розгорнуто ітераційне (кожного з 14 днів на ранок, день, вечір) застосування запропонованої математичної моделі кластеризації інформаційних повідомлень.

Технологічний базис python та результати реалізації кожного етапу математичної моделі рис.2.4 у відповідності до архітектури її програмної реалізації рис.2.5 наводяться далі.

1. Первинна обробка інформаційних повідомлень. – *конвеєр підготовчих етапів NLP*: фільтрація; нормалізація; токенизація; видалення стоп-слів; лематизація – реалізовано методами бібліотек: spacy, nltk, re. Візуалізація результатів проведено з використанням методів бібліотек: matplotlib, seaborn.

Вхідні інформаційні повідомлення (рис.2.8) підлягали первинній фільтрації – відокремлення стрічки новин, прибирання зайвих символів (динаміку процесів див.2.9).

Нормалізація даних передбачає зведення усіх повідомлень до малої літеральної форми.

```
<script type="text/javascript">window.client&#x2D;p = 189.130.54.189;/</script>  
<title>Новина Україна - останні новини дня в Україні сьогодні, свіжачі огляди | РБК-Україна</title><meta content="Новина Україна, українські новини, новина України сьогодні, українські новини, новина України онлайн, новина контент">Більше новин України сьогодні онлайн! Останні новини дня в Україні і світі, новини політики, економіки, спорту, фінансів та ділових новин в Інтернеті." name="description"/><link href="https://www.rbc.ua/rus/" style="type="text/css">  
    <!--before--><webkit-box-sizing: border-box;box-sizing: border-box;>  
    .container {max-width:980px;width:980px;min-width:760px;margin:0 auto;display:flex;align-items: stretch;justify-content:space-between;}  
    .branding-layout .container {position:relative;z-index:3;}  
    .branding-layout .main-menu {position:absolute;z-index:3;max-width: 980px;width: 980px;min-width: 980px;margin: 44px auto 20px;}  
    .branding-layout {background-position: center 57px;}  
    .middle-col {width:100%;max-width:680px;order:1;margin:0;}  
    .right-col {width:300px;min-width:300px;order:2;margin:0;position:relative;}  
    .right-sticky {position: -webkit-sticky;position: sticky;top:140px;overflow: hidden;vertical-align: baseline;margin-top:20px;}  
    .MIXADVERT.NET {margin-bottom:20px;}  
    #catfish-wrapper{width:100%;bottom:0;position: fixed;z-index:999;text-align: center;}  
    #catfish-wrapper #catfish-close {position: absolute:right: 0;top: 0;z-index: 1000;cursor: pointer;width: 32px;height: 32px;}  
    @media (max-width: 1024px) {.container {width:980px};}  
    @media (max-width: 1024px) {.container {width: 980px;min-width:980px}.middle-col {max-width:680px;padding-left:20px;}}  
  
    </style>  
<link href="https://www.rbc.ua/rus/news" hreflang="uk" rel="alternate"/>  
<link href="https://www.rbc.ua/ukr/news" hreflang="ru" rel="alternate"/>
```

Новина України

13:14

Благословие війни проти України: СБУ заочно засудила митрополита УПЦ КП

13:29

"Metinvest Politehnika" оголосив про набір на фахову передачу oscity

13:25

У Берліні дали оптимістичний прогноз щодо 18-го пакету санкцій проти Росксії

Виділена стрічка новин <https://www.rbc.ua/rus/news>

Рисунок 2.9. Інформаційні повідомлення

Токенізація відбувається на рівні відокремлення стоп-слів, оскільки надалі реалізується у т.ч. частотний аналіз тексту.

Лематизація дозволяє уніфікувати однакові терміни для спрощення частотного аналізу. Етапи нормалізації, токенізації та лематизації демонструють результати рис. 2.10 в динаміці трансформації тексту з переліку рис.2.9.

modified_text =	новини україни	благословив війну проти україни	сбу заочно засудила мітрополита упц ім	"метінвест політехніка" оголосив про набір на фахову передавчу освіту
Кількість слів:	635			
Кількість унікальних слів:	456			

Нормалізація інформаційних повідомлень

[illegible]

Токенізація, видалення стоп-слів та лематизація інформаційних повідомлень

Рисунок 2.10. Інформаційні повідомлення

2. Кластеризація першого рівня – сферична передбачає встановлення приналежності інформаційних повідомлень до сфер цінностей і діяльності соціуму (рис.2.1). Зазначений етап реалізовано із впровадженням технології великих мовних моделей (LLM – Large Language Model) та відповідної технологійної бази python.

3. Кластеризація другого рівня – *сферична тональність* передбачає встановлення тональності інформаційних повідомлень в кожній сфері

цінностей і діяльності соціуму (рис.2.1) за категоріями: негатив (negative), нейтральне (neutral), позитив (positive) в програмному модулі реалізовано за допомогою інструментів 3-х бібліотек:

1. Textblob:TextBlob;
2. vaderSentiment.vaderSentiment: SentimentIntensityAnalyzer;
3. nltk.sentiment.vader: SentimentIntensityAnalyzer.

Мовна узгодженість українського інформаційного простору за повідомленнями із можливостями вбудованих бібліотек зазначених засобів з визначення тональності python досягнута шляхом перекладу оригінальних інформаційних повідомлень на англійську мову методами deep_translator: GoogleTranslator.

Кожен із аналізаторів тональності експериментальним шляхом налаштовувався на суб'єктивність оцінки: позитив / нейтральний / негативний – до української аудиторії (поріг спрацювання часткових та узагальненого показників з діапазону -1...+1). В основі базових технологій з визначення тональності є токени (слова), що породжують негативні емоції. Налаштування порогових спрацювань спрямовано на врахування особливостей національної аудиторії.

Надалі рішення про тональність приймалось за критерієм 2/3 на користь негативної оцінки у разі неузгодженості рішень. Зазначена логіка описується фрагментом коду:

```
def tonality_rezalt(number_Positive, number_Negative, number_Neutral):
    if number_Negative >= 2: tonality_return = 'Negative'
    if (number_Positive >= 2) and (number_Negative <= 1): tonality_return =
'Positive'
    if (number_Neutral >= 2) and (number_Negative <= 1): tonality_return =
'Neutral'
    return tonality_return
```

Результатом є тональність інформаційних повідомлень для кожної сфери та на кожен час повідомлення. Результати роботи програмного модуля для одного зрізу у часі подані на рис.2.11.

```
result_list = [{'title': 0, 'href': 'Міністрів освіти, охорони здоров'я та культури можуть замінити -',  
df = title href about  
0 0 Міністрів освіти, охорони здоров'я та культури... Neutral  
1 1 Росіяни уперше атакували Харків дроном нового ... Negative  
2 2 ЦПД: Росія посилює кампанію з дискредитації є... Neutral  
3 3 Гейм, сет і матч: вчимося краще розуміти велик... Positive  
4 4 Фінал "Гри в кальмара" "потопив" акції корейсь... Negative  
5 5 Росіяни вдарили дроном по Херсону: поранено чо... Neutral  
6 6 Земля утримує вдвічі більше тепла, ніж 20 рокі... Negative  
7 7 Чеська родина заробляє на деталях для російськ... Neutral  
8 8 Під час атаки дронів на Кривий Ріг ворог поціл... Negative  
9 9 В Україні подорожчала картопля, а буряк подеше... Neutral  
10 10 Винищувачі F-35 з осені захищатимуть ключовий ... Neutral  
11 11 В Україні зріс тіньовий ринок сигарет: у Kanta... Negative  
12 12 Трамп змусив Канаду скасувати податок на велик... Negative  
13 13 "Чутливе слово" – і товар зупиняється: Китай б... Negative  
14 14 У Чорнобильському заповіднику показали окопи ч... Negative  
15 15 Фонд держмайна продає підприємство на Дніпропе... Neutral  
16 16 Захисник збірної України з хокею перейшов у по... Neutral  
17 17 У Росії третина таксопарків опинилася під загр... Negative  
18 18 ДБР: Чотирьох військових з київської частини а... Negative  
19 19 Європейські порти забиті вщент: мита Трампа сп... Negative  
date and time = 30/06/2025 13:44:43
```

30/06/2025 15:41:26

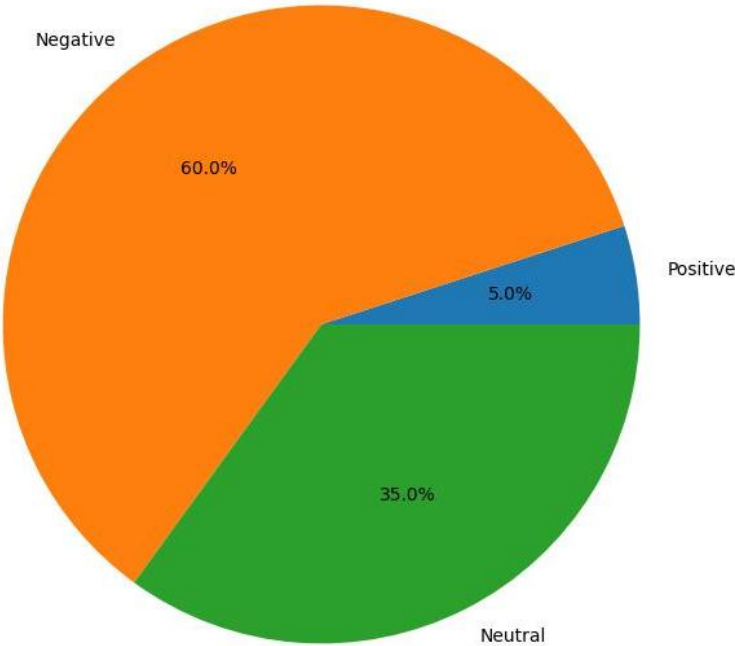


Рисунок 2.11. Результати визначення тональності інформаційних повідомлень

Оцінювання ефективності результатів кластеризації ІІ за тональністю реалізовано на статистиці у 100 повідомлень для кожної сфери діяльності соціуму. Порівняння автоматизованих рішень здійснювалось за узгодженою оцінкою експертів. Експертами виступали посадові особи –

споживачі результатів кластеризації (див. акти впровадження додатку А). Розрахунки показали ймовірність вірної ідентифікації $P = 0,87$ (ефективність – похибки I роду) та ймовірність промаху (втрати – похибки II роду) $F = 0,2$. Для ймовірностних задач оцінювання активності інформаційного контенту зазначені показники ефективності є прийнятними для використання результатів кластеризації.

4. Вторинна обробка кластеризованих інформаційних повідомлень – частотний аналіз кластерів сферичної тональності – реалізовано з використанням бібліотек python: `spacy`, `nlTK`.

Частотний аналіз дозволяє оцінити домінантну тематику інформаційних повідомлень та динаміку їх зміни у часі. Кінцева мета – отримати об'єктивні індикатори активності інформаційного контенту у глобальному інформаційному просторі – частота появи токенів в ІІІ.

Частотний аналіз проведено за різним метриками та у послідовності графічної візуалізації результатів та формування часових рядів.

Результати частотного аналізу у вигляді графіків на конкретний дискретний момент часу показані на рис.2.12.

На час моніторингу в домінанті токени / слова: Україна, дрон, росія, деталь, що відображають домінантний контент – атаки дронів по Україні. Аналіз частоти інших термінів показує прийнятність та доцільність фокусування в частоті повторення слів лише на домінантному контенті. Про це свідчать діаграми частоти та розподілу токенів / слів усіх типів. Тобто, використання в якості показника активності інформаційного контенту частоти повторення слів є виправданим та адекватним.

За отриманими даними в ході 14-денного моніторингу за частотною тональністю інформаційних повідомлень можна провести просторово-часовий OLAP – аналіз отриманих та сформованих даних. Результати візуалізації у форматі OLAP кубів подані на графіках рис.2.13.

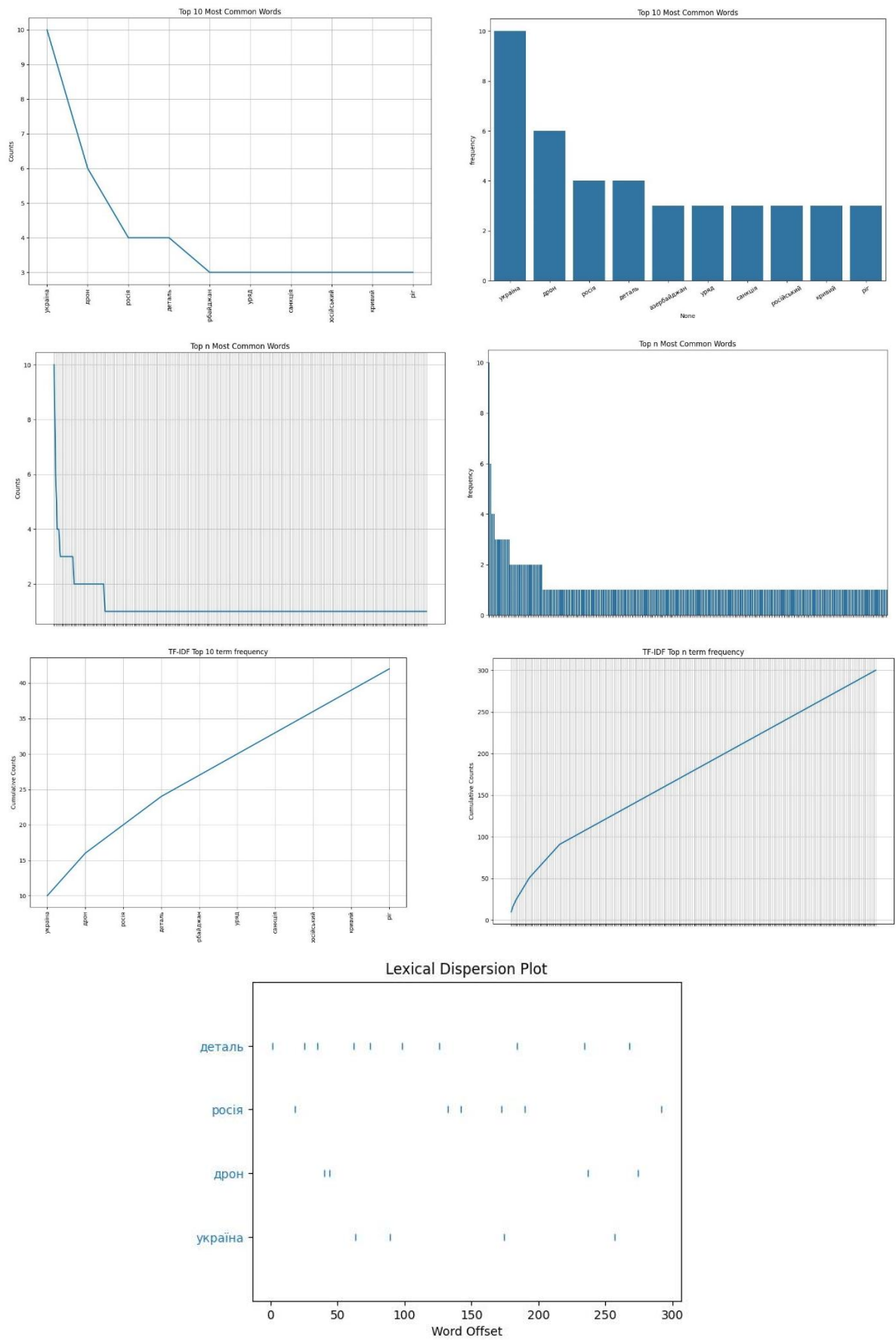
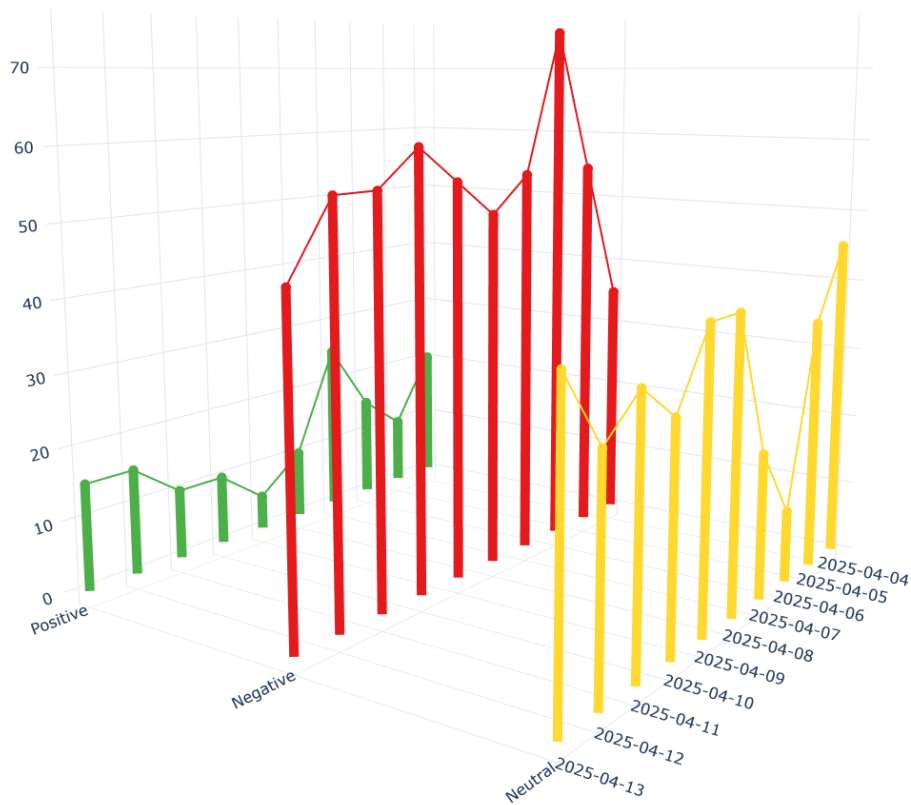
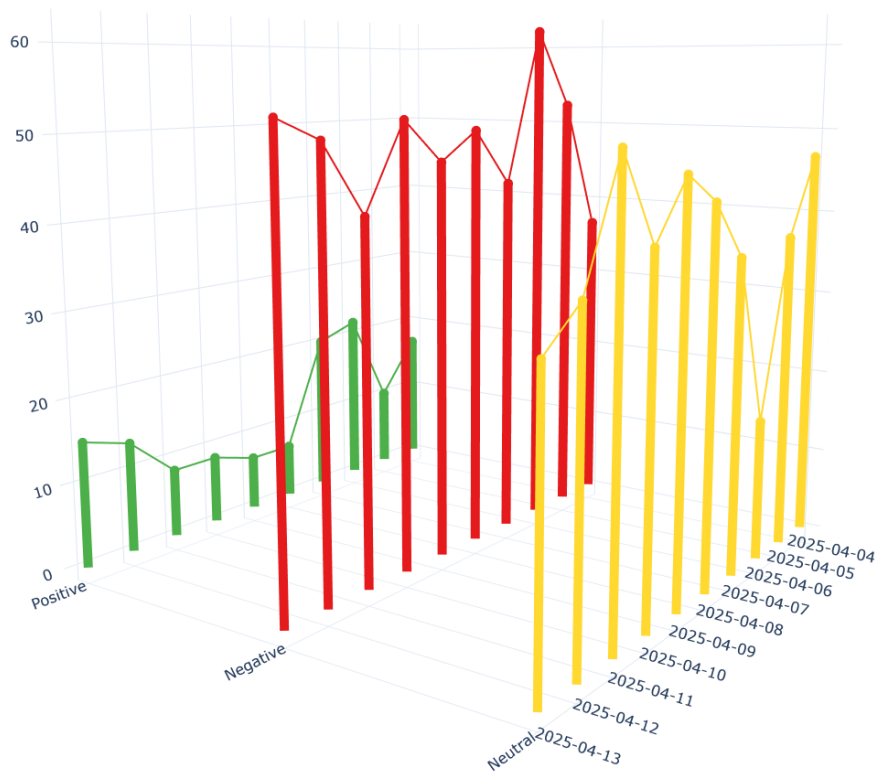


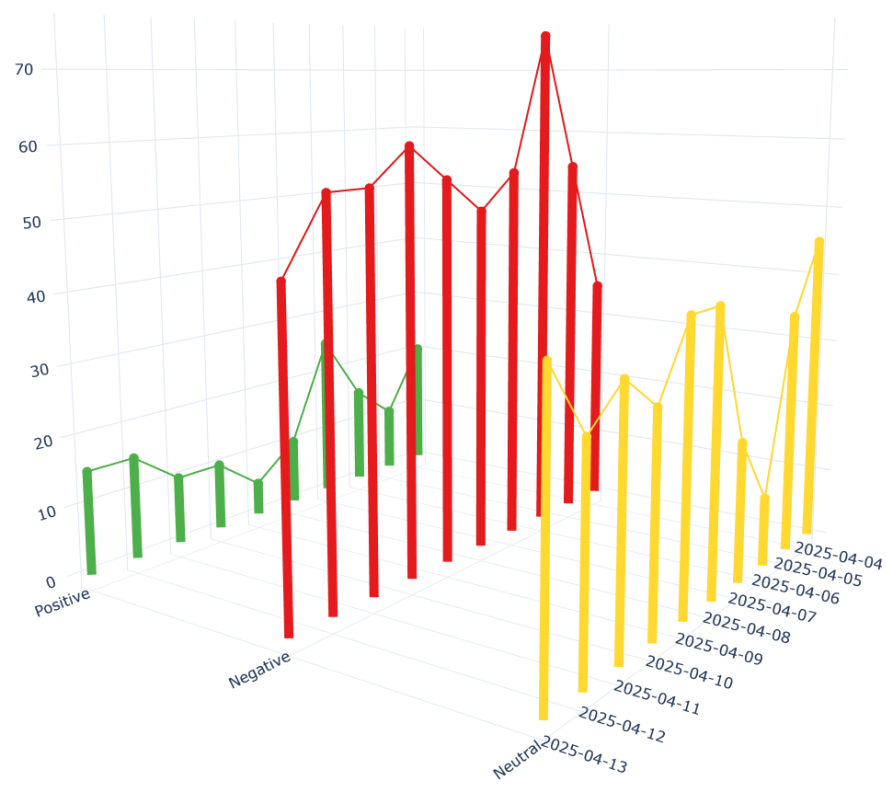
Рисунок 2.12. Результати частотного аналізу ІІІ



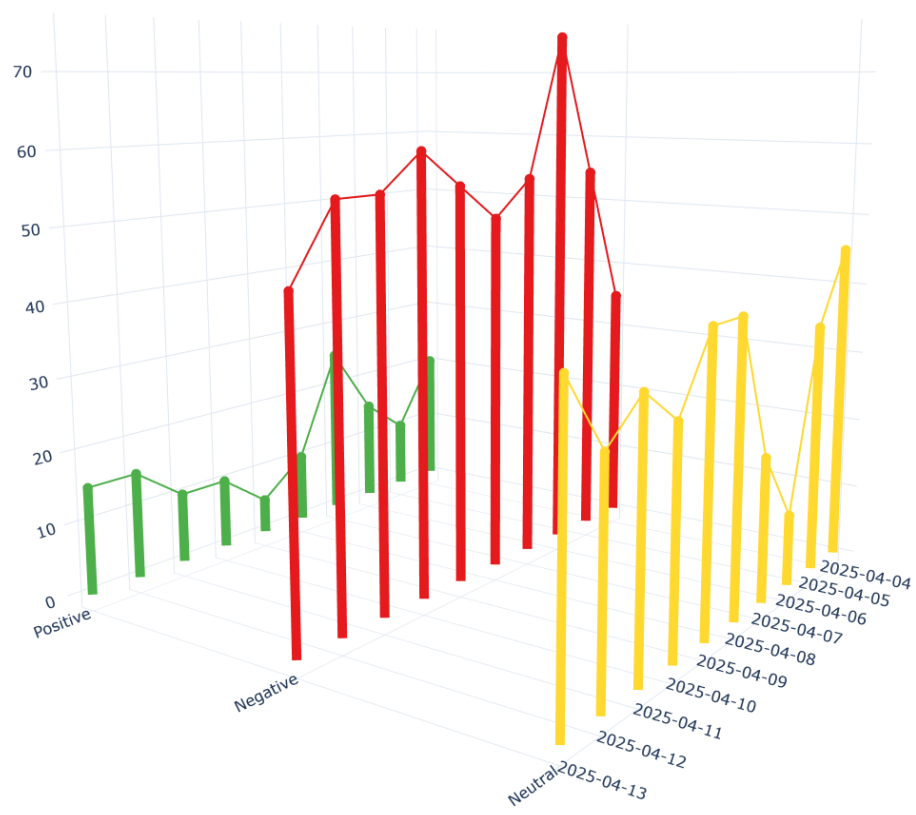
Зміна у часі тональності на ранок



Зміна у часі тональності на обід



Зміна у часі тональності на вечір



Середня за день зміна тональності

Рисунок 2.13. OLAP – аналіз отриманих та сформованих даних

Аналіз графіків рис.2.13 дозволяють сформулювати наступні висновки. Наразі в інформаційному просторі кількість негативних повідомлень привалює над позитивними. Це спостерігається увесь період моніторингу. Співвідношення позитив / негатив / нейтральні повідомлення протягом доби мають відносно стабільний баланс. Протягом часу зміна активності контенту за тональністю має коливальний монотонний тренд з випадковим шумом. Пікова (аномальна) кількість негативних повідомлень спостерігаються на фактах резонансних подій – для періоду моніторингу зокрема [68, 69]: 04.04.25, вечір повітряна атака на Рив Ріг (05.04.25 – ранок аномалії в кількості негативних інформаційних повідомлень) [<https://fakty.com.ua/ua/proisshestvija/20250405-vybuhy-u-kryvomu-rozi-4-kvitnya-2025-roku-shho-vidomo/>]; 06.04.25, ранок Київ – повітряна нічна атака ракет та БПЛА [<https://fakty.com.ua/ua/proisshestvija/20250406-vybuhy-u-kyievi-u-troh-rajonah-stolyczi-stalysya-pozhezhi-cherez-raketnu-ataku-rf/>]. При чому друга подія на фоні першої має нижчий рівень активності.

Таким чином, сформовані індикатори активності інформаційного контенту як частота тональності інформаційних повідомлень відображають реальні події і динаміку інформаційних повідомлень, а відтак є придатними для практичного застосування.

Висновки до 2 розділу

У ході досліджень отримала подальший розвиток математична модель кластеризації інформаційних повідомлень, яка базується на частотному аналізі їх тональності з використанням методологій обробки природної мови (Natural Language Processing (NLP)) з підтримкою великих мовних моделей; OLAP візуалізації результатів кластеризації та вирізняється встановленою системою індикаторів активності інформаційного контенту за сферами діяльності соціуму з ієрархічною компресією вхідних Big Data масивів, що обумовлює модель бази даних для їх зберігання. Це забезпечує підвищення повноти аналізу інформаційних повідомлень в глобальних інформаційних мережах шляхом врахування множини факторів за сферами діяльності соціуму.

Для визначення функціональної придатності до практичного застосування, верифікації та оцінювання ефективності запропонованої математичної моделі кластеризації інформаційних повідомлень розроблено програмний модуль моніторингу інформаційних джерел та кластеризації інформаційних повідомлень.

Розрахунки показали наступні показники ефективності кластеризації за тональністю інформаційних повідомлень: ймовірність вірної ідентифікації / ефективність $P = 0,87$; ймовірність промаху / втрати $F = 0,2$. Зазначені показники ефективності є прийнятними для практичного використання результатів кластеризації.

Результати OLAP аналізу довели, що наразі в інформаційному просторі кількість негативних повідомлень привалює над позитивними. Це спостерігається увесь період моніторингу. Співвідношення позитив / негатив / нейтральні повідомлення протягом доби мають відносно стабільний баланс. Протягом часу зміна активності контенту за тональністю має коливальний монотонний тренд з випадковим шумом. Пікова (аномальна) кількість негативних повідомлень спостерігаються на

фактах підтверджених резонансних подій. Таким чином, сформовані індикатори активності інформаційного контенту як частота тональності інформаційних повідомлень відображають реальні події і динаміку інформаційних повідомлень, а відтак є придатними для практичного застосування.

Таким чином, запропоновані рішення з кластеризації інформаційних повідомлень є ефективними та придатними до практичного застосування.

РОЗДІЛ 3.

МАТЕМАТИЧНА МОДЕЛЬ БАГАТОФАКТОРНОГО ПРОГНОЗУВАННЯ АКТИВНОСТІ КОНТЕНТУ В ГЛОБАЛЬНИХ ІНФОРМАЦІЙНИХ МЕРЕЖАХ

В результаті моніторингу інформаційного простору (ОМ табл.2.1) та застосування математичної модель кластеризації інформаційних повідомлень маємо часові ряди (time series) параметрів (2.8) – (2.10). А саме, величини частот / кількість повідомлень, віднесених до сфер цінностей рис.2.1, які розділені на тональності (негатив, позитив, нейтральне) і прив'язані до дискретних моментів часу моніторингу ОМ.

Надалі зазначені інформаційні масиви підлягають обробці: підготовчі етапи обробки time series (дослідження властивостей, визначення статистичних характеристик, виявлення та локалізація впливу аномалій); побудова і навчання моделі поведінки (обробка ретроспективних біг дата масивів); прогнозування розвитку time series поза межами інтервалу спостереження.

Поточні та прогнозовані результати приводяться до інтегрованого показника за нелінійною схемою компромісів. Інтегрований показник аналізується на наявність біфуркаційного інтервалу (асимптотичне зростання), що може спричинити біфуркаційний перехід у настроях та діях соціуму під впливом негативних інформаційних повідомлень в різних сферах. Сумісний сигнатурний та окремий аналіз поведінки часових рядів з OLAP візуалізацією дозволяє оцінити причини появи біфуркаційних інтервалів – наслідків. Це може бути використано для своєчасної реакції у протидії негативним інформаційним повідомленням. Зазначені процеси реалізуються з використанням запропонованої математичної моделі багатофакторного прогнозування активності контенту в глобальних інформаційних мережах, яка базується на обробці та аналізі часових рядів

з використанням методів статистичного та глибинного машинного навчання і відрізняється від відомих рішень:

запропонованими підходами до виявлення і очищення вхідних даних від аномальних вимірів;

використанням нелінійних за параметрами моделей в методах статистичного навчання;

модифікацією згорткових нейронних мереж для обробки часових рядів із впровадженням нелінійної схеми компромісів в моделі нейронної структури та моделюванням навчальної вибірки для екстраполяції за методами статистичного навчання;

визначенням інтегрованої оцінки активності інформаційного контенту за нелінійною схемою компромісів як результат багатофакторного аналізу вектору факторів за сферами діяльності соціуму;

пошуком прогнозованих біфуркаційних інтервалів як критичної активності інформаційного контенту;

Це забезпечує підвищення достовірності (ймовірності виявлення і точності прогнозу) домінантної активності контенту в глобальних інформаційних мережах.

Головна ідея запропонованого рішення полягає у застосуванні для обробки сформованих time series модифікованих методів двох класів: статистичного навчання; глибинного навчання. Кожен з двох класів методів обробки застосовується до time series кожної групи і тональності. В межах кожного класу методів обробки time series реалізується агрегація отриманих оцінок за нелінійною схемою компромісів. Тим самим формується узагальнена оцінка міри ІВ відносно єдиного об'єкту інформаційного впливу.

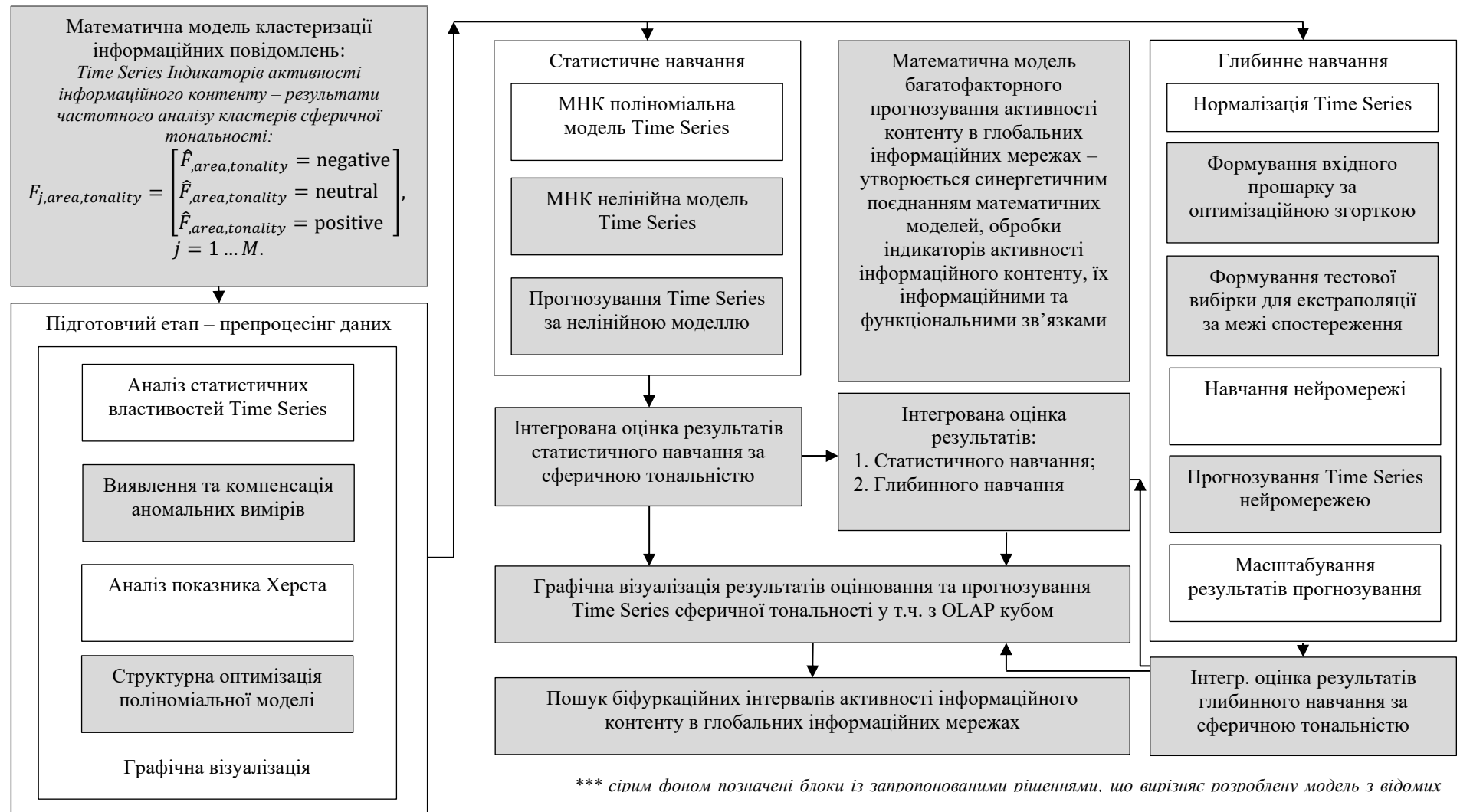


Рисунок 3.1. Структурна схема етапів математичної моделі багатфакторного прогнозування активності

Таким чином, маємо три (за кількістю класів методів обробки) узагальнені оцінки та динаміку їх розвитку на ділянці спостереження і прогнозу. Отримана надмірність підвищує достовірність результуючої оцінки та використовується для емпіричного аналізу результатів моніторингу інформаційного простору.

Структурна схема зазначених етапів, що відображає сенс, інформаційні і функціональні зв'язки окремих етапів математичної моделі багатофакторного прогнозування активності контенту в глобальних інформаційних мережах подана на рис.3.1. Зміст етапів розкривається відповідними математичними моделями пунктів цього розділу.

3.1. Процеси багатофакторного прогнозування активного інформаційного контенту. Підготовчий етап.

3.1.1. Первинний аналіз властивостей часових рядів та визначення їх статистичних характеристик

Первинний аналіз Time Series $F_{j,area,tonality}$, $j = 1 \dots N$ (2.8), (2.10) проведено з метою визначення їх властивостей:

- рівнодискретність / нерівнодискретність;
- рівноточність / нерівноточність;
- наявність / відсутність випадкових складових похибок;
- наявність / відсутність аномалій, їх тип – грубі виміри, пропуски;
- стохастичні властивості вибірки – стаціонарність / нестаціонарність
- наявність / відсутність тренду, його тип: лінійний, нелінійний, асимптотичний, коливальний (сезонний).

Для реалізації первинного аналізу застосовано об'єктивно суб'єктивні методи: візуалізації; аналізу; оцінки статистичних характеристик, розрахунків і аналіз показника Херста.

Розрахунок статистичних характеристик проведено в припущенні наявності тренду за методикою [41, 54].

1. Виділення систематичної складової Time Series – тренду.
2. Розрахунок числових характеристики закону розподілу випадкової складової похибок в Time Series, з оцінок математичного сподівання m_y , дисперсії D_y , середнього квадратичного відхилення (СКВ) σ_y . За необхідності визначається достовірності апроксимації R^2 (coefficient of determination) – індикатор якості реалізації п.1.

$$\begin{aligned}
 m_y &= \frac{1}{n} \sum_{i=1}^n \Delta_i, \\
 D_y &= \frac{1}{n-1} \sum_{i=1}^n (\Delta_i - m_y)^2, \sigma_y = \sqrt{D_y}, \\
 R^2 &= 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}, \\
 \Delta_i &= |y_i - \hat{y}_i|, \\
 \bar{y}_i &= \frac{1}{n} \sum_{i=1}^n y_i.
 \end{aligned} \tag{3.1}$$

В 3.1. прийняті позначення: y_i – вимірне значення елементу часового ряду – позначає обраний для обробки індикатор з множини $F_{j,area,tonality}$, $j = 1 \dots N$; $\hat{y}_i$ – оцінка виміряного значення.

3. На стаціонарність часовий ряд перевіряється з контролем умов: математичне очікування $m\{y_t\} = a = const$, дисперсія $D\{y_t\} = \sigma^2 = const$.

4. Знаходження закону розподілу випадкової складової реалізується за допомогою графічного відображення часового ряду у формі гістограми – частоти потрапляння реалізацій y_i в множину порційних довірчих інтервалів в межах діапазону зміни дисперсії D_y часового ряду.

Показник / індекс Херста дозволяє аналізувати тенденції розвитку часового ряду за ретроспективною вибіркою з виявленням факту зростання / убуття / невизначеність [41]. Результати аналізу показника Херста використовується для подальшої обробки і аналітики даних.

Алгоритм розрахунку показника Херста для часового ряду y_i .

1. Для часового ряду y_i визначаються математичне сподівання m_y , та СКВ σ_y (3.1).

2. Часовий ряд y_i розбивається на множину інтервалів $y_j, j = 1 \dots m$, з $l = 1 \dots n_j < N$ вимірів.

3. На кожному інтервалі y_j визначається лінійне відхилення:

$$\delta_j = \sum_{l=1}^{n_j} (y_{jl} - m_y). \quad (3.2)$$

4. Надалі розраховується величина розмаху часового ряду:

$$R = \max \delta_j - \min \delta_j. \quad (3.3)$$

6. Розрахунок показника Херста здійснюється за виразом:

$$H = \frac{\log(R / \sigma_y)}{\log(N/2)}. \quad (3.4)$$

Показник Херста змінюється в межах $[0, 1]$ і є мірою схильності часового ряду до трендів у поведінці:

якщо $H > 0,5$ – то ретроспективна поведінка часового ряду ймовірно буде зберігатись і у майбутньому;

якщо $H < 0,5$ – ймовірно часовий ряд змінить свою поведінку у майбутньому;

значення $H = 0,5$ – свідчить про непередбачуваність / випадковість у поведінці часового ряду у майбутньому.

Фрагмент файлу вхідних даних $F_{j,area,tonality}$ наведено на рис.4.7.

Виділення систематичної складової часових рядів (п.1 алгоритму розрахунку статистичних характеристик) здійснювався за класичним алгоритмом методу найменших квадратів [41, 57]. Результати візуалізації вхідних даних наведено на графіках табл.4.1.

Результати розрахунків для конкретної реалізації Time Series наведено на графіках табл.4.1.

Аналіз отриманих результатів дозволяє сформулювати перелік властивостей Time Series $F_{j,area,tonality}$, $j = 1 \dots N$. Отримані дані, що відображають активність контенту в глобальних інформаційних мережах є

- рівнодискретними – оскільки частота вимірювання відповідає значенню один раз (або три рази з оцінкою середнього) на день;
- рівноточні – відбувається за єдиними підходами у встановленні $F_{j,area,tonality}$ відносно сталого переліку ОМ і характеризують єдине явище – активність інформаційного контенту;
- мають випадкові похибки з нормальним законом;
- присутні аномальні виміри типу «грубі виміри», пропуски даних – відсутні;
- спостерігається стаціонарні властивості вибірки;
- присутній тренд нелінійного асимптотично-коливального характеру із сезонними повтореннями.

Виявлені властивості визначають зміст і порядок реалізації подальших етапів обробки сформованих часових рядів.

3.1.2. Очищення часового ряду від аномальних вимірів

Первинний аналіз властивостей часових рядів, що підлягають обробки довів наявність аномальних вимірів (AB) типу грубі виміри. Дослідження довели, що причинами появи AB в Time Series $F_{j,area,tonality}$ є короткочасні інформаційні приводи, що публікуються в глобальних інформаційних мережах і які не мають подальшої актуальності / підтримки.

Наявність AB в Time Series, залежно від їх кількості, абсолютного значення, обсягу вибірки може значно погіршувати результати обробки та прогнозування розвитку часового ряду. Тому в тренді процесів обробки Time Series типу $F_{j,area,tonality}$ передбачено алгоритми виявлення та відновлення AB.

Враховуючи що цей процес відбуватиметься на Big Data масивах, поверх основних етапів обробки – такі алгоритми мають бути ефективні, відносно прості, швидкими та враховувати апіорну невизначеність та наявність інформації про властивості вхідних даних..

Враховуючи зазначене у ході досліджень розроблено три алгоритму, побудовані на використанні принципу ковзного вікна.

1. Алгоритм попередньої обробки – sliding_wind.

Суть алгоритму полягає у реалізації згладжування тривіальним ковзним вікном. У вікні реалізовано розрахунок статистичного середнього, або середньозваженого. Розмір ковзного вікна визначається вимогою квазілінійності тренду, що описує зміну досліджуваного процесу у ковзному вікні.

Етапи алгоритму sliding_wind.

1. Вхідною є статистична вибірка $X = \{x_i\}$, $i = 1 \dots n$.
2. Формування ковзного вікна розміром N_{win} вимірів.
3. Визначення оцінки статистичного середнього:

$$\hat{x}_j = \frac{1}{Nwin} \sum_{j=1}^{Nwin} x_j, j = 1 \dots Nwin. \quad (3.5)$$

4. Формування очищеної вибірки $\hat{X} = \{\hat{x}_i\}$, $i = 1 \dots n$ шляхом заміни елементів вибірки X оцінками статистичного середнього $\hat{x}_j$: $x_{i=Nwin} = \hat{x}_j$, починаючи з останнього виміру, включеного до ковзного вікна.

5. Зміщення ковзного вікна на наступний вимір вибірки праворуч – на $i = i + 1$.

6. Повторення п. 2-5 в межах вхідної статистичної вибірки $i = 1 \dots n$.

7. Після обробки повної вибірки, з метою корегування даних в межах першого ковзного вікна з вхідної статистичної вибірки формується підвибірка об'ємом $2Nwin$. Надалі на підвибірці здійснюється повторення п.2-5 в зворотному до попереднього порядку напрямі.

II. Алгоритм контролю змін статистичних властивостей вибірки – алгоритм medium.

Суть алгоритму полягає у виявленні аномальних вимірів шляхом визначення еталонних параметрів статистичної вибірки (математичне сподівання, середньоквадратичне відхилення) без аномалій та їх порівняння з поточними статистичними характеристиками. Виявлені аномалії замінюються статистичним середнім. Еталонні та поточні статистичні параметри розраховуються за вибірками, що утворюють початкове та поточне ковзні вікна відповідно.

Етапи алгоритму medium.

1. Вхідною є статистична вибірка $X = \{x_i\}$, $i = 1 \dots n$.

2. В межах першого ковзного вікна розміром $Nwin$ вимірів здійснюється визначення еталонних оцінок: статистичного середнього (математичне сподівання), дисперсії та середньоквадратичного відхилення похибки:

$$\begin{aligned}\hat{x}_j^{\text{etalon}} &= \frac{1}{Nwin} \sum_{j=1}^{Nwin} x_j, \hat{D}_j^{\text{etalon}} = \frac{1}{Nwin-1} \sum_{j=1}^{Nwin} (\hat{x}_j - x_j)^2, \\ \hat{\sigma}_j^{\text{etalon}} &= \sqrt{\hat{D}_j}, j = 1 \dots Nwin\end{aligned}\quad (3.6)$$

3. Формування наступного ковзного вікна розміром $Nwin$ вимірів.

4. Визначення поточних оцінок статистичного середнього (математичне сподівання), дисперсії та середньоквадратичного відхилення похибки для ковзних вікон:

$$\begin{aligned}\hat{x}_j &= \frac{1}{Nwin} \sum_{j=1}^{Nwin} x_j, \hat{D}_j = \frac{1}{Nwin-1} \sum_{j=1}^{Nwin} (\hat{x}_j - x_j)^2, \\ \hat{\sigma}_j &= \sqrt{\hat{D}_j}, j = 1 \dots Nwin\end{aligned}\quad (3.7)$$

5. Якщо умова виконується

$$\hat{\sigma}_j > Q \hat{\sigma}_j^{\text{etalon}}, (Q - \text{ваговий коефіцієнт налаштування}) - \quad (3.8)$$

то поточний j -й вимір, що був доданий до ковзного вікна приймається аномальним.

Аномальний j -й вимір замінюється у вибірці $X = \{x_i\}$ на оцінку математичного сподівання поточного ковзного вікна

$$x_i = \hat{x}_j \quad (3.9)$$

Якщо умова $\hat{\sigma}_j > Q \hat{\sigma}_j^{\text{etalon}}$ не виконується, то модифікації вибірки $X = \{x_i\}$ не здійснюється.

6. Повторення п. 3-5 в межах вхідної статистичної вибірки $i = 1 \dots n$.

7. Результатом є очищена вибірка $\hat{X} = \{\hat{x}_i\}$, $i = 1 \dots n$.

III. Алгоритм контролю змін динамічних властивостей вибірки – контролю швидкості з МНК.

Суть алгоритму полягає у виявленні аномальних вимірів шляхом визначення еталонних параметрів швидкості зміни динаміки тренду та їх порівняння з поточними характеристиками з використанням МНК. Еталонні значення швидкості розраховуються за вхідною статистичною вибіркою, а поточні – за вибірками ковзного вікна. Порівняння реалізується з урахуванням статистичних похибок оцінювання.

Етапи алгоритму контролю швидкості з МНК.

1. Вхідною є статистична вибірка $X = \{x_i\}$, $i = 1 \dots n$.
2. В межах статистичної вибірки за алгоритмом МНК із поліноміальною моделлю визначається еталонна швидкість для вибірки:

$$Speed^{etalon} = c_1, \quad (3.10)$$

з МНК поліному $T(x) = c_0 + c_1x + c_2x^2 + \dots$

3. Формування ковзного вікна розміром $Nwin$ вимірів.
4. Визначення поточних оцінок статистичного середнього (математичне сподівання), дисперсії та середньоквадратичного відхилення похибки для ковзних вікон:

$$\begin{aligned} \hat{x}_j &= \frac{1}{Nwin} \sum_{j=1}^{Nwin} x_j, \\ \hat{D}_j &= \frac{1}{Nwin-1} \sum_{j=1}^{Nwin} (\hat{x}_j - x_j)^2, \\ \hat{\sigma}_j &= \sqrt{\hat{D}_j}, j = 1 \dots Nwin \end{aligned} \quad (3.11)$$

5. Визначаються контрольовані параметри аномальності, що масштабуються до розміру вибірок отримання оцінок та з урахуванням похибок вимірів:

$$\begin{aligned} \text{Ind}_1 &= |Speed^{\text{etalon}} \cdot \sqrt{n}|, \\ \text{Ind}_2 &= |Q \cdot \hat{\sigma}_j \cdot Speed^{\text{etalon}} \cdot \sqrt{Nwin}|, \end{aligned} \quad (3.12)$$

(Q – ваговий коефіцієнт налаштування)

6. Якщо умова виконується

$$\text{Ind}_2 > \text{Ind}_1, \quad (3.13)$$

то поточний j -й вимір, що був доданий до ковзного вікна приймається аномальним.

Аномальний j -й вимір замінюється у вибірці $X = \{x_i\}$ на МНК оцінку

$$x_i = T(x = \hat{x}_j). \quad (3.14)$$

Якщо умова $\text{Ind}_2 > \text{Ind}_1$ не виконується, то модифікації вибірки $X = \{x_i\}$ не здійснюється.

6. Повторення п. 3-5 в межах вхідної статистичної вибірки $i = 1 \dots n$.

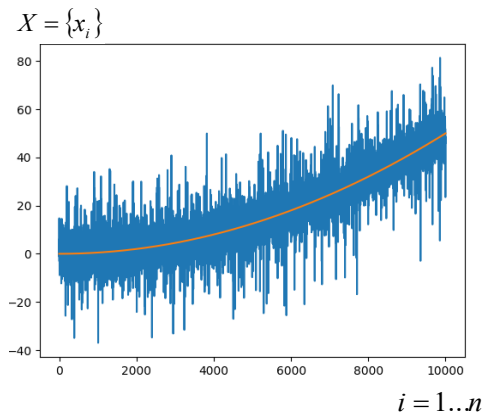
7. Результатом є очищена вибірка $\hat{X} = \{\hat{x}_i\}$, $i = 1 \dots n$.

Результати моделювання.

Моделюванню підлягала статистична вибірка з $n = 10000$ вимірів. У вибірці присутній квадратичний тренд, похибка виміру із нормальним законом розподілу. Нормальний закон має нульове математичне очікування та середньоквадратичне відхилення $\sigma_x = 5,0$ од. Аномальні виміри рівномірно розподілені у межах вибірки і для наведених результатів розрахунку складають 10%. Модель тренду та стохастичних складових – адитивна. Розрахунки проведені з реалізацією алгоритму у формі прикладного скрипта мовою python [34, 37] з використанням можливостей бібліотек numpy, matplotlib.

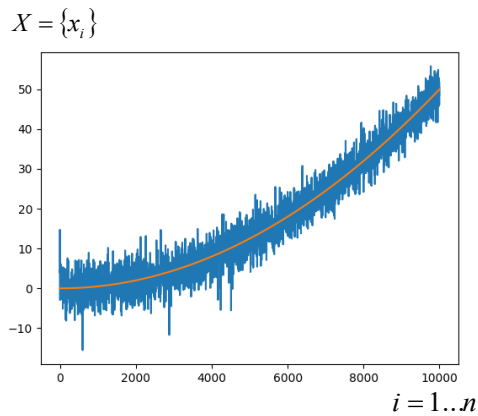
Результати досліджень подані на рис.3.4. На рис.3.4а представлено графік вхідної статистичної вибірки $X = \{x_i\}$, $i = 1 \dots n$ з нормальним шумом і аномаліями, а також лінія тренду, отримана за алгоритмом методу найменших квадратів (МНК) [41]. Середньоквадратичне відхилення вхідної статистичної вибірки в присутності аномалій складає $\sigma_X = 6.64$ од.

На рис.3.4б подано графік очищеної від аномалій за алгоритмом `sliding_wind` статистичної вибірки, на 3.4в – за алгоритмом `medium`, на рис.3.5г – за алгоритмом контролю швидкості з МНК.



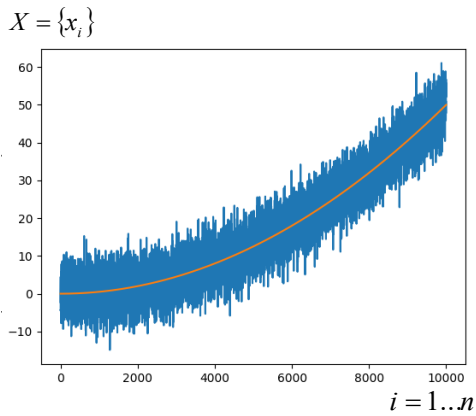
Вхідна вибірка $\sigma_X = 6,64$ од.

а).



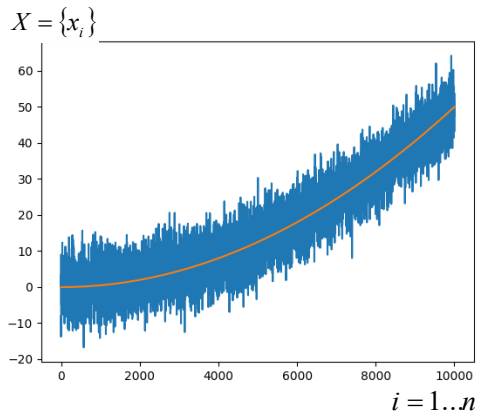
Очищена вибірка $\sigma_X = 2,95$ од.

б).



Очищена вибірка $\sigma_X = 4,45$ од.

в).



Очищена вибірка $\sigma_X = 4,71$ од.

г).

Рисунок 3.4. Результати досліджень запропонованих алгоритмів

Порівняння графіків рис.1а та рис.1б,в,г, а також зменшення середньоквадратичного відхилення похибки очищеної статистичної вибірки свідчить про дієвість запропонованих алгоритмів.

Узагальнені статистичні характеристики із зменшення похибки очищених даних за трьома запропонованими алгоритмами можна подати у вигляді табл.3.1.

Таблиця 3.1

Узагальнені статистичні характеристики розроблених алгоритмів

Вхідна вибірка з аномаліями	sliding_wind	medium	MNK
$\sigma_x = 6,64$ од.	$\sigma_x = 2,95$ од.	$\sigma_x = 4,45$ од.	$\sigma_x = 4,71$ од.

Аналіз даних таблиці показує, що за показником точності (мінімальним значенням середньоквадратичного відхилення) є алгоритм sliding_wind. Він же є найбільш продуктивним та не має параметрів для налаштування. Однак його використання для динамічних процесів із значною нелінійністю може призводити до зміщеності результуючих оцінок. Алгоритми medium та контролю швидкості з MNK мають протилежні до sliding_wind властивості. Вибір для використання конкретного алгоритму визначається властивостями вибірки та аномальними вимірами – об’єм, похибки, характер трендової моделі.

Порівняння розроблених підходів із більш складними аналогами, що співставні за складністю обчислень з MNK на вибірці, що розглядалась – в 10000 вимірів не дали прийнятних показників оперативності розрахунків.

3.1.3. Структурна оптимізація моделі даних

Оптимізація структури моделі має на меті мінімізацію динамічної і випадкової помилки результатів оброблення часових рядів. Предметом

оптимізації є математична модель, що на ділянці спостереження з високим ступенем точності інтерпретує домінантний тренд даних та володіє високими прогностичними властивостями поза межами інтервалу спостереження Time Series. Зазначений процес реалізується з контролем класичних індикаторів, які включають наступний перелік [28].

Глобальне лінійне відхилення оцінки – (mean absolute error (MAE)) – є відносною, оцінюється методом порівняння з мінімізацією:

$$\Delta = \frac{1}{n-1} \sum_{i=1}^n |y_i - \hat{y}_i|. \quad (3.15)$$

Достовірності апроксимації R^2 (coefficient of determination) – змінюється в межах $0 \dots 1$ та має бути максимальним (див. також (3.1)):

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (3.16)$$

де n – об'єм вибірки $\bar{y}_i = \frac{1}{n} \sum_{i=1}^n y_i$, y_i – виміряне значення, $\hat{y}_i$ – оцінка виміряного значення.

Однак їх визначення передбачає послідовний процес тестування кожної структури моделі на апостеріорних даних, що потребує значних обчислювальних витрат на Big Data масивах. Така процедура є сталою і не дозволяє оперативно адаптувати структуру моделі під зміну властивостей вхідних даних.

Тому, в роботі запропоновано реалізацію процесу оптимізації моделі здійснювати на підставі аналізу значень похідних вищих порядків, з урахуванням випадкових похибок Time Series.

Суть запропонованого підходу полягає у наступному. Нехай часовий ряд $F_{j,area,tonality}$ позначає вибірка реалізацій

$$y_1, y_2, y_3, \dots, y_m. \quad (3.17)$$

Необхідно за вибіркою (3.17) визначити порядок експериментальної кривої k , що дозволить оптимальним чином сформувати поліноміальну модель даних

$$F(t) = C_0 + C_1 t + C_2 t^2 + \dots + C_n t^n = \sum_{i=0}^{n_{opt}} C_i t^i, \quad (3.18)$$

де C_i – коефіцієнти апроксимуючого полінома, $n_{opt} = k$ – оптимальний (дорівнює порядку експериментальної кривої) ступінь полінома.

Основна ідея підходу полягає у використанні властивості, що ускладнення характеру досліджуваного процесу приводить до появи додаткових ненульових прискорень (похідних вищих порядків) [76]. Ускладнення в моделі вигляду (3.18) враховується шляхом додавання додаткових коефіцієнтів C_i , суть яких – похідні i -го порядку (C_0 – початкове значення координати, C_1 – швидкість зміни координати, C_2 – прискорення і т. ін.).

Можна стверджувати, що порядок поліноміальної моделі (3.18) знаходиться в прямій залежності від порядку останньої ненульової (в умовах відсутності похибок вимірів) похідної експериментальної вибірки (3.17). Таким чином, базовою ознакою для визначення порядку моделі в умовах відсутності похибок виміру координат є порядок останньої її ненульової похідної.

У загальному вигляді обчислення похідної за дискретною послідовністю здійснюється згідно з виразом:

$$y_j^{(p)} = \frac{y_{j+1}^{(p-1)} - y_j^{(p-1)}}{\Delta t}, j = \overline{1 \dots m}, p = \overline{1 \dots n}, \quad (3.19)$$

де $y_j^{(p)}, y_{j+1}^{(p-1)}, y_j^{(p-1)}$ – p -а і $(p-1)$ -а похідні в моменти часу j та $j+1$ відповідно; Δt – інтервал отримання вимірів вибірки (3.17).

Відповідно до базового виразу (3.19) кінцеві рівняння для пошуку похідних за дискретними даними до п'ятого порядку включно набудуть вигляду:

$$\begin{cases} y_j^{(1)} = \frac{y_{j+1} - y_j}{\Delta t}, \\ y_j^{(2)} = \frac{y_{j+2} - 2y_{j+1} + y_j}{\Delta t^2}, \\ y_j^{(3)} = \frac{y_{j+3} - 3y_{j+2} + 3y_{j+1} - y_j}{\Delta t^3}, \\ y_j^{(4)} = \frac{y_{j+4} - 4y_{j+3} + 6y_{j+2} - 4y_{j+1} + y_j}{\Delta t^4}, \\ y_j^{(5)} = \frac{y_{j+5} - 5y_{j+4} + 10y_{j+3} - 10y_{j+2} + 5y_{j+1} - y_j}{\Delta t^5}. \end{cases} \quad (3.20)$$

Раніше було зазначено, що послідовність (3.17) отримана експериментальним шляхом, що зумовлює наявність у координатах y_j , ($j = \overline{1 \dots m}$) випадкової похибки вимірів. Тоді модель виміру координати можна подати у вигляді

$$y_j = y_{jd} + \xi, \quad (3.21)$$

де y_{jd} – дійсне значення координати, яке, як правило, невідоме; ξ – випадкова похибка виміру, що підпорядкована нормальному закону розподілу з відомими числовими характеристиками: математичним сподіванням $m_\xi = 0$ і дисперсією σ_ξ^2 (квадрат середньоквадратичного відхилення (СКВ) похибки), яка характеризує потенційні точності вимірювача. Виходячи із зазначеного, модель p -ї похідної за координатою може бути записана в такий спосіб:

$$y_j^{(p)} = y_{jd}^{(p)} + \xi^{(p)}, \quad (3.22)$$

де $y_{jd}^{(p)}$ – дійсне значення p -ї похідної за координатою; $\xi^{(p)}$ – випадкова похибка розрахунку p -ї похідної з нормальним законом розподілу, нульовим середнім ($m_{\xi^{(p)}} = 0$) і дисперсією $\sigma_{\xi^{(p)}}^2$, яка характеризує похибку перетворень (3.20).

У випадку рівності нулю дійсного значення p -ї похідної (що раніше було прийнято за базову ознаку порядку вимірної послідовності) модель (3.22) трансформується до вигляду:

$$y_j^{(p)} = \xi^{(p)}. \quad (3.23)$$

Надалі слід визначити похибку перетворень (3.20) двома способами: теоретично; експериментально.

Теоретичне визначення дисперсії базується на теоремі про дисперсію суми (різниці) для рівнодискретних та рівноточних випадкових величин:

$$\sigma_{y^{(p)}_T}^2 = \frac{2\sigma_y^2}{\Delta t^{2(p)}}, \quad (3.24)$$

де σ_y^2 – дисперсія похибок виміру координат вибірки (3.17), яка, як правило, відома і характеризує потенційні точності вимірника.

Експериментальне визначення дисперсії базується на обробці серії реалізацій. Для розглянутої задачі вважатимемо, що вирази (3.19) і відповідно формули (3.20) послідовно застосовуються до кожного компонента вибірки (3.17). У результаті цього отримаємо підвибірки:

$$y_j^{(1)}, (j = \overline{1 \dots m-1}), y_j^{(2)}, (j = \overline{1 \dots m-2}), \dots, \\ y_j^{(p)}, (j = \overline{1 \dots m-p}). \quad (3.25)$$

Тоді експериментальне визначення дисперсії перетворень (3.19) реалізується згідно з виразом:

$$\sigma_{y^{(p)}_E}^2 = \frac{1}{(m-p)-1} \sum_{j=1}^{m-p} (y_j^{(p)} - m_{y^{(p)}}), \quad (3.26)$$

де $y_j^{(p)}$ – компоненти вибірок p -ї похідної (див. (3.25)); $m_{y^{(p)}}$ – математичне сподівання для p -ї похідної вибірок (3.25), визначене відповідно до виразу

$$m_{y^{(p)}} = \frac{1}{m-p} \sum_{j=1}^{m-p} y_j^{(p)}.$$

Вираз (3.25) дасть правильний результат тільки тоді, коли аналізована вибірка з набору (3.26) характеризуватиме стаціонарний випадковий процес. В інших випадках спостерігатиметься зміщеність оцінок, що розраховуються. Відстежити наявність зміщеності в оцінках експериментально розрахованої дисперсії перетворення (3.19) можливо реалізувати шляхом порівняння її значення з теоретично розрахованою дисперсією. Таким чином, порядок експериментальної моделі можна визначити за порядком p -ї похідної, для якої абсолютна різниця між теоретичним $\sigma_{y^{(p)}_T}$ і експериментальним $\sigma_{y^{(p)}_E}$ значенням СКВ похибки визначення p -ї похідної $\Delta^{(p)}$

$$\Delta^{(p)} = \left| \sigma_{y^{(p)}_E} - \sigma_{y^{(p)}_T} \right| - \quad (3.27)$$

буде мінімальною.

Відповідно до викладеного оптимальне значення порядку поліноміальної моделі визначається умовою:

$$\begin{cases} n_{opt} = k = p, & \text{при} \\ \Delta^{(p)} = \left| \sigma_{y^{(p)}_E}^2 - \sigma_{y^{(p)}_T}^2 \right| = \min \text{ із } \Delta^{(p)}, & p = \overline{1 \dots m} \end{cases} \quad (3.28)$$

Алгоритм визначення порядку поліноміальної моделі включає етапи:

1. За компонентами вибірки (3.17) визначити підвибірки (3.25) згідно з виразами (3.20).
2. Визначити теоретичне та експериментальне значення дисперсії розрахунку p -х похідних, використовуючи вирази (3.24) і (3.26) відповідно.
3. Розрахувати різниці (3.27) для кожної пари СКВ похибок визначення p -х похідних.
4. Прийняти рішення про оптимальний порядок поліноміальної моделі відповідно до умови (3.28).

Слід зазначити, що сформований у такий спосіб алгоритм досить простий у реалізації і вимагає значно менше обчислювальних витрат у порівнянні з відомими підходами [76], коли для визначення порядку експериментальної кривої слід одержати декілька моделей і проаналізувати низку критеріїв для вироблення остаточного рішення.

Верифікація розробленого алгоритму проводилась методом імітаційного моделювання. Під час дослідження покладалося, що обробці підлягала вибірка в 21 вимір координати y з темпом оновлення інформації $\Delta t = 2$ сек і СКВ похибки вимірювання $\sigma_y = 0.1y$.од. Аналіз розробленого алгоритму здійснювався для кривих до п'ятого порядку включно. Результати досліджень подані в табл.3.2.

Таблиця 3.2

Результати верифікації алгоритму визначення
порядку поліноміальної моделі

Параметр	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 5$	n_{opt}
$k = 1$	0.024	0.027	0.033	0.037	0.039	1
$k = 2$	23.62	0.027	0.033	0.037	0.039	2
$k = 3$	$16 \cdot 10^2$	67.47	0.033	0.037	0.039	3
$k = 4$	$86 \cdot 10^3$	$60 \cdot 10^2$	$25 \cdot 10$	0.037	0.039	4
$k = 5$	$42 \cdot 10^5$	$40 \cdot 10^4$	$28 \cdot 10^3$	$12 \cdot 10^2$	0.039	5

У табл.3.2 подані значення параметра $\Delta^{(p)}$ для похідних до п'ятої включно ($p = \overline{1 \dots 5}$) при зміні порядку вимірної кривої ($k = \overline{1 \dots 5}$). Останній стовпець табл.3.2 характеризує прийняте рішення про оптимальне значення порядку згладжуючого полінома n_{opt} .

Аналіз поданих у табл.3.2 даних дозволяє зробити висновок, що умова (3.28) є дієвою, а розроблений алгоритм реалізує формування оптимального (рівного порядку експериментальної кривої) значення поліноміальної моделі за наявності похибок вимірювання. Отже запропонований підхід дозволяє реалізувати оперативне (за один прохід обчислень по вибірці вимірів) визначення порядку поліноміальної моделі за експериментальними даними.

3.2. Статистичне навчання з нелінійними моделями

В практиці обробки часових рядів наразі відомо три класу методів: статистичне навчання / згладжування; апроксимація; глибинне навчання.

Статистичне навчання, наприклад з використанням МНК спрямовано на отримання параметрів моделі заздалегідь сформованої структури (навчання моделі) як результат обробки ретроспективних даних.

Модель, фактично забезпечує виділення тренду досліджуваного процесу з мінімізацією випадкової складової похибок.

Апроксимація / регресійний аналіз, наприклад алгоритмом ARIMA та його модифікації, експоненціальне згладжування тощо. Передбачає відновлення дискретного набору Time Series аналоговою моделлю із відвідуванням кожної точки ряду.

Глибинне навчання – реалізується із використанням нейронних мереж різного класу і має на меті відтворення досліджуваного процесу по отриманих ретроспективним даним.

При природний характер Time Series в імплементації частоти $F_{j,area,tonality}$ та результати аналізу реальних даних свідчить про їх випадковий характер із динамічним трендом. Тому за базовий алгоритм обробки метод найменших квадратів з класу методів статистичного навчання.

Разом із тим відомий доволі успішний досвід обробки часових рядів, навіть із спотворюючими випадковими факторами з використанням методів апроксимації та глибинного навчання [38-54]. Тому, зважаючи на ймовірнісний характер отримуваних оцінок в роботі використовується три класи методів для задач оцінювання і прогнозування розвитку активності контенту за індикаторами частоти $F_{j,area,tonality}$. Таке поєднання передбачається використати для підвищення інтервальної точності – зменшення величини довірчого інтервалу прогнозованих значень. Це планується досягти на підставі сумісного аналізу отримуваних результатів обробки від методів трьох раніше зазначених класів.

Відомо, що для рівноточної, рівнодискретної вибірки вимірів з нормальним законом розподілу похибок оптимальна за критерієм мінімуму дисперсії оцінка досліджуваного процесу може бути отримана з використанням класичного методу найменших квадратів (МНК) [41].

Вхідною інформацією для МНК є типовий Time Series (табл.3.2) – дискретні значення досліджуваного процесу за присутності випадкових похибок – y , у дискретні моменти часу – x .

Таблиця 3.3.

Вхідні дані для МНК

Значення аргументу і функції	Номер вимірювального вузла: значення функції відповідно до значення аргументу			
	1	2		n
x	x_1	x_2		x_n
y	y_1	y_2		y_{n1}

Відомо, що досліджуваний процес можливо описати функціональною залежністю $y(x)$ – моделлю даних. Визначення типу та структури моделі розглянуто в п.3.1.2 та буде уточнено нижче.

Необхідно встановити параметри функціональної залежності $y(x)$ – конкретизувати її вигляд, знизивши похибки вимірів: «навчити модель за параметрами». Тобто, визначити параметри моделі s зі зниженням впливу випадкової складової вимірів y :

$$y(x) = c_0 + c_1x + c_2x^2 + c_3x^3 + + c_nx^n ... \tag{3.29}$$

Нехай досліджуваний процес описується поліномом:

$$T(x) = \sum_{i=0}^m c_i \phi_i(x_i), (m \geq 2n), \tag{3.30}$$

де c_i – коефіцієнти апроксимуючої функції; $\phi_i(x_i)$ – базисні функції апроксимуючого полінома, m, n – порядок апроксимуючого полінома і

кількість значень у вибірці відповідно. Тоді, задача статистичного навчання полягає у визначенні коефіцієнтів апроксимуючого полінома.

Критерій оптимальності МНК: сума квадратів різниць між дискретними значеннями і оціненими параметрами повинна бути мінімальною:

$$\delta(x, c) = \sum_{i=1}^n q_i (y_i - T(x_i))^2 \Rightarrow \min, \quad (3.31)$$

де $q_i = 1/\sigma_{y_i}^2$ – вага зворотно пропорційна дисперсії i -го виміру – $\sigma_{y_i}^2$; y_i – i -й вимір (значення) контрольованого параметру; $T(x_i)$ – оцінка виміру з номером x_i ; c – параметр, що включає невідомі коефіцієнти c_i поліному – підлягають «навчанню».

Для визначення мінімуму функції (3.31) необхідно знайти m її часткових похідних за параметрами полінома дорівняти їх нулю і розв’язати утворену систему рівнянь:

$$\frac{\partial \delta(x, c)}{\partial c_0} = 0; \quad \frac{\partial \delta(x, c)}{\partial c_1} = 0; \quad \dots \quad \frac{\partial \delta(x, c)}{\partial c_n} = 0. \quad (3.32)$$

Розв’язкам системи рівнянь (3.32) є матрична форма МНК:

$$\begin{aligned} \hat{\bar{C}} &= (F^T F)^{-1} F^T \bar{Y}, \\ \hat{\bar{T}} &= F \hat{\bar{C}}. \end{aligned} \quad (3.33)$$

$\hat{\bar{C}}[m \times 1]$ – вектор коефіцієнтів «навченого» полінома – результатів МНК; $F[n \times m]$ – матриця значень базисних функцій полінома; $\bar{Y}[n \times 1]$ – вектор значень / вимірюваних параметрів; $\hat{\bar{T}}[n \times 1]$ – вектор оцінок вимірюваних значень – результатів МНК.

Специфікою і недоліком класичного МНК (3.33) є вимога отримання системи рівнянь (3.32) в класі лінійних. Це забезпечується класом моделей з лінійними по параметрах відношеннями аргумент / функція. Тобто, для визначення параметрів моделі c – мають бути застосовано лише лінійні операції. Це виконується в обмеженому класі функцій типу (3.30). За таких умов відтворення нелінійності досліджуваного процесу, що відображає Time Series формату табл.3.3 реалізується шляхом додавання високоступеневих компонент до моделі (3.30). Цим досягається методична / динамічна точність моделі. Однак такий шлях має низку недоліків / обмежень. По-перше, збільшення ступеня полінома (3.30) призводить до збільшення кількості розрахункових операцій МНК (див. алгоритм (3.33), зокрема матриця $F[n \times m]$). По-друге отримати адекватну модель можливо на ділянці спостереження, що не гарантує його високу адекватність на ділянці прогнозування (не гарантує високі прогностичні властивості моделі). По-третє, для певних процесів, наприклад асимптотичні, періодичні / сезонні, особливо із зміною параметрів періодичності підбір структури полінома (3.30) та ще і з високими прогностичними властивостями є нездійсненою задачею. Задача згладжування із класичним МНК значно ускладнюється з позицій формування поліному (3.30) для процесів із комбінованою поведінкою, наприклад асимптотичний, поліноміальний та періодичний характер в межах однієї вибірки.

В той же час, реальні часові ряди $F_{j,area,tonality}$ мають саме комбінаторний характер з сегментарним представленням лінійних, квадратичних, асимптотичних та періодичних ділянок.

Враховуючи вищезазначене, в роботі використано та удосконалено підхід до побудови нелінійних за параметрами моделей із статистичним навчанням їх параметрів [54, 72, 73, 77].

Початковою інформацією для реалізації алгоритмів статистичної навчання є дискретна вибірка – часовий ряд Time Series формату $F_{j,area,tonality}$, що для зручності позначено у (див. (3.17) табл. 3.2)

Вважається відомою нелінійна за параметрами теоретична модель $f(t, c)$, яка може бути заданою у вигляді певної функціональної залежності. Вигляд нелінійної моделі формується виходячи із властивостей Time Series, виявлених на етапі первинного аналізу.

Необхідно побудувати, тобто визначити параметри нелінійної моделі досліджуваного процесу $f(t, c)$ за статистичною вибіркою (3.17), табл.3.2.

Надалі за вибіркою вхідних параметрів y , наприклад за МНК формується експериментальна модель

$$\hat{z}(t) = \Phi(y), \quad (3.34)$$

де Φ позначає операції визначення параметрів поліноміальної моделі $\hat{z}(t)$, формату (3.30) за вибіркою експериментальних даних y відповідності до обраного алгоритму статистичного навчання.

Отже отримано дві моделі: нелінійна $f(t, c)$ – визначеної структури, але з невідомими параметрами; $\hat{z}(t)$ – узгоджена / відома за параметрами з вибіркою y .

Головна ідея статистичного навчання з нелінійними за параметрами моделями полягає у визначенні параметрів $f(t, c)$ з відомих параметрів поліному $\hat{z}(t)$. Для реалізації сформованої ідеї обрано метод диференціальних перетворень (ДП), що забезпечує отримання аналітичних рішень широкого кола нелінійних задач практики [54, 72, 73, 77].

Диференціальні перетворення в загальному випадку – це функціональні перетворення вигляду:

$$Z(k) = P\{z(t)\}_{t^*} = \frac{H^k}{k!} \left[\frac{d^k z(t)}{dt^k} \right]_{t^*}, \quad (3.35)$$

$$z(t) = f(t, c), \quad (3.36)$$

де t^* – значення аргументу, при якому проводиться перетворення; $Z(k)$ – дискретна функція аргументу $k = 0, 1, 2, \dots$; H – відрізок аргументу, на якому розглядається функція $z(t)$; $f(t, c)$ – апроксимуюча (у т.ч. нелінійна) функція; c – множина вільних коефіцієнтів c_i .

Вираз (3.35) забезпечує отримання за оригіналом $z(t)$ його диференціальне зображення $Z(k)$ – пряме ДП. За оберненим ДП (3.36) відбувається відновлення оригіналу $z(t)$. Диференціальне зображення $Z(k)$ називається диференціальним спектром (ДС) або Р-спектром, а значення $Z(k)$ при конкретних значеннях аргументу – дискретами ДС (Р-дискретами). У найпростішому випадку функція $f(t, c)$ має вигляд багаточлена, а відновлення оригіналу зводиться до підсумовування дискрет Р-спектра у вигляді відрізка ряду Тейлора (основні, або диференціально–тейлорівські (ДТ) перетворення).

Для визначення за параметрами поліноміальної моделі $\hat{z}(t)$ шуканих параметрів нелінійної моделі $f(t, c)$ застосуємо баланс диференціальних спектрів (БДС) – прирівнювання дискрет ДС з однаковими аргументу k . Таким чином утворюється система рівнянь вигляду

$$\begin{cases} \hat{Z}(0) - F(0, c_1) = 0, \\ \hat{Z}(1) - F(1, c_2) = 0, \\ \hat{Z}(2) - F(2, c_3) = 0, \\ \dots \dots \dots \\ \hat{Z}(k) - F(k, c_m) = 0. \end{cases} \quad (3.37)$$

Кількість рівнянь в системі (3.37) має дорівнювати кількості невідомих параметрів c нелінійної моделі $f(t, c)$. Тож структура

поліноміальної моделі $\hat{z}(t)$ має забезпечувати формування визначеної системи рівнянь (3.37).

Первинний аналіз даних $F_{j,area,tonality}$ показав їх комбінаторний характер з представленням лінійних, квадратичних, асимптотичних та періодичних ділянок. Для збереження високих прогностичних властивостей моделей, що реалізують тренд даних $F_{j,area,tonality}$ надалі реалізовано статистичне навчання параметрів комбінаторних нелінійних моделей трьох класів: експоненціальна модель в комбінаториці з лінійним поліномом (3.38); тригонометрична двокомпонентна модель (3.39).

$$f(t, a) = a_0 + a_1 t + a_2 e^{a_3 t},$$

$$c = \{a_0, a_1, a_2, a_3\} - \text{параметри моделі}$$
(3.38)

$$f(t, c) = a_0 \cos \omega t + b_0 \sin \omega t,$$

$$c = \{a_0, b_0, \omega\} - \text{параметри моделі}$$
(3.39)

Експоненціальна модель (3.38) містить чотири шуканих параметра $c = \{a_0, a_1, a_2, a_3\}$. Тож для формування системи рівнянь вигляду (3.37) в алгоритмі МНК (3.33) необхідно реалізувати поліном (3.30) як мінімум третього ступеня. Додавання інших компонент до полінома, визначається умовою (3.28).

Застосовуючи прямі ДП (3.35) до нелінійної моделі (3.38) та поліноміальної моделі третього ступеня вигляду (3.30), конкретизованого до (3.41) відповідно до БДС (див. приклад системи (3.37)) маємо систему рівнянь для пошуку параметрів нелінійної моделі $f(t, a)$ – (3.41).

$$z(t, c) = c_0 + c_1 t + c_2 t^2 + c_3 t^3$$
(3.40)

$$\begin{cases} c_0 - (a_0 + a_2) = 0, \\ c_0 - (a_1 + a_2 + a_3) = 0, \\ c_0 - \frac{a_2 a_3^2}{2} = 0, \\ c_0 - \frac{a_2 a_3^3}{6} = 0. \end{cases} \quad (3.41)$$

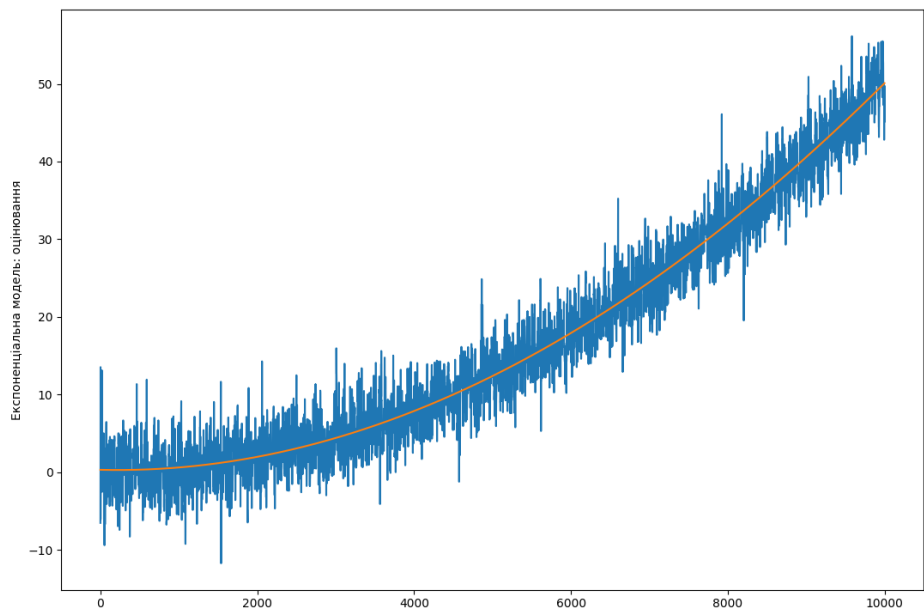
Розв'язкам системи рівнянь (3.41) є шукані параметри нелінійної моделі (3.38):

$$\begin{aligned} c &= \{a_0 = c_0 - a_2, \\ a_1 &= c_1 - (a_2 * a_3), \\ a_2 &= \frac{2c_2}{a_3^2}, \\ a_3 &= \frac{3c_3}{c_2}. \end{aligned} \quad (3.42)$$

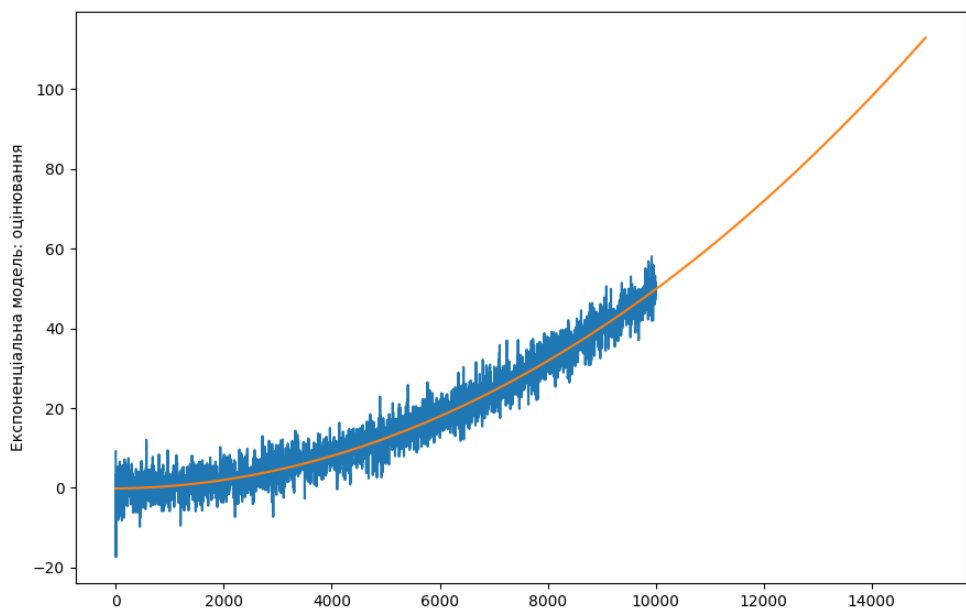
Таким чином маємо експоненціальну модель (3.38), параметри якої визначені за навченими згідно норми МНК параметрами полінома (3.40) відповідно до виразів (3.42).

Верифікація запропонованих рішень проводилась на модельних даних з 10000 вимірів із випадковим шумом: нульовим середнім та СКВ в 5 у.о. Вибірка забруднена 10% аномальними вимірами типу грубі значення. Закон зміни модельних даних – асимптотична крива, що зростає. Результати оцінювання та прогнозування подані на графіках рис.3.5 а та рис.3.5 б відповідно.

Значення показників ефективності нелінійної експоненціальної та поліноміальної моделей наведені в табл.3.4. 3.5.



а)



б)

Рисунок 3.7. Графіки оцінювання та прогнозування часових рядів з нелінійною експоненціальною моделлю

Застосовувались також чисельні методи розв'язку задачі пошуку параметрів експоненціальної моделі за вибіркою вимірів. Це підходи потребують апріорних даних про первинні наближення шуканих параметрів, що не є можливим для реальних процесів обробки часових

рядів типу $F_{j,area,tonality}$. Тож результати, придатні для застосування чисельних методів не були отримані. Хоча в якості первинних наближень можливо розглядати параметри моделі, отримані запропонованим підходом. Даний напрям в сумісному використанні нелінійних моделей та чисельних методів буде застосовано в подальшому.

Аналіз даних табл.3.4, 3.5 показав стабільність поліноміальної та експоненціальної моделі даних на ділянці оцінювання. Однак, за динамічною точністю нелінійна за параметрами модель має незначні покращення, що важливо на ділянці прогнозування (див дані табл.3.5).

Таблиця 3.4

Показники ефективності моделей на ділянці оцінювання

Показник	Поліноміальна модель	Експоненціальна модель
Кількість елементів вибірки	10000	10000
Математичне сподівання ВВ	7.904787935331123e-14	7.904787935331115e-14
Дисперсія ВВ	3.6702065914811867e-28	3.6702065914811904e-28
СКВ ВВ	1.9157783252456234e-14	1.9157783252456924e-14
Динамічна похибка моделі	4.6709103193845	4.7526642753216795
Коефіцієнт детермінації (ймовірність апроксимації)	0.9633860728043558	0.9616629184356816
Час на розрахунки, с	0.008940935134887695	0.005113363265991211

Таблиця 3.5

Показники ефективності моделей на ділянці прогнозування

Показник	Поліноміальна модель	Експоненціальна модель
Кількість елементів вибірки	15000	15000
Математичне сподівання ВВ	8.029851072244298e-12	-1.8829382497642655e-13
Дисперсія ВВ	7.972209846948842 e-20	6.615311136464127e-26
СКВ ВВ	8.928723227286667 e-10	2.572024715368056e-13
Довірчий інтервал прогнозованих значень	0.4464361613643333e-05	1.2860123576840278e-09
Час на розрахунки, с	0.01014566421508789	0.01014566421508789

Трансцендентна двокомпонентна модель (3.39) дозволяє утворити систему рівнянь типу (3.37) конкретного вигляду, базуючись на поліноміальній моделі (3.30):

$$\begin{cases} c_0 - a_0 = 0, \\ c_1 - b_0 \omega = 0 \\ c_2 + a_0 \omega^2 = 0. \end{cases} \tag{3.43}$$

Розв’язок системи (3.43) дає параметри нелінійної моделі (3.39):

$$c = \{a_0 = c_0, \omega = \sqrt{c_2/c_0}, b_0 = c_1/\sqrt{c_2/c_0}\}. \tag{3.44}$$

Отже, сформовано трансцендентну двокомпонентну модель (3.39) параметри якої визначені з МНК поліноміальної моделі за БДС.

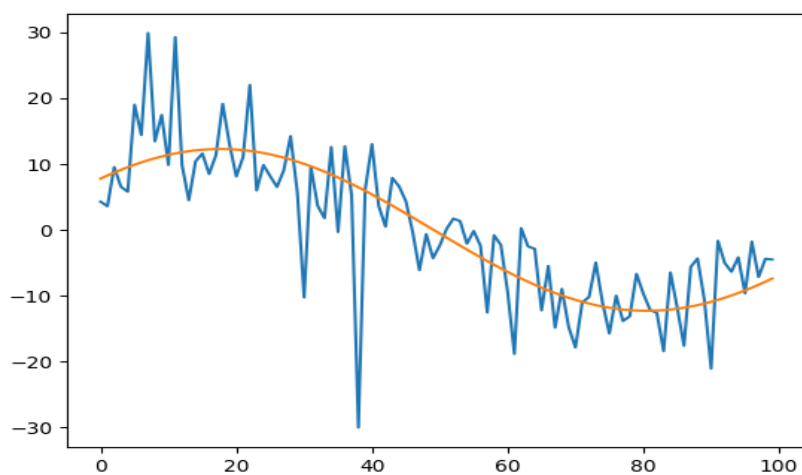
Дослідження моделі (3.39) з параметрами (3.44) здійснено на синтезованих даних: об'єм вибірки 100 значень; характер тренду – періодичний; значення time series забруднені стохастичним шумом з нормальним законом розподілу, середньоквадратичним відхиленням 5 у.о. та 10% аномальних вимірів. Результати досліджень подано на рис.3.8.

На рис.3.8. вісь ігрек позначає вимірний параметр, вісь ікс – аргумент зміни параметра. Рис.3.8.а. – графік вхідної вибірки (синій колір) з ідеальним періодичним трендом (помаранчевий колір). Результати застосування статистично навченого полінома подані на рис.3.8.б. – синій колір крива вхідних даних, помаранчевий – виявлений тренд. Результати прогнозування за моделлю, отриманою відповідно до двокомпонентної трансцендентної моделі подані на рис.3.8,в.

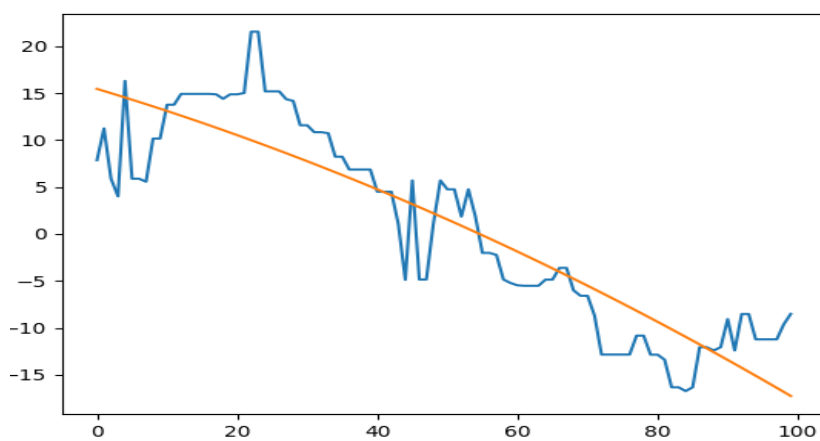
Показники якості нелінійної за параметрами моделі рис.3.8.б складають: абсолютне лінійне відхилення 5,78 у.о.; коефіцієнт детермінації 0,91 у.о. Показники якості поліноміальної моделі 2 порядку (рис.3.8.в), для заданих умов і обмежень складають: абсолютне лінійне відхилення 7,65 у.о, та коефіцієнт детермінації 0,76 у.о. Досягнення показників якості нелінійної моделі, забезпечуються поліномом не менше 6 ступеня.

Отже, наведені результати досліджень доводять ефективність застосування нелінійних за параметрами моделей.

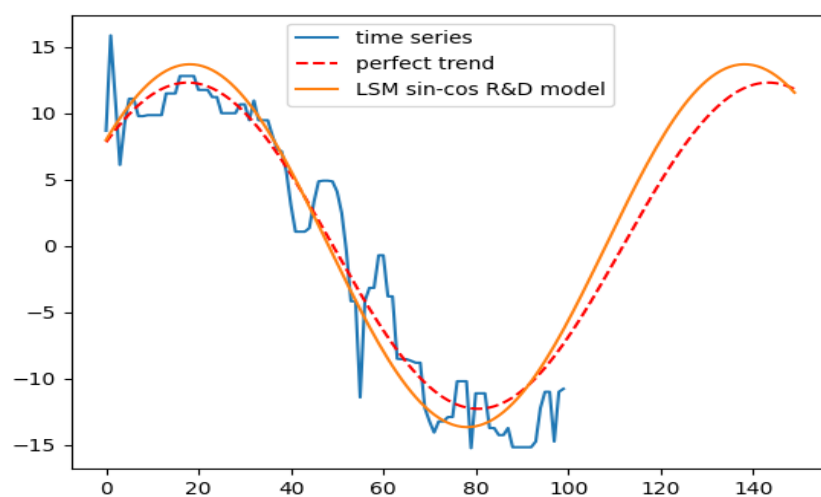
Моделі (3.38), (3.39) отримані за БДС, що обмежує структур і властивості нелінійних за параметрами моделей поліномом обмеженого ступеня. Такі ж обмеження накладаються на структуру і нелінійної за параметрами моделі. Подолати таке обмеження пропонується введенням двоетапної процедури побудови нелінійної за параметрами моделі.



a)



б)



в)

Рисунок 3.8. Графіки оцінювання та прогнозування часових рядів з трансцендентною двокомпонентною моделлю

Суть її в доповненні БДС чисельними методами пошуку параметрів нелінійних моделей де в якості початкових наближень використати отримані за БДС результати. Зазначене реалізовано в пошуку параметрів комбінаторної моделі: трансцендентна із лінійним поліномом.

Подані рішення демонструють порядок побудови різних типів нелінійних за параметрами моделей в доповненні поліноміальних. Вибір конкретної структури моделі для визначення її параметрів з методом статистичного навчання, що має високі показники достовірності на ділянці спостереження і володіє високими прогностичними властивостями відбувається за алгоритмом, наведеним на структурній схемі рис.3.9.



Рисунок 3.9. Алгоритм вибору та оцінювання параметрів моделі даних для методів статистичного навчання

1. Вхідною інформацією є Time Series $F_{j,area,tonality}$ (2.8) та / або $\hat{F}_{j,area,tonality}$ (2.10);
2. За вхідною інформацією (2.8) та / або (2.10) визначається оптимальний ступень полінома (3.30) для МНК (3.33) за аналізом значень похідних вищих порядків (умова (3.28));
3. Здійснюється навчання (визначення параметрів) полінома (3.30) за алгоритмом МНК (3.33) з обробкою даних п.1 алгоритму;
4. Реалізується розрахунок показників якості (KPI) навченої за МНК моделі даних, вир. (3.15), (3.16) та показника Херста (3.4)
5. Аналіз KPI навченої за МНК поліноміальної моделі (п.3,4);
6. Залежно від результатів аналізу п.5 обирається модель для прогнозування активності контенту в глобальних інформаційних мережах з переліку: поліноміальна; експоненційна; трансцендентна, або чисельний розв'язок задачі визначення параметрів нелінійної моделі. Для останньої в якості значень первинних наближень шуканих параметрів обирається результати побудови нелінійної за параметрами моделі.

3.3. Обробки часових рядів з використанням оптимізаційних нейронних мереж

В сучасній практиці обробки Time Series активно розвивається напрям застосування штучних нейронних мереж. Парадигма глибинного навчання в обробці часових рядів полягає у навчанні нейронної мережі на вибірці поточної реалізації (в межах інтервалу спостереження) та її використання для прогнозування поведінки часового ряду на ділянці спостереження та в перспективі (екстраполяція) / ретроспективі (інтерполяція). Результатом є згенерований часовий ряд з максимально схожими властивостями навидь до рівня окремої реалізації (див. рис.3.10).

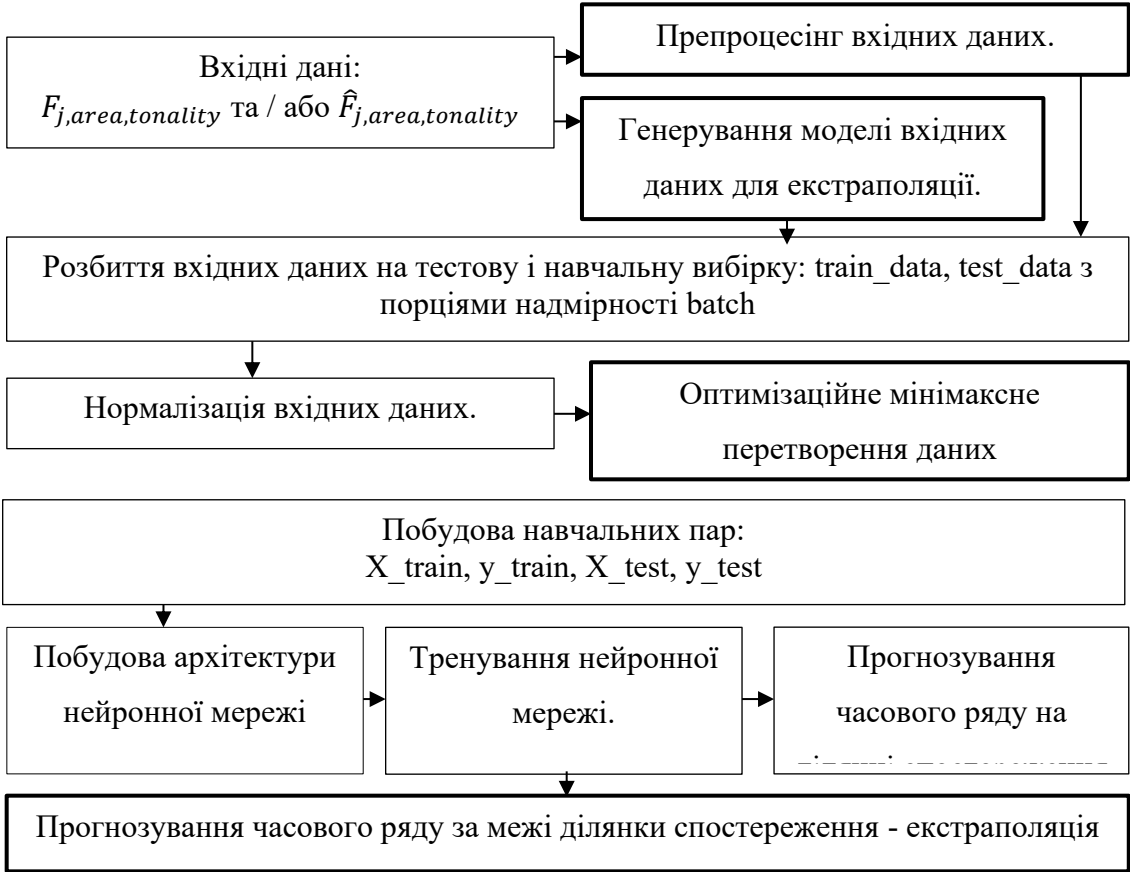


Рисунок 3.10. Схема обробки Time Series за методами глибинного навчання

У ході досліджень запропоновано модифікацію архітектури нейронної мережі та порядок формування тестової вибірки для прогнозування активності контенту в глобальних інформаційних мережах.

Дослідження мали фокус в згорткових нейронних мережах. Такий вибір обумовлений досить успішним досвідом їх застосування для обробки великих масивів Time Series із навчанням за накопиченою вибіркою вимірів. Це відповідає суті задачі прогнозування активності інформаційного контенту, коли для розв’язку задачі маємо вибірку Time Series формату $F_{j,area,tonality}$ (2.8) та / або $\hat{F}_{j,area,tonality}$ (2.10) – тобто одну реалізацію, з якої формується DataSet для навчання нейронної мережі.

Схема обробки Time Series за методами глибинного навчання передбачає реалізацію наступних етапів.

Класичні методології і технології глибинного навчання на Time Series мають специфіку переважно емпіричної реалізації та, як слідство, необхідність удосконалення наступних задач.

1. Формування метрики втрат, що мінімізуються в процесі навчання. Адже показники ефективності обробки / відтворення / екстраполяції часових рядів відрізняються від процесів обробки, наприклад природної мови та зображень. Обробка часового ряду орієнтовані на цільову задачу, а показники ефективності можуть включати вимоги мінімізацію випадкових похибок, або тотожність до вхідних даних тощо.

2. Структура вхідного прошарку та подальших прошарків і схема класичного навчання орієнтовані на вимогу «жорсткої» тотожності / відповідності навчальної та тестової вибірок. Однак, для обробки часового ряду це далеко не завжди прийнятна, оскільки може включати вимогу мінімізації випадкових похибок вхідних даних. У цьому сенсі більш прийнятною вбачається вимога мінімаксного наближення навчальної та тестової частини датасет.

3. Архітектура нейронної мережі за кількості внутрішніх прошарків, нейронів (окрім вхідного та вихідного прошарків), тип нейронної мережі залежить від структури вхідних даних цільовим завданням обробки та вимогами до показників ефективності. Архітектурні рішення, наразі, визначаються здебільшого емпірично та не мають аналітичних моделей. Практика сучасності демонструє домінантне використання трьох типів архітектурних рішень нейронних мереж: згорткові; рекурентні; трансформери великих мовних моделей – LLM. Але обумовленого аналітичного рішення стосовно вибору ефективного рішення для їх застосування ще не сформовано.

4. Обробка часових рядів методами глибинного навчання не поодиноким здійснюється на відносно малих структурах даних. Це викликано практичними сферами існування часових рядів. Тому,

застосовують розбиття даних на партії / batch для тиражування даних. Процес розбиття також реалізується переважно емпіричними методами.

5. Головною парадигмою нейронних мереж є робота з навчальними парами: навчальної та тестової вибірки. Прогнозування відбувається за поточним екземпляром, що має належати множині тестової вибірки. Для задач прогнозування поведінки часового ряду за межами інтервалу спостереження / екстраполяція це означає наявність даних про поведінку часового ряду за межами спостереження. Це є проблемним, оскільки для цього власне і будується мережа. Практика і актуальні дослідження вирішують цю задачу таких напрямках. Застосування рекурентних мереж, де попереднє значення використовується в якості елемента тестової вибірки для прогнозування – аналог рекурентної фільтрації з усіма її недолікам. Використання алгоритмів ковзного вікна. Перспективним в цьому сенсі вбачається застосування методів статистичного навчання з усіма запропонованими в роботі модифікаціями в поєднанні з методами імітаційного моделювання.

Фокус дисертаційних досліджень спрямовано на вирішення задач 1, 2, 5 із сформованого переліку:

формування метрики втрат, що мінімізуються в процесі навчання;

додавання оптимізаційного прошарку до нейронної мережі шляхом препроцесінгу вхідних даних;

застосування методів статистичного навчання із запропонованими в роботі модифікаціями в поєднанні з методами імітаційного моделювання для формування тестової вибірки для екстраполяції.

Реалізація зазначених задач в блоках схемі рис.3.6 виділено грубою лінією контуру.

Реалізація зазначених задач здійснювалась з урахуванням особливостей обробки часових рядів і властивостей масивів даних $F_{j,area,tonality}$. Зокрема наявність тренду із випадковим шумом.

Враховуючи зазначена за метрику втрат / похибки $()$, що мінімізується в процесі навчання обрано середньоквадратичну помилку: середньоквадратична похибка – Mean Squared Error (MSE).

Загальновідомо, що базова ідеологія навчання штучних нейронних мереж полягає у досягненні в ідеальному випадку абсолютної тотожності тренувальної і тестової частин / навчальної пари вхідного масиву даних / dataset. Це контролюється через мінімізацію метрики втрат. В базовій реалізації штучних нейронних мереж зазначене відображається в адитивності вхідних даних, що формують вихідний параметр окремого нейрона. Це своєрідна «жорстка» вимога тотожності вхід – вихід для навчальної пари dataset. Більш природнім вбачається «м'яка» вимога максимального наближення вхід – вихід навчальної пари. Це природні реалії процесів навчання та ідентифікації. Особливо це виправдано, коли вхідні дані, для нашого випадку Time Series, забруднені випадковим шумом. Тож наближення вхід – вихід в максимальній схожості, враховує ймовірнісні властивості вхідних даних. Реалізувати логіку «максимальної схожості» в альтернативу абсолютної тотожності навчальної пари пропонується із застосуванням логіки багатокритеріальної оптимізації з використанням нелінійної схеми компромісів.

В роботі запропоновано додати до класичної архітектури штучної нейронної мережи *прошарок препроцесінгу вхідних даних*, яких реалізує багатокритеріальну оптимізаційну логіку. Для відзнаки утворених таким чином нейронних мереж авторами запропоновано називати їх *оптимізаційними нейронними мережами*.

Нелінійна схема компромісів для розв'язку багатокритеріальних задач на дискретних даних реалізується за згортою часткових критеріїв [66]:

$$I = \sum_{l=1}^b \gamma_{0l} (1 - y_{0l})^{-1} \rightarrow \min, \quad (3.45)$$

де $l = 1 \dots b$ – кількість включених у згортку частинних критеріїв оптимальності; γ_{0l} – нормований ваговий коефіцієнт; y_{0l} – нормований частинний критерій.

Згортка (3.45) дозволяє отримати інтегровану оцінку як агрегація множини часткових критеріїв оптимальності. Модель (3.45) забезпечує нелінійною схемою компромісів за багатьох супересливих критеріїв і, порівняно з відомими аналогами, має доведені властивості [66]. Зокрема: рішення належить області Паретто та реалізує мінімаксну схему розв'язку багатокритеріальної задачі; на обмеженому інтервалі розгляду критеріїв згортка унімодальні.

Дослідження довели, що для обробки Time Series, забруднених шумом ваговий коефіцієнти γ_{0l} варто брати в зворотній пропорції до середньоквадратичного відхилення похибки вхідних даних $\gamma_{0l} = 1/\sigma_{y_{0l}}$. Цим досягається об'єктивна міра довіри до вхідних вимірів з урахуванням випадкової похибки їх отримання.

Згортка є такою, що мінімізується, тож має різні напрямки екстремуму для значень з тестової вибірки для тих, що наближаються до нуля (мінімізація) і для тих, що наближаються до одиниці (максимізація). Для цього пропонується реалізувати нормалізацію даних, що устримляються до одиниці з вимогою їх мінімізації. Тоді згортка для умовного нуля залишається незмінною (3.46), а для умовної одиниці трансформується до вигляду:

$$I = \sum_{l=1}^b \gamma_{0l} (1 - (1/y_{0l}))^{-1} \rightarrow \min, \quad (3.46)$$

Поріг умовної одиниці пропонується встановити, наприклад на рівні більше за 0.6.

Таким чином, вхідні дані y_0 нейронної мережі підпорядковані препроцесінгу відповідно до моделі перетворень:

$$I_{opt} = \begin{cases} \text{якщо } y_0 < 0.6 & \gamma_{0l}(1 - y_{0l})^{-1} \\ \text{якщо } y_0 > 0.6 & \gamma_{0l}(1 - (1/y_{0l}))^{-1} \end{cases} \quad (3.47)$$

Вираз (3.47) де-факто утворює додатковий до класичної архітектури вхідний прошарок нейронної мережі.

Навчена нейронна мережа інтерпретує (прогнозує) дані на часовому інтервалі існування тестової вибірки – на інтервалі спостереження. Для задач екстраполяції – прогнозування розвитку контрольованого процесу поза межами інтервалу спостереження – слід також мати тестову вибірку. Для цього пропонується застосування методів статистичного навчання із запропонованими в роботі модифікаціями в поєднанні з методами імітаційного моделювання.

Тоді тестова вибірка для екстраполяції матиме дві компоненти: оригінальна складова – вхідні дані $F_{j,area,tonality}$ на інтервалу спостереження та синтезована складова на інтервалі прогнозу y_{model} :

$$y_{test_ext} = [F_{j,area,tonality}, y_{model}], \quad (3.48)$$

Синтезована складова $y_{test_{out}}$ формується як адитивна суміш МНК оцінок вхідних параметрів – тренд y_{model0} та випадкова нормальна похибка φ із середньоквадратичним відхиленням оригінальної вибірки:

$$y_{model} = [y_{model0} + \varphi_i], i = 1 \dots m. \quad (3.49)$$

Запропоновані рішення стосовно метрики втрат, вхідного прошарку та процесу екстраполяції можуть бути застосовані для будь-якого типу нейронних мереж, що використовуються для обробки часових рядів: згорткові, рекурентні, LLM. В роботі наведено приклад розробки і

застосування згорткової нейронної мережі із впровадженням запропонованих новацій.

Дослідження довели ефективність архітектури оптимізаційної нейронної мережі згорткового типу та структури вхідних даних з параметрами: поділ вхідного датасет на порції $batch_size = 200$ вимірів; кількість нейронів вхідного, прихованих та вихідного прошарків становила: $n_neurons = [512, 256, 128, 64]$. Архітектура нейронної мережі побудована відповідно до регресу ступеневого ряду за основою двійка. Параметри структури та поділу вхідного масиву даних встановлені емпіричним шляхом та є найкращими для задач прогнозування розвитку часових рядів формату $F_{j,area,tonality}$.

Архітектура оптимізаційної нейронної мережі згорткового типу для обробки, прогнозування та екстраполяції Time Series подана на рис.3.11.

Дослідження ефективності запропонованих рішень здійснювалось методами математичного моделювання на реальних даних.

Умови досліджень. Обробці підлягала вибірка із 348 вимірів з монотонним трендом та випадковими помилками: нульове математичне очікування; середньоквадратичне відхилення $\sigma = 0.99$.

Екстраполяція здійснювалась на 100 вимірів, отже навчання реалізовувалось в межах 248 вхідних вимірів. Тестова вибірка для екстраполяції формувалась як адитивна суміш тренду від МНК та випадкового нормального шуму.

Модель тренду вхідної вибірки з параметрами, визначеними за алгоритмом МНК з препроцесінгом, описаним вище має вигляд (3.49).

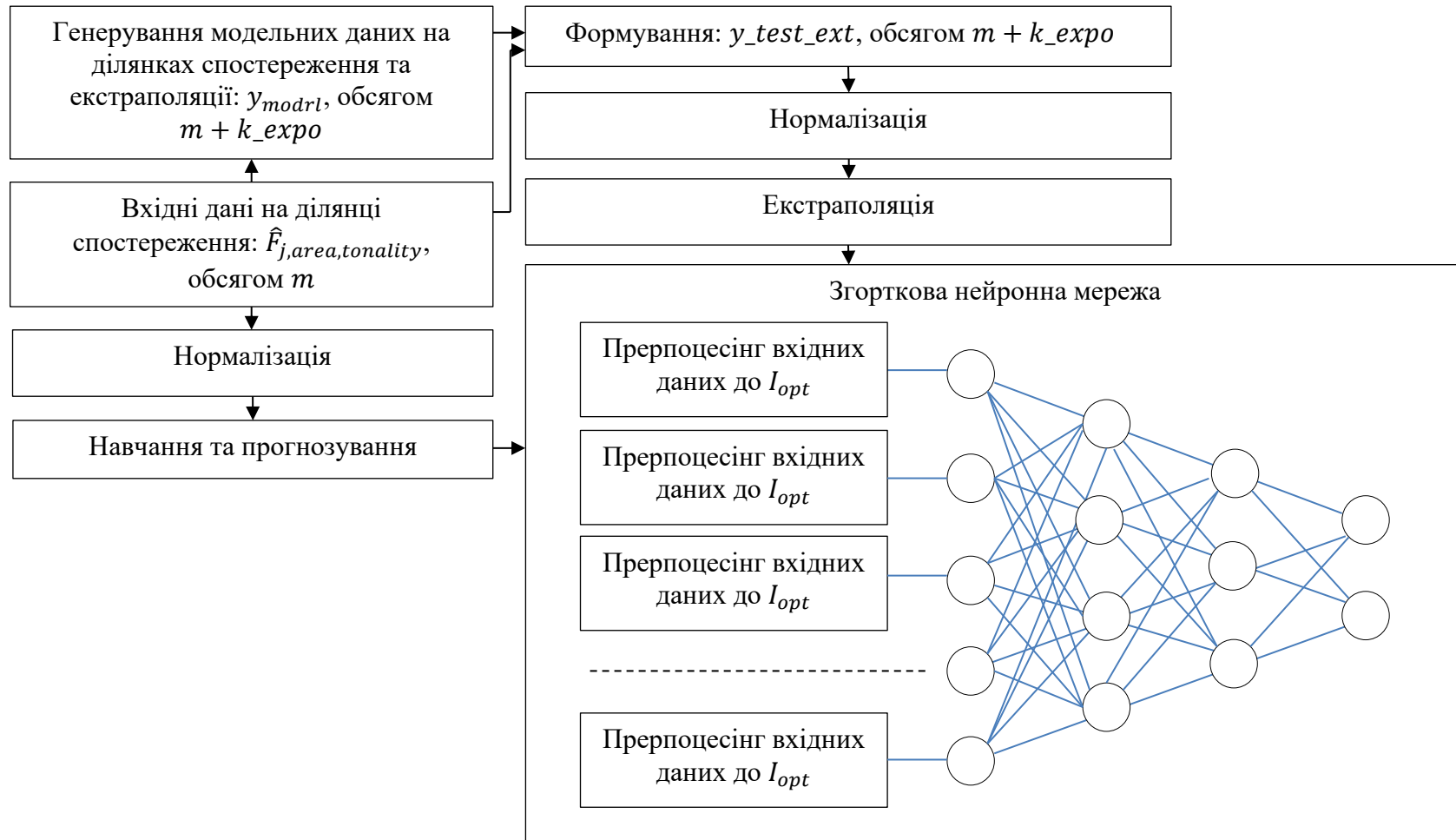


Рис.3.11. Архітектура оптимізаційної нейронної мережі згорткового типу для обробки, прогнозування та екстраполяції Time Series

$$\begin{aligned}
F(t) = & 28.03423083418919 + 0.004773157735137576t \\
& + 0.0008240042970011226t^2 \\
& - 3.204450289838111(e - 06)t^3 \\
& + 7.543074486030252(e - 10)t^4 \\
& + 6.87528764332385(e - 12)t^5
\end{aligned} \tag{3.49}$$

Динаміку зміни вхідної вибірки відображає графік рис.3.12.

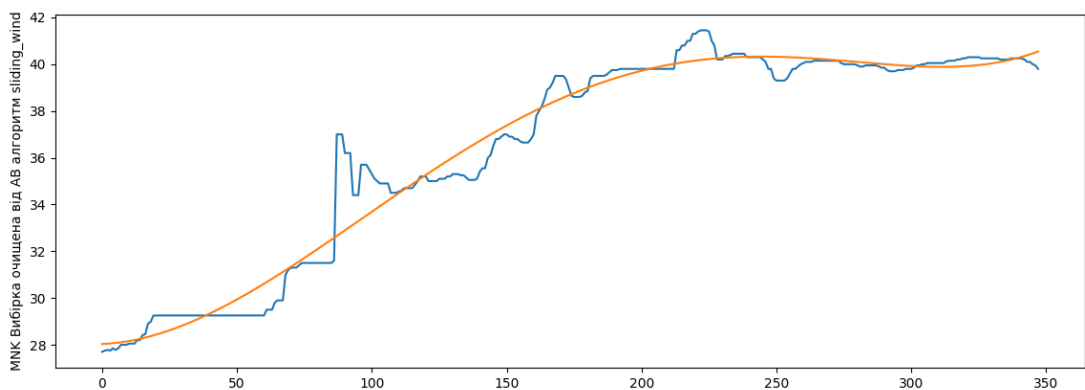


Рис.3.12. Вхідні дані та тренд МНК

На графіку рис.3.12 та на подальших графіках вісь абсцис позначає номер виміру дискретної вибірки / час – t , а вісь ординат – величину контрольованого параметру $F(t)$ в позначеннях (3.48).

Параметри загорткової нейронної мережі в модельному експерименті: поділ вхідного датасет на порції $batch_size = 200$ вимірів; епохи навчання $epochs = 100$; кількість нейронів вхідного, прихованих та вихідного прошарків становила: $n_neurons = [512, 256, 128, 64]$. Для генерування тестової вибірки для екстраполяції параметрів досліджуваного процесу поза межами інтервалу спостереження до тренду МНК (3.52), додано випадкова нормальна помилка із змінним значенням $\sigma = 1.0, 0.9, 0.1$.

Параметри навчання нейронної мережі формувались в діапазоні 0-248 вимірів. Тестова вибірка для екстраполяції формувалась як комбінація

реальних вхідних даних (інтервал 0-248) та синтезованих даних (інтервал 249-348). Екстраполяція забезпечувалась для 100 вимірів (інтервал 249-348) за синтезованими описаним способом тестовою екстраполяційною вибіркою з оцінюванням ефективності на відомій вхідній вибірці (348 вимірів).

Графік тестової вибірки для екстраполяції, як комбінація з реальних та синтезованих даних при $\sigma = 0.1$ подана на графіку рис.3.13.

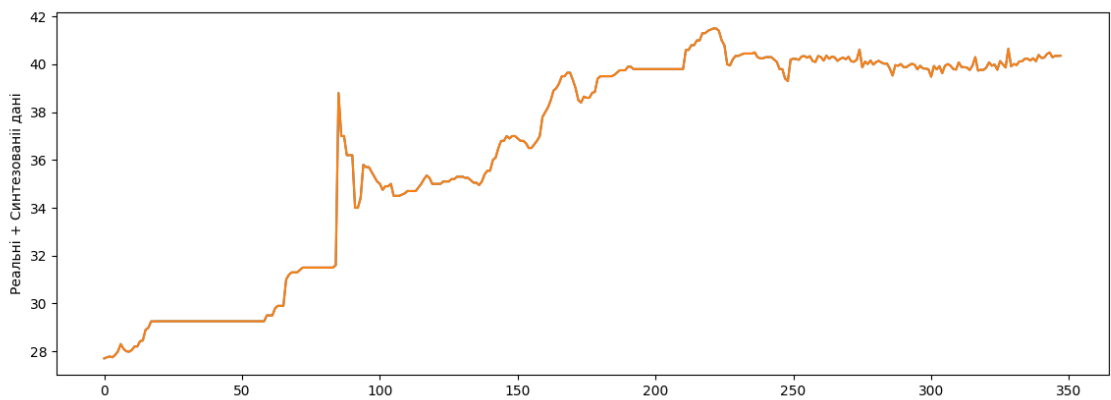


Рис.3.13. Графік тестової вибірки для екстраполяції

Контрольованими параметрами прогресу навчання та ефективності прогнозування і екстраполяції виступали: середня абсолютна похибка – Mean Absolute Error (MAE); середньоквадратична похибка – Mean Squared Error (MSE) / Root Mean Squared Error (RMSE); коефіцієнт детермінації / ймовірність апроксимації – R-squared (R2) Score; втрати – loss.

Результати досліджень подані в табл.3.5 – 3.7 та на рис.3.14-3.16. Аналіз наведених результатів показав, що встановлена архітектура нейронної мережи та підготовка даних для визначених умов є прийнятними для процесів прогнозування та екстраполяції параметрів часового ряду не менше ніж 0,5 інтервалу спостереження. Про це доводять високі показники коефіцієнту конкордації (зберігається на рівні 0,9) та відносно низький рівень похибок моделі (MSE) та середньоквадратичного відхилення випадкової помилки (MAE).

Таблиця 3.5

Результати досліджень оптимізаційної нейронної мережі

Параметр	Динаміка навчання			Прогнозування	Екстраполяція
	1	22	100		
$\sigma = 1.0$					
Без оптимізаційного шару					
MAE	2.2293	0.0604	0.0080	0.00770	1.30636
MSE	21.4090	0.0072	9.3980e-05	14. 35234e-05	2.48986
RMSE	-	-	-	0.011980	1.57792
R2	-50.7250	0.9825	0.9998	0.999523	0.86950
loss	21.4090	0.0072	9.3980e-05	-	-
Із оптимізаційним шаром					
MAE	2.2253	0.3043	0.0034	0.00287	0.66466
MSE	21.2251	0.1233	2.6881e-05	3.41680e-05	0.73367
RMSE	-	-	-	0.00584	0.85654
R2	-50.2806	0.7022	0.9999	0.99990	0.96154
loss	21.2251	0.1233	2.6881e-05	-	-

Таблиця 3.6

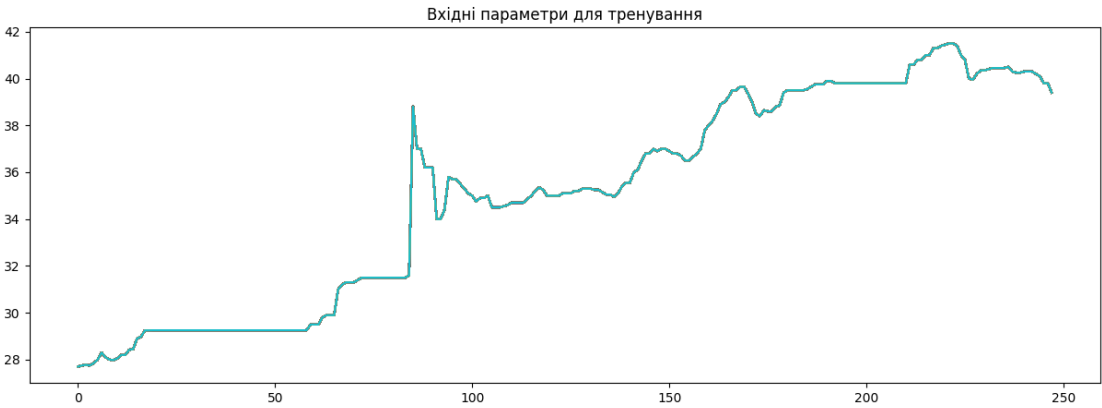
Результати досліджень оптимізаційної нейронної мережі

Параметр	Динаміка навчання			Прогнозування	Екстраполяція
	1	22	100		
$\sigma = 0.9$					
Без оптимізаційного шару					
MAE	2.0302	0.2032	0.0058	0.006060	0.973266
MSE	17.9008	0.0758	6.7122e-05	0.000130	1.446210
RMSE	-	-	-	0.011414	1.202585
R2	-42.2490	0.8168	0.9998	0.999597	0.924202
loss	17.9008	0.0758	6.7122e-05	-	-
Із оптимізаційним шаром					
MAE	2.0968	0.1127	0.0198	0.014862	0.646525
MSE	19.2580	0.0193	4.7718e-04	0.000352	0.727511
RMSE	-	-	-	0.018762	0.852942
R2	-45.5280	0.9533	0.9988	0.999012	0.961870
loss	19.2580	0.0193	4.7718e-04	-	-

Таблиця 3.7

Результати досліджень оптимізаційної нейронної мережі

Параметр	Динаміка навчання			Прогнозування	Екстраполяція
	1	22	100		
$\sigma = 0.1$					
Без оптимізаційного шару					
MAE	1.9222	0.2149	0.0063	0.006659	0.115665
MSE	16.5236	0.0731	9.5741e-05	7.494345e-05	0.042147
RMSE	-	-	-	0.008656	0.205299
R2	-38.9217	0.8234	0.9998	0.999814	0.997790
loss	16.5236	0.0731	9.5741e-05	-	-
Із оптимізаційним шаром					
MAE	1.9933	0.0625	3.6219e-06	0.000962	0.086257
MSE	17.8145	0.0118	0.0013	2.571392e-06	0.039888
RMSE	-	-	-	0.001603	0.199721
R2	-42.0404	0.9715	3.6219e-06	0.999993	0.997909
loss	17.8145	0.0118	1.0000	-	-



а)

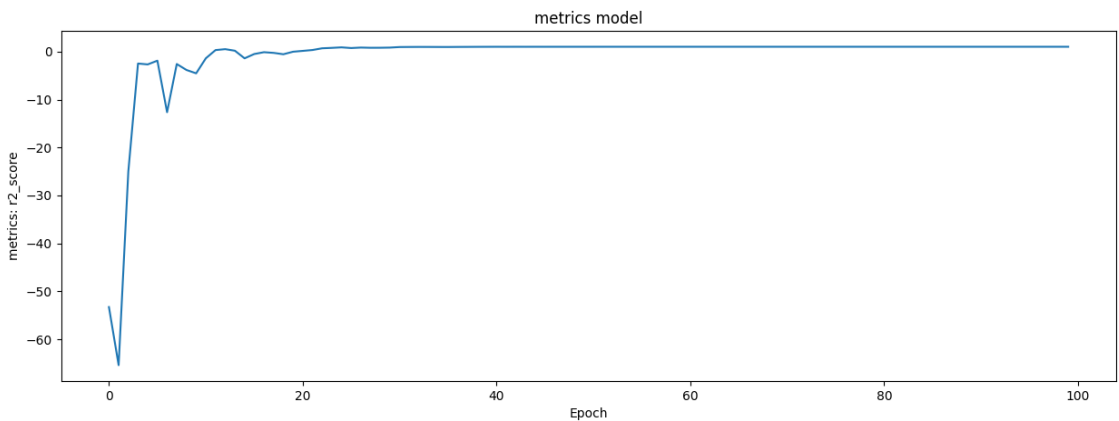
Вхідні параметри для тренування на ділянці спостереження



б)

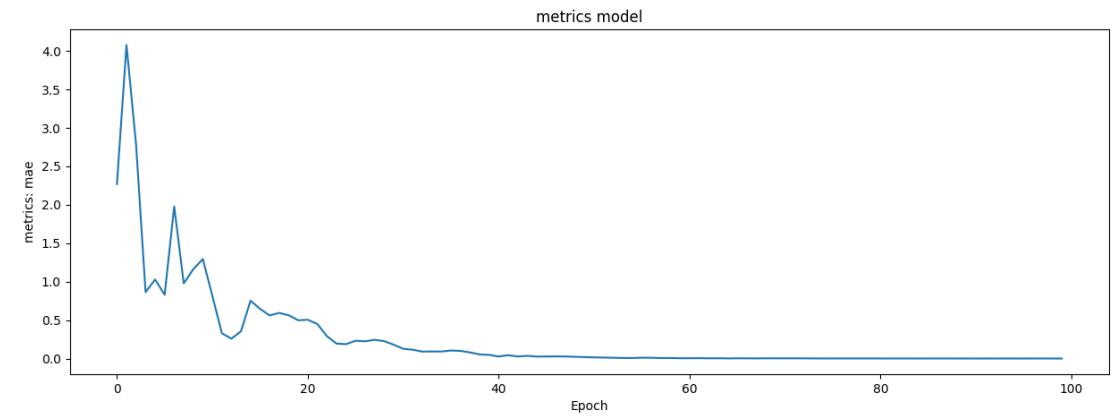
Вхідні параметри для тестування на ділянці спостереження

Рис.3.14. Графік вхідних даних



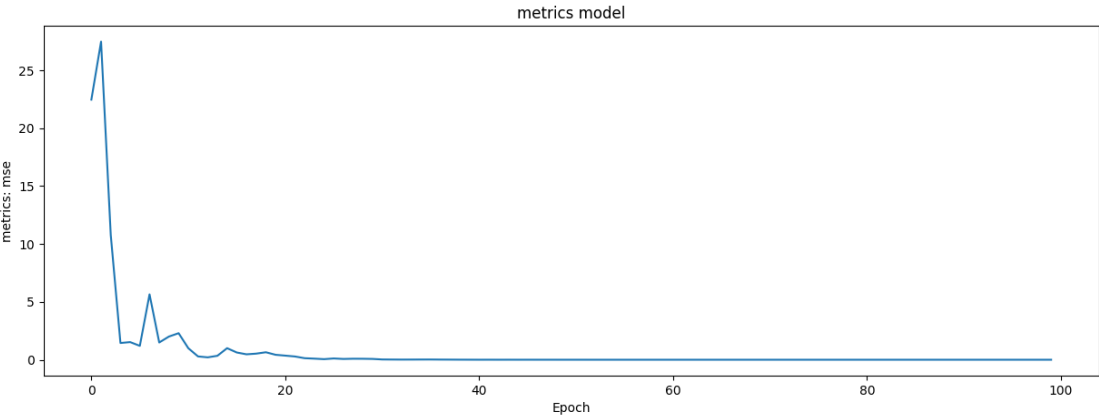
а)

Коефіцієнт детермінації



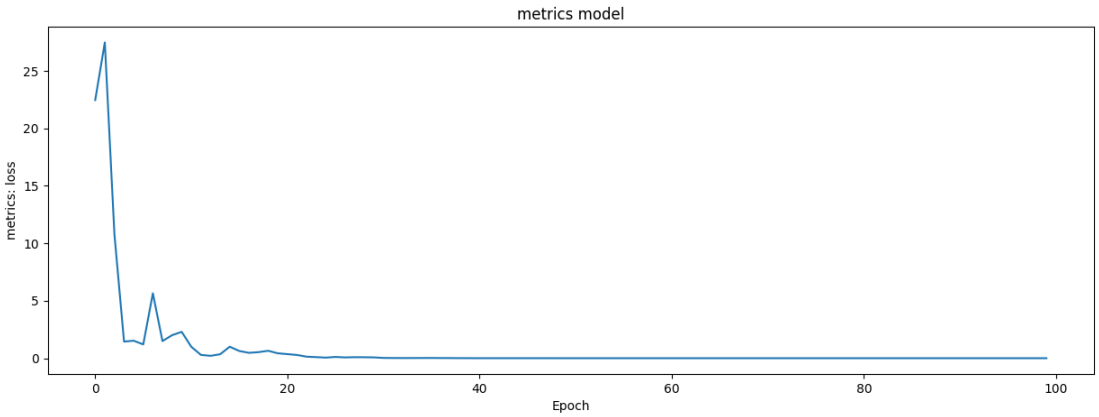
б)

Середня абсолютна похибка



в)

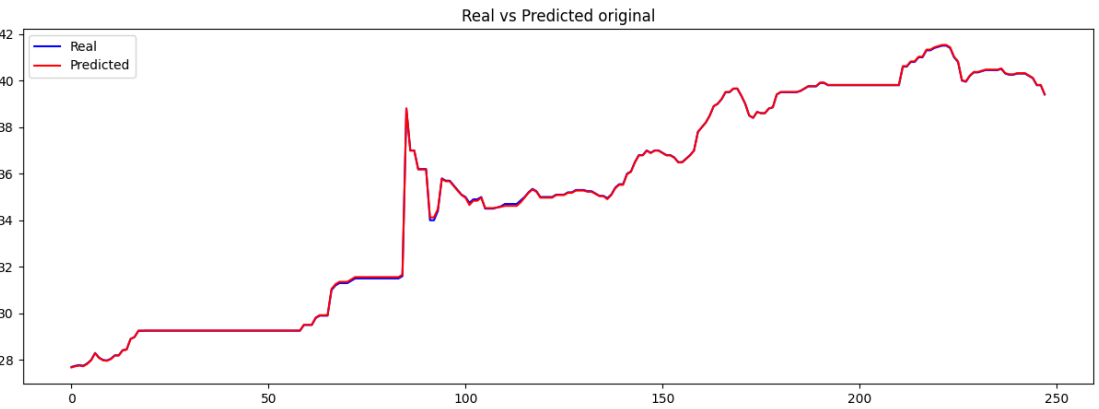
Середньоквадратична похибка



г)

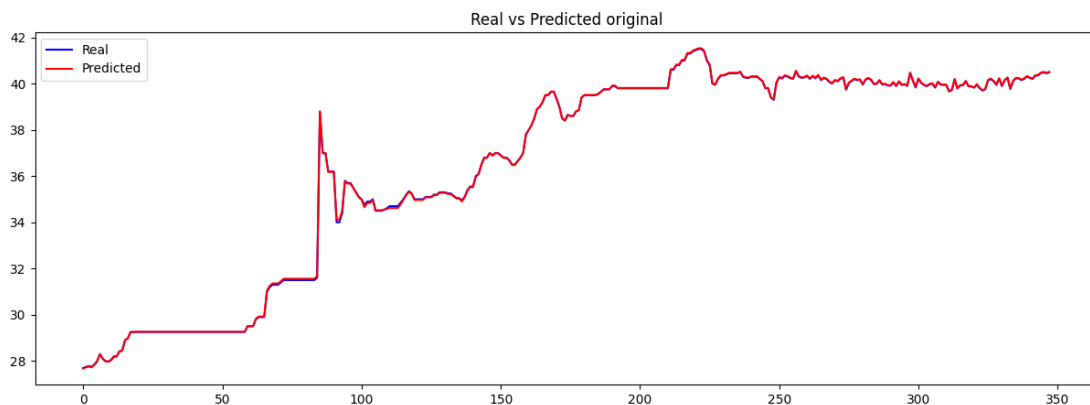
Втрати

Рис.3.15. Динаміка показників ефективності навчання



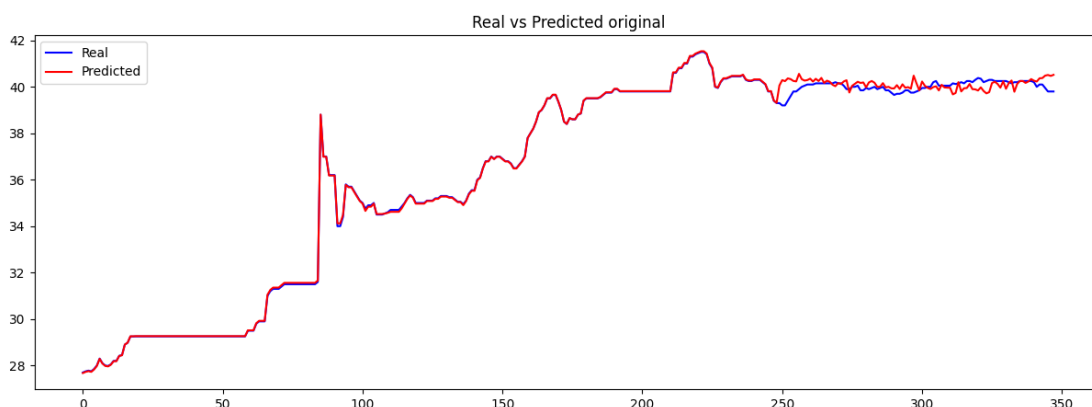
а)

Результати прогнозування



б)

Результати екстраполяції



в)

Результати екстраполяції та реальні дані

Рис.3.16. Графік результатів прогнозування та екстраполяції

На ділянці екстраполяції спостерігається типове зростання похибок за показниками MSE та MAE із збереженням значення ймовірності апроксимації – R-squared (R^2) на високому рівні 0,9. Отже, сформовані рішення та отримані результати придатні для практичного застосування.

Для інтервалу спостереження і екстраполяції має місце виграв за усіма показниками в точності від застосування оптимізаційної нейронної мережі. Зростання величини випадкової складової похибки в даних лише посилює виграв який становить від 5 до 20%, залежно від показника та похибки вхідних даних.

В якості тестової вибірки запропоновано застосування МНК. Поєднання методу статистичного та глибинного навчання має позитивні результати та значні перспективи, які обумовлені наступним. Нейронна мережа здатна акумулювати значні обсяги інформації про реалізацію часового ряду як випадкового процесу. Цим досягається гнучкий механізм відтворення прогнозованих та екстрапольованих даних з аналогією із накопичуванням статичної інформації за методом статистичних випробувань Монте-Карло. Тобто нейронна мережа здатна вміщувати в собі досвід значної кількості реалізацій випадкових процесів формату часового ряду. При цьому МНК вдало доповнює нейронну мережу своїми прогностичними властивостями за одної єдиної реалізації випадкового процесу. Наведені дослідження реалізовані з навчанням нейронної мережі за однією реалізацією часового ряду. На обмеженому складі даних навчання реалізовано у традиційний спосіб розбиття вхідних даних на рівні порцій – `batch_size`. Варто очікувати позитивні результати в оптимізаційному прошарку та у підтримці МНК для навчання нейронних мереж за великими даними. Для практичної задачі прогнозування активності інформаційного контенту це вбачається у накопиченні досвіду розв'язку такої задачі в нейронній мережі за ретроспективними даними.

Досвід застосування оптимізаційного прошарку лише на вхідному рівні довів стійкість результатів до випадкових похибок даних. В подальшому передбачається синтез нейронних мереж з операцією згортки, що поширюється на інші прошарки на нейронному рівні.

3.4. Інтегрована оцінка для багатофакторного прогнозування активності контенту в глобальних інформаційних мережах

Інтегрована оцінка для багатофакторного прогнозування активності контенту в глобальних інформаційних мережах забезпечує агрегацію множини індикаторів активності до єдиного показника, в якому враховано

інформацію про сферичні тональності $F_{j,area,tonality}$, $j = 1 \dots N$. Мета агрегації – підтримка процесу пошуку прогнозованих біфуркаційних інтервалів як критичної активності інформаційного контенту.

Вхідною інформацією для визначення інтегрованої оцінки та багатофакторного прогнозування активності контенту в глобальних інформаційних мережах є багатовимірні часові ряди (див. рис.3.1):

первинні часові ряди частот тональності інформаційних повідомлень за сферами $F_{j,area,tonality}$, $j = 1 \dots N$;

результатів статистичного навчання за сферичною тональністю;

результатів глибинного навчання за сферичною тональністю;

результатів статистичного навчання та глибинного навчання.

Запровадження та оцінювання динаміки зміни інтегрованої оцінки за вхідними масивами $F_{j,area,tonality}$ має істотне прикладне значення. Це виявляється в агрегації індикаторів інформаційної активності за головними сферами діяльності соціуму. Тож баланс тональності інформаційних повідомлень з відсутністю асимптотичного зростання інтегрованого показника свідчить про стабільність соціальних настроїв без негативних дій та біфуркаційних переходів.

Пошук прогнозованих біфуркаційних інтервалів як критичної активності інформаційного контенту пропонується реалізувати методами фрактального аналізу часових рядів з контролем індексу Херста (див. вир. (3.4)) [00]. Значення індексу Херста розраховуватиметься за інтегрованою оцінкою вхідних показників $F_{j,area,tonality}$.

Інтегровану оцінку інформаційної активності за множиною показників $F_{j,area,tonality}$ пропонується визначати за оптимізаційною моделлю нелінійної схеми компромісів [66].

$$It_scor_j^{tonality} = \sum_{area=1}^7 \gamma_{0l} (1 - F_{j,area,tonality0})^{-1}, \quad (3.50)$$

$$j = 1 \dots N$$

де: γ_{0l} – ваговий коефіцієнт значущості сфери з переліку *area*, який визначає її пріоритет в інтегрованому показнику. Для рівнозначної оцінки сфер – приймає значення одиниці;

$F_{j,area,tonality0}$ – нормований показник.

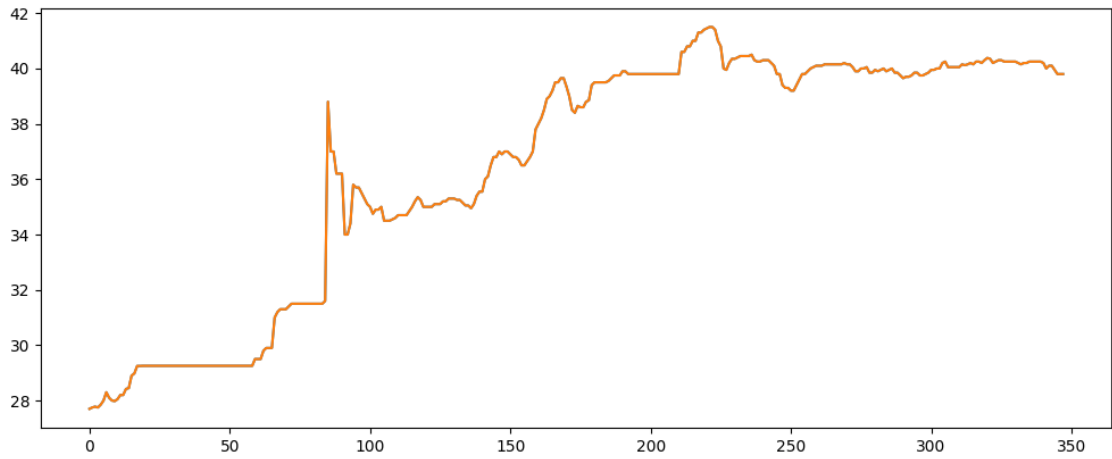
Інтегрований показник (3.50) утворює три часові ряди, залежно від тональності: *negative, neutral, positive*.

Нормування часових рядів $F_{j,area,tonality}$ реалізується відносно глобального максимального значення в межах усіх масивів $\max(F_{j,area,tonality})$:

$$F_{j,area,tonality0} = \frac{F_{j,area,tonality}}{\max(F_{j,area,tonality}) + \Delta}, j = 1 \dots N. \quad (3.51)$$

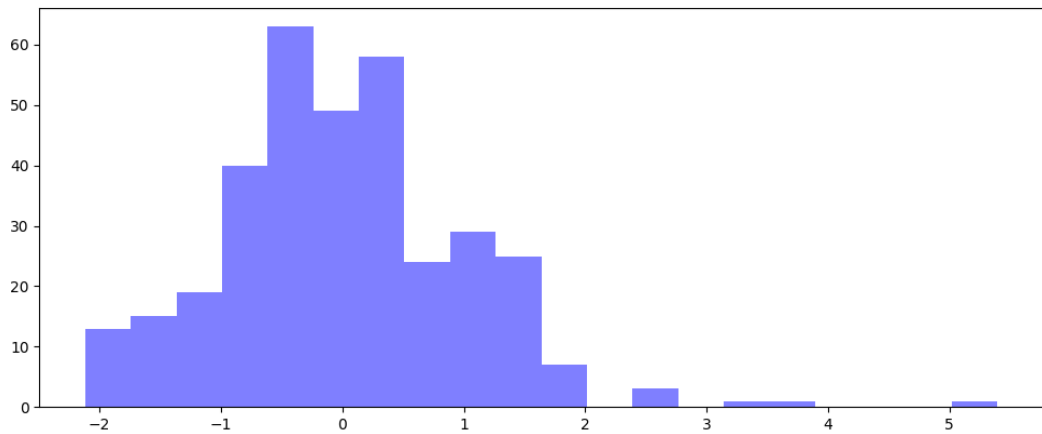
Параметр Δ – коефіцієнт запасу приймає значення $\Delta = 0.1 \dots 0.3$ та призначений для позбавлення операції ділення на нуль в згортці (3.50), якщо результат нормалізації $F_{j,area,tonality} / \max(F_{j,area,tonality})$ дає одиницю.

Дослідження ефективності запропонованих рішень проводилось методами математичного моделювання. Аналізу підлягали три часові ряди, забруднені пропусками та аномаліями. Характеристики часового ряду: кількість вимірів 348. Передбачалось, що часові ряди відображали різні показники одного і того ж фізичного явища. Статистичні характеристики вимірів, їх графіки та гістограми законів розподілу випадкових похибок подані на графіках рис.3.17-3.19.



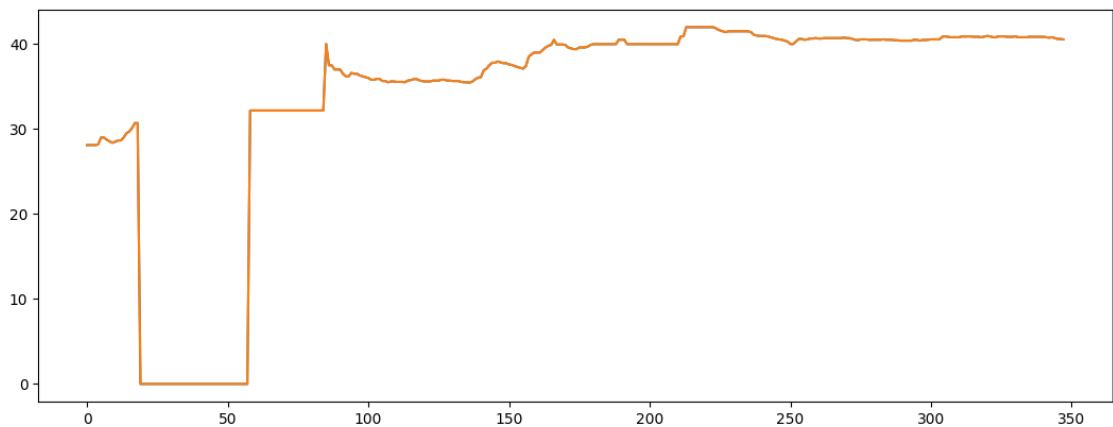
а)

Графік дискретного процесу в аналоговій формі



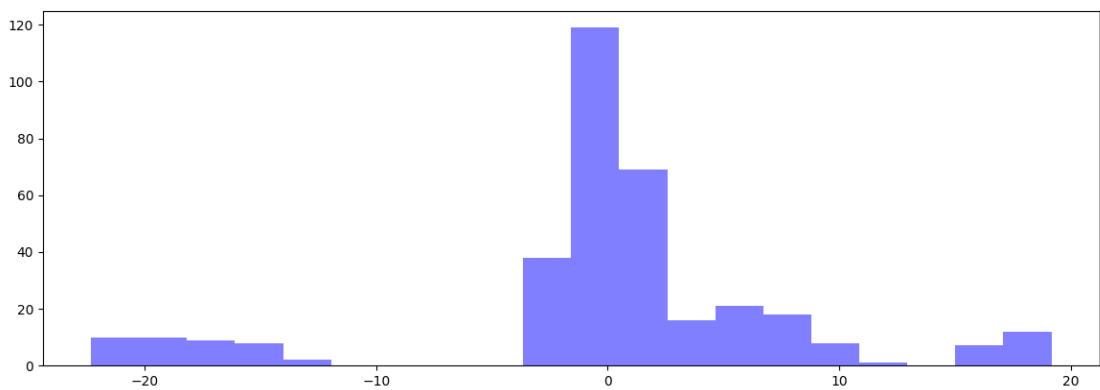
б)

Гістограма закону розподілу випадкової похибки без тренду:
 математичне сподівання випадкової похибки: -0.09548375755047012 ;
 дисперсія випадкової похибки: 0.9891767028531205 ;
 СКВ випадкової похибки: 0.9945736286736747
 Рис.3.17. Характеристики першого часового ряду



а)

Графік дискретного процесу в аналоговій формі



б)

Гістограма закону розподілу випадкової похибки без тренду:

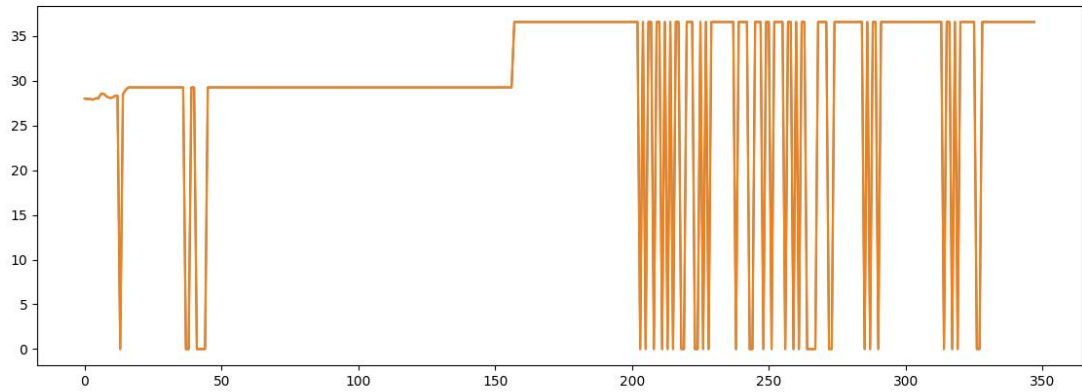
математичне сподівання випадкової похибки:

0.22432316911765326;

дисперсія випадкової похибки: 63.896746894523595;

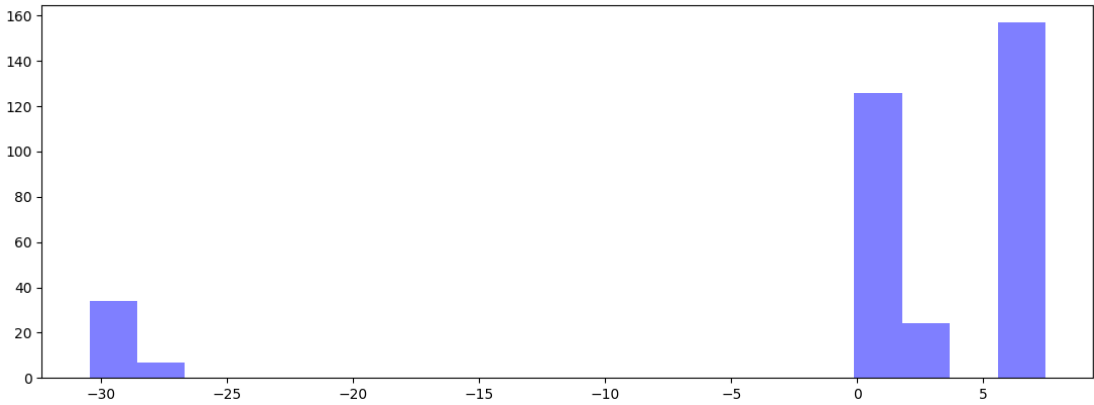
СКВ випадкової похибки: 7.993544075973035.

Рис.3.18. Характеристики першого часового ряду



а)

Графік дискретного процесу в аналоговій формі



б)

Гістограма закону розподілу випадкової похибки без тренду:
 математичне сподівання випадкової похибки: 1.9970139022351727;
 дисперсія випадкової похибки: 123.68764833066759;
 СКВ випадкової похибки: 11.12149487841754.

Рис.3.19. Характеристики першого часового ряду

На графіках рис.3.17-3.19 подано зміну контрольованого часового ряду (рис.3.17,а – рис.3.19,а – вісь ординат позначає величину значення часового ряду, вісь абсцис – часовий аргумент), гістограми закону розподілу випадкової складової (рис.3.17,б - рис.3.19,б – вісь ординат позначає частоту потрапляння, реалізації випадкової похибки в порційний

довірчий інтервал діапазону, відкладеного на осі абсцис) та числові характеристики випадкової складової.

Аналіз графіків та статистичних характеристик рис.3.17-3.19 показує наступне. Часові ряди мають монотонний тренд, що зростає, випадкові складові (залишки / шум) з нормальним законом розподілу та аномалії типу грубі виміри (див. залишки вимірів на гістограмі поза межами довірчого інтервалу в три СКВ рис.3.17,б – рис.3.19,б) та пропуски даних. Цікавим є факт нормальності закону розподілу випадкових похибок, що для реальних даних свідчить про множину випадкових факторів, що впливають на наслідки досліджуваного процесу.

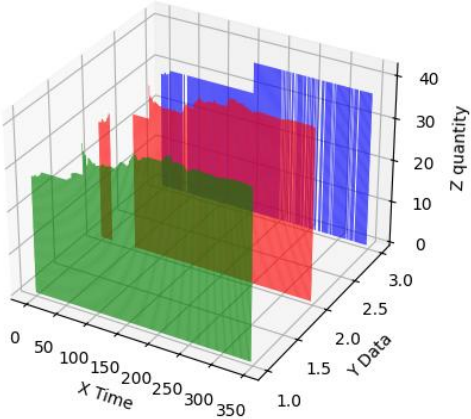
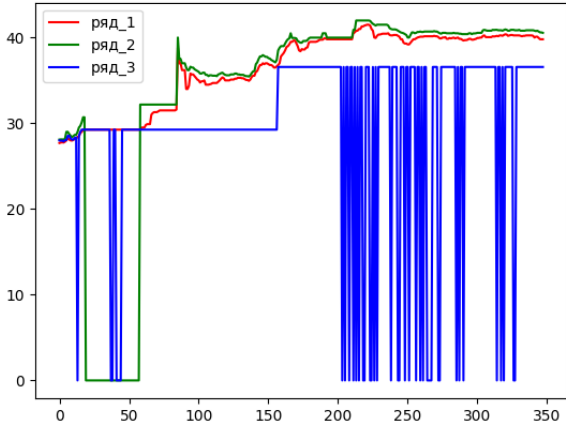
Враховуючи виявлені властивості, вхідні дані підлягали препроцесінгу шляхом відновлення пропусків. Пропуски, для конкретного випадку, замінювались попередніми даними. Результати відновлення, порівняно із вхідними даними та нормалізації вхідних даних подано на графіках табл. 3.8.

В табл.3.8 подано графіки 3d і 2d візуалізації часових рядів, що підлягають обробці. На всіх графіках вісь абсцис відображає час, а вісь ординат – величину контрольованого параметру. Просторова форма за третю координату має перелік трьох параметрів. В останньому стовпчику таблиці – подано значення індексу Херста, розрахованого ковзним вікном з множинною кількістю інтервалів (перше значення) та шляхом розбиття вибірки на два рівні інтервали (друге значення). Останнє є менш чутливим до пропусків та аномальних викидів.

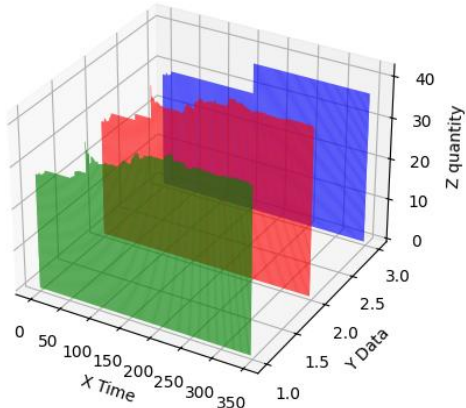
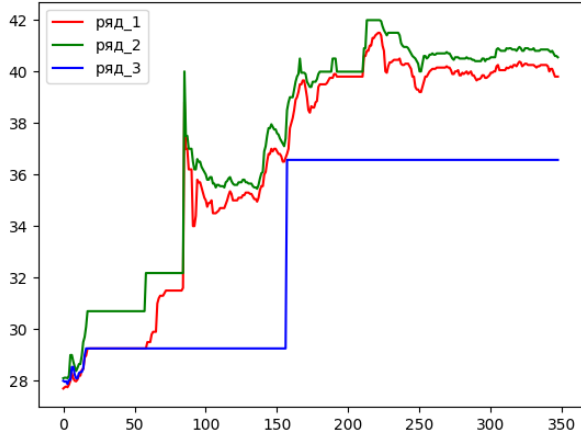
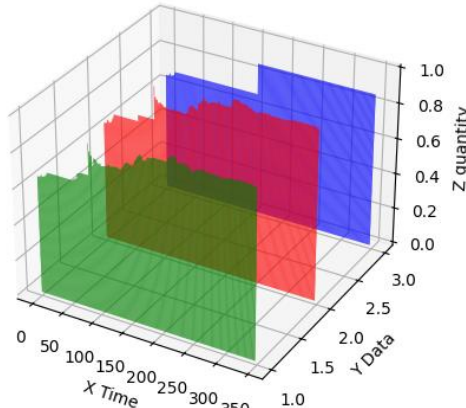
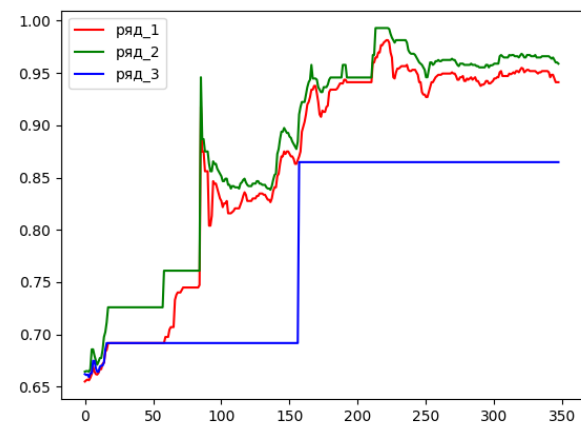
Наведені результати – графіки та індекс Херста, їх аналіз свідчить про збереження загальних тенденцій в даних після їх препроцесінгу. Більш того, відновлення даних позитивно відображається в адекватності індексу Херста. Отже, проведенні операції відновлення та нормалізації є прийнятними для подальших етапів оброблення і формування інтегрованої оцінки для багатофакторного прогнозування активності контенту в глобальних інформаційних мережах.

Таблиця 3.8

Результати досліджень препроцесінгу даних: відновлення, нормалізація

3d візуалізація	2d візуалізація	Індекс Херста, Н
Вхідні дані до відновлення пропусків		
		Ряд_1
		0.5098
		Ряд_2
		0.3215
		Ряд_3
		0.2307

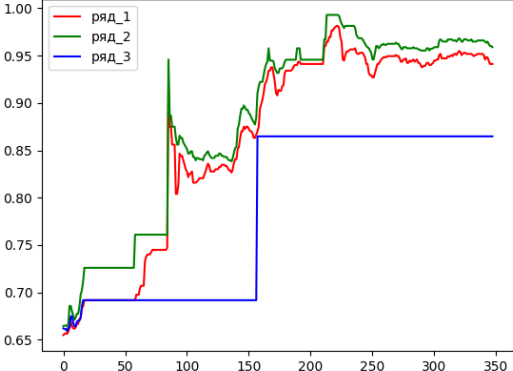
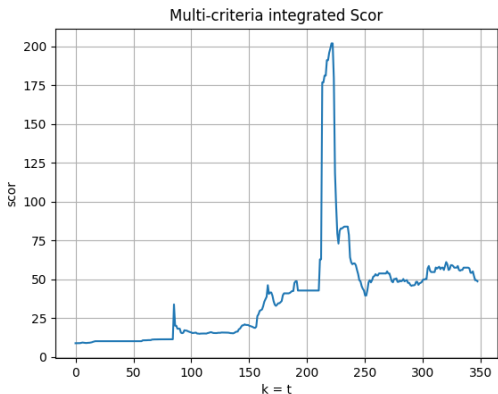
Продовження таблиці 3.8

3d візуалізація	2d візуалізація	Індекс Херста, Н
Вхідні дані після відновлення пропусків		
		Ряд_1
		0.5579; 0.5098
		Ряд_2
		0.5647; 0.5207
		Ряд_3
		0.4801; 0.3095
Вхідні дані після нормалізації		
		Ряд_1
		0.5579; 0.5098
		Ряд_2
		0.5647; 0.5207
		Ряд_3
		0.4801; 0.3095

Подальші дослідження, результати яких внесені до табл. 3.9-3.12, репрезентують процес отримання інтегрованого показника в різній комбінаториці підготовки та / або обробки первинних даних. Це реалізується з метою дослідження можливостей і закономірностей в прогнозуванні поведінки досліджуваного процесу як суперпозиція впливу множинних факторів.

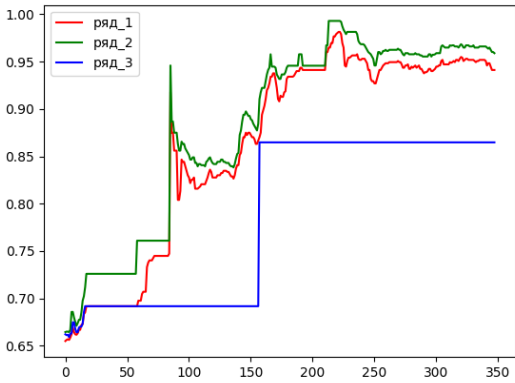
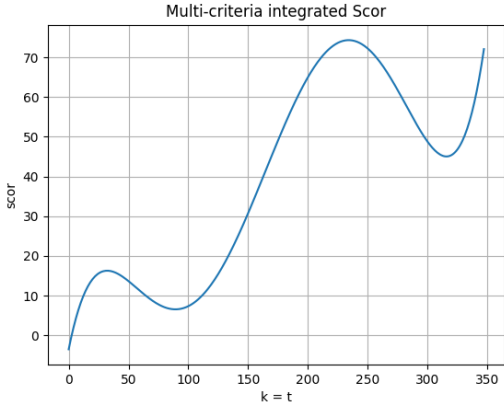
Таблиця 3.9

Результати розрахунку інтегрованого показника

Вхідні дані	Інтегрований показник
Вхідні дані відновлені та нормалізовані. Інтегрований показник в оригінальному вигляді.	
	
Індекс Херста, Н	
Ряд_1	-
0.5579; 0.5098	0.8653; 0.3805
Ряд_2	
0.5647; 0.5207	
Ряд_3	
0.4801; 0.3095	

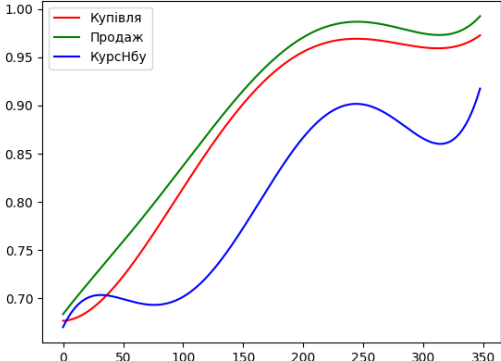
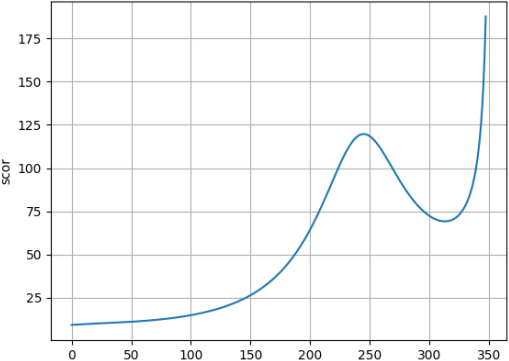
Таблиця 3.10

Результати розрахунку інтегрованого показника

Вхідні дані		Інтегрований показник	
Вхідні дані відновлені та нормалізовані. Інтегрований показник оброблений МНК.			
			
Індекс Херста, Н			
Ряд_1		r2_score = 0.521	
0.5579; 0.5098		- ; 0.3447	
Ряд_2			
0.5647; 0.5207			
Ряд_3			
0.4801; 0.3095			

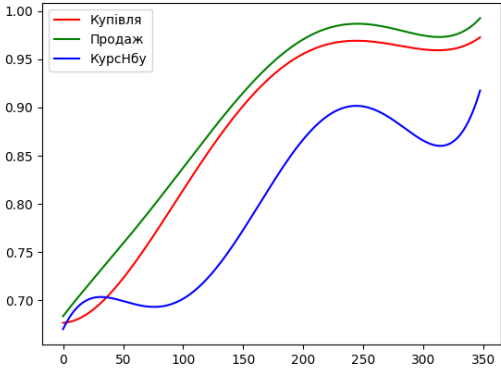
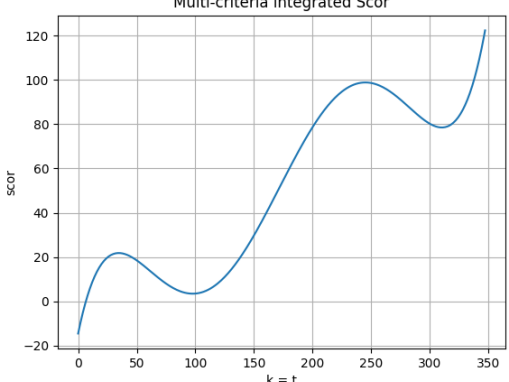
Таблиця 3.11

Результати розрахунку інтегрованого показника

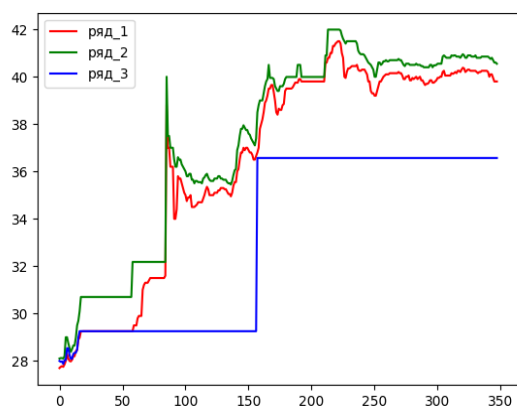
Вхідні дані	Інтегрований показник
Вхідні дані відновлені, нормалізовані та оброблені МНК. Інтегрований показник в оригінальному вигляді.	
	
Індекс Херста, Н	
Ряд_1, r2_score=0.963	-
0.0157; 0.607	0.3722; 0.4227
Ряд_2, r2_score=0.957	
-0.5854; 0.672	
Ряд_3, r2_score=0.898	
0.0305; 0.399	

Таблиця 3.12

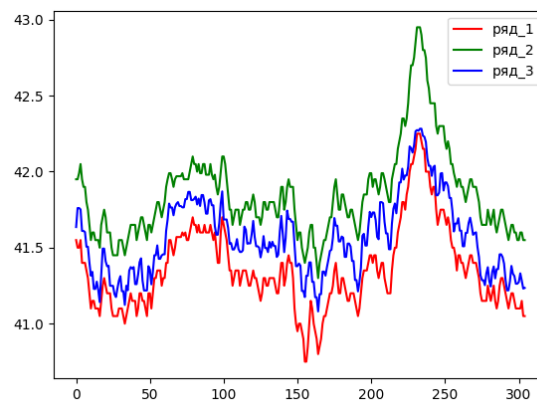
Результати розрахунку інтегрованого показника

Вхідні дані	Інтегрований показник
Вхідні дані відновлені, нормалізовані та оброблені МНК. Інтегрований показник в оригінальному вигляді.	
	
Індекс Херста, Н	
Ряд_1, r2_score=0.963	r2_score=0.916
0.0157; 0.607	- ; 0.4088
Ряд_2, r2_score=0.957	
-0.5854; 0.672	
Ряд_3, r2_score=0.898	
0.0305; 0.399	

На графіках рис.3.17. подано графіки досліджуваного процесу на ділянці спостереження та його розвиток поза межами цієї ділянки. Наведені графіки дозволяють оцінити дійсну динаміку поведінки досліджуваного процесу та порівняти її із результатами розрахунків інтегрованого показника.



а) ділянка спостереження



б) розвиток процесу

Рис.3.17. Графіки досліджуваного процесу на ділянці спостереження та його розвиток поза межами цієї ділянки

Для сумісного аналізу результатів розрахунків, поданих на графіках рис.3.17 та в табл.3.9-3.12 варто нагадати про значення показника Херста: $H > 0.5$ – збереження динаміки процесу; $H < 0.5$ – процес змінить спрямованість; $H = 0.5$ – випадковий білий шум.

Для всіх варіантів обробки інтегрований показник за індексом Херста свідчить про зміну динаміки індикаторів у майбутньому ($H < 0.5$). При цьому візуальний аналіз схиляє до висновку збереження тенденцій асимптотичного зростання трендів (рис.17.6,а). Однак, реальний розвиток досліджуваного процесу (рис. рис.3.17,б) свідчить на користь зміни його динаміки в бік сезонного коливання відносно статистичного середнього із аномаліями та випадковим шумом.

Таким чином, запропонований підхід в процесах обробки та композиції часткових індикаторів до інтегрованого показника за нелінійною схемою компромісів є дієвим та прийнятним для практичного спрямування.

Найближчим аналогом продемонстрованого підходу може бути багатовимірний регресійний аналіз. Однак, його застосування для нелінійних процесів із випадковими похибками виявився неефективним.

Використання в якості схеми агрегації часткових показників нелінійної схеми компромісів (3.50) забезпечує мінімаксну логіку врахування часткових

показників. Тобто принцип міркування «як найкраще – як найгірше». Це є природною моделлю суб'єктивного судження на об'єктивно виміряні процеси реальної дійсності – оцінювання міри активності інформаційного контенту в глобальних мережах. Тобто на етапах формалізації, моделювання та розрахунків є факт підвищення адекватності інтерпретації досліджуваного явища.

Варто зазначити, що інтегрований показник розраховано та використано в умовних одиницях, без нормалізації. Такий підхід є цілком виправданий, оскільки аналізу підлягає динаміка показника із пошуку біфуркаційних інтервалів – ділянки часу (номерів вимірів) де відбувається асимптотичне зростання інтегрованого показника за результатами контролю індексу Херста (див. рис.1.5). Часова шкала інтегрованого показника не зазнала змін та є абсолютною.

Наведені результати досліджень дозволяють підтвердити комбінаторику результатів обробки первинних індикаторів та інтегрованого показника у відповідності до схеми рис.3.1.

Отже, наведена послідовність дій та використані математичні моделі в комплексі дозволяють реалізувати розрахунок інтегрована оцінка для багатофакторного прогнозування активності контенту в глобальних інформаційних мережах.

3.5. Формалізований опис та індикатори ефективності математичної моделі багатофакторного прогнозування активності контенту в глобальних інформаційних мережах

У ході досліджень вперше запропоновано математичну модель багатофакторного прогнозування активності контенту в глобальних інформаційних мережах, яка базується на обробці та аналізі часових рядів з використанням методів статистичного та глибинного машинного навчання.

Розроблена математична модель є композицією часткових математичних моделей, представлених етапами структурної схеми рис. 3.1.

1. Підготовчий етап з аналізом та препроцесінгом даних.

2.1. Виявлення властивостей вхідних даних: розрахунок статистичних характеристик – вирази (3.1); аналіз фрактальних властивостей з контролем індексу Херста (3.4).

2.2. Очищення часового ряду від аномалій реалізується з алгоритмами, що базуються на реалізації ковзного вікна: із заміною вимірів оцінками математичного очікування (3.5); контролю статистичних характеристик вимірів (3.8); аналіз змін динаміки контрольованого процесу (3.12).

2.3. Структурна оптимізація моделі за швидким (в межах одного проходу верхівки часового ряду) алгоритмом контролю значень похідних високих порядків (3.28).

2. Обробка та прогнозування часових рядів методами статистичного та глибинного навчання.

2.1. Статистичне навчання з нелінійними моделями реалізовано із впровадженням методу диференціально-нетейлорівських перетворень для синтезу: експоненціальної моделі (3.38); трансцендентної двокомпонентної моделі (3.39).

2.2. Обробки часових рядів методами глибинного навчання з модифікацією згорткових нейронних мереж для обробки часових рядів із впровадженням нелінійної схеми компромісів в моделі нейронної структури (3.45) та моделюванням навчальної вибірки для екстраполяції за методами статистичного навчання (3.48), (3.49).

3. Розрахунок інтегрованої оцінки (3.50) для багатфакторного прогнозування активності контенту в глобальних інформаційних мережах з пошуком біфуркаційних інтервалів.

Запропонована математична модель відрізняється від відомих аналогів інноваційними рішеннями: запропонованими підходами до виявлення і очищення вхідних даних від аномальних вимірів; використанням нелінійних

за параметрами моделей в методах статистичного навчання; модифікацією згорткових нейронних мереж для обробки часових рядів із впровадженням нелінійної схеми компромісів в моделі нейронної структури та моделюванням навчальної вибірки для екстраполяції за методами статистичного навчання; визначенням інтегрованої оцінки активності інформаційного контенту за нелінійною схемою компромісів як результат багатфакторного аналізу вектору факторів за сферами діяльності соціуму; пошуком прогнозованих біфуркаційних інтервалів як критичної активності інформаційного контенту;

Поєднання в єдиному інформаційно-процесному середовищі запропонованих та відомих рішень забезпечує отримання синергетичного ефекту в підвищення достовірності (точності оцінювання і прогнозу) та повноти (множини врахованих факторів) визначення домінантного інформаційного контенту в глобальних інформаційних мережах.

Точність оцінювання та прогнозування часових рядів формату $\hat{F}_{j,area,tonality}$ – частотна тональність інформаційних повідомлень відстежувалась за метриками: середньоквадратична похибка – Mean Squared Error (MSE); коефіцієнт детермінації / ймовірність апроксимації – R-squared (R2) Score.

Узагальнені результати досліджень запропонованих рішень, порівняно із традиційними підходами зведені до табл. 3.15. Наведені результати отримані методами імітаційного моделювання.

Дослідження довели, що запропоновані рішення в обробці та прогнозування часових рядів забезпечують підвищення точностних характеристик.

Показники повноти врахування множини факторів при визначенні домінантного інформаційного контенту визначались за двома напрямками:

за методами факторного аналізу [28];

за методами теорії інформації – інформаційний підхід [41];

Факторний аналіз спрямований на врахуванні декомпозиційного поділу інформаційних повідомлень на множину із семи сфер діяльності соціуму (рис.2.1.). Для визначення показника інформативності за факторним підходом використано середньоквадратичні відхилення $\sigma_{\hat{F}_{area,tonality}}$ та дисперсії $\sigma_{\hat{F}_{area,tonality}}^2$ вимірів частотної тональності $\hat{F}_{j,area,tonality}$ (2.9), (2.10).

Таблиця 3.13

Узагальнені результати досліджень

Підхід	Етап		
	Препроцесінг	Оцінювання	Прогнозування
Традиційні рішення	MSE=6.640	-	-
Запропоновані рішення	MSE=2.950	-	-
Виграш	0.55	-	-
Традиційна поліноміальна модель N=10000	-	R2=0.963	MSE=8.928e-10
Нелінійна модель N=10000	-	R2=0.961	MSE=2.572e-13
Виграш	-	0.0022	3.47e3
Традиційна нейронна мережа N=248	-	MSE=14.352e-05 R2=0.999	MSE=2.489 R2=0.869
Оптимізаційна нейронна мережа N=248	-	MSE=3.416e-05 R2=0.999	MSE=0.733 R2=0.961
Виграш	-	MSE=0.76 R2=0	MSE=0.70 R2=0.10

Дисперсія вимірів, віднесена до обсягу вибірки вимірів $j = 1 \dots M$ дає похибку оцінювання параметрів $\hat{F}_{j,area,tonality}$. Зазначені параметри формують кореляційну матрицю похибок оцінювання частотної тональності інформаційних повідомлень для сфер діяльності соціуму:

$$R_F^{array} = \begin{bmatrix} \frac{\sigma_{\hat{F}_{1,tonality}}^2}{\sqrt{N}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{\sigma_{\hat{F}_{2,tonality}}^2}{\sqrt{N}} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\sigma_{\hat{F}_{3,tonality}}^2}{\sqrt{N}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\sigma_{\hat{F}_{4,tonality}}^2}{\sqrt{N}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{\sigma_{\hat{F}_{5,tonality}}^2}{\sqrt{N}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{\sigma_{\hat{F}_{6,tonality}}^2}{\sqrt{N}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{\sigma_{\hat{F}_{7,tonality}}^2}{\sqrt{N}} \end{bmatrix}, \quad (3.52)$$

$$R_F^{array} = diag \left[\frac{\sigma_{\hat{F}_{array,tonality}}^2}{\sqrt{N}} \right].$$

На відміну від (3.52) класичний однофакторний аналіз частотної тональності $\hat{F}_{j,tonality}$ в аналогічних позначеннях утворює похибку оцінювання:

$$R_F = \left[\frac{\sigma_{\hat{F}_{tonality}}^2}{\sqrt{N}} \right]. \quad (3.53)$$

До (3.52) та (3.53) застосовується розрахунок детермінанту, що і є індикатором, в якому враховано декомпозиційний поділ інформаційних повідомлень на множину із семи сфер діяльності соціуму (рис.2.1):

$$\det|R_F^{array}|, \det|R_F|. \quad (3.54)$$

Аналіз показників (3.54) здійснювався методами імітаційного моделювання з умовами: $N = 1000 \dots 10000$, $\sigma_{\hat{F}_{area,tonality}}^2 = 25$. Результати досліджень наведені в табл.3.14 та на рис.3.18.

Таблиця 3.14

Значення параметрів (3.54)

N	1000	2000	3000	4000	5000
$\det R_F^{array} $	0.193010	0.017059	0.004127	0.001507	0.000690
$\det R_F $	0.790569	0.559016	0.456435	0.395284	0.353553
Виграш	0.75589	0.96948	0.99095	0.99618	0.99804

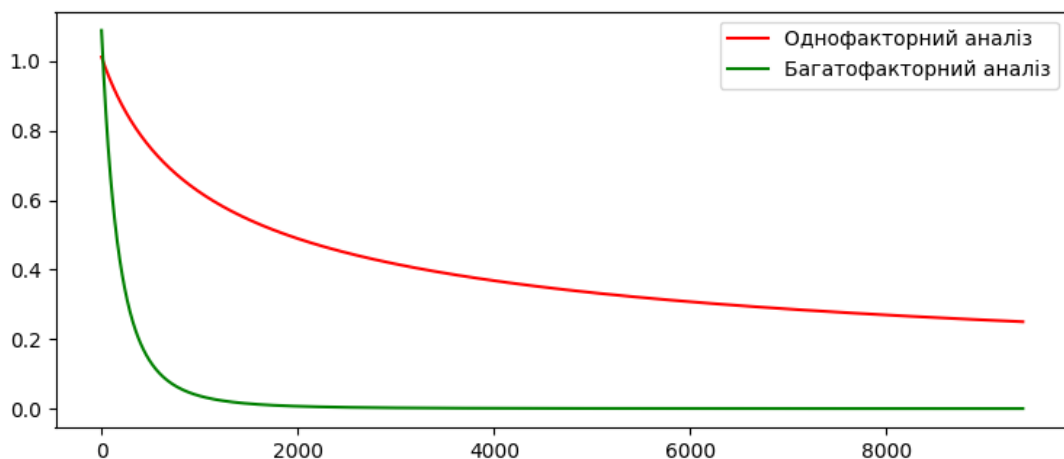


Рис.3.18. Графіки зміни параметрів (3.54) – вісь ординат від об'єму вибірки N – вісь абсцис.

Аналіз даних табл.3.14 та графіків рис.3.18 демонструють виграш у точності оцінювання частотної тональності інформаційних повідомлень

$\hat{F}_{j,area,tonality}$ у випадку багатофакторного аналізу порівняно із однофакторним. Тобто збільшення факторів для аналізу – повноти, природним чином збільшує кількість аналізованої інформації, що підвищує точність оцінювання. Зазначена закономірність є обумовленою та підтверджується базовими положеннями теорії ймовірностей (див. закон великих чисел [41]). Тож отримання позитивного ефекту від декомпозиції інформаційних повідомлень на множину сфер діяльності соціуму із подальшою їх агрегацією в інтегрованому показнику є обумовленим і виправданим рішенням.

Інформаційний підхід в оцінювання повноти врахування множини факторів при визначенні домінантного інформаційного контенту базується на головних положеннях теорії інформації. Це виявляється у зміні кількісного значення інформації в оцінюванні домінантного інформаційного контенту за рахунок збільшення аналізованих факторів – сфер діяльності соціуму.

За джерело інформації приймаємо вибірку частотної тональності інформаційних повідомлень для семи сфер діяльності соціуму $\hat{F}_{j,area,tonality}$ та вибірку частот тональності без сферичного поділу інформаційних повідомлень $\hat{F}_{j,tonality}$.

Джерела інформації характеризується ймовірностями $P(\hat{F}_{j,area,tonality})$ та $P(\hat{F}_{j,tonality})$. Для випадку, коли джерело це вибірка вимірів – реалізації випадкового процесу, за ймовірність джерела пропонується уявляти ймовірність потрапляння виміру до довірчої області випадкового закону розподілу вимірів. Наприклад, для нормального закону ймовірність 0.97 відповідає довірчому інтервалу $\pm 3\sigma_{\hat{F}_{j,area,tonality}}$ та $\pm 3\sigma_{\hat{F}_{j,tonality}}$.

Для уведених позначень, можемо встановити кількісні характеристики інформації $I(F)$ для n джерел за К. Шеноном [87]:

$$I(F) = - \sum_{i=1}^n \log P(F_i) \quad (3.55)$$

Логарифмічна функція $I(F)$, (3.55) забезпечує виконання очевидних властивостей кількісної міри інформації: якщо повідомлення відоме, тобто $P(F) = 1$, то кількість інформації в повідомленні $I(F) = \log P(F) = 0$. Збільшення кількості повідомлень призводить до збільшення кількості інформації.

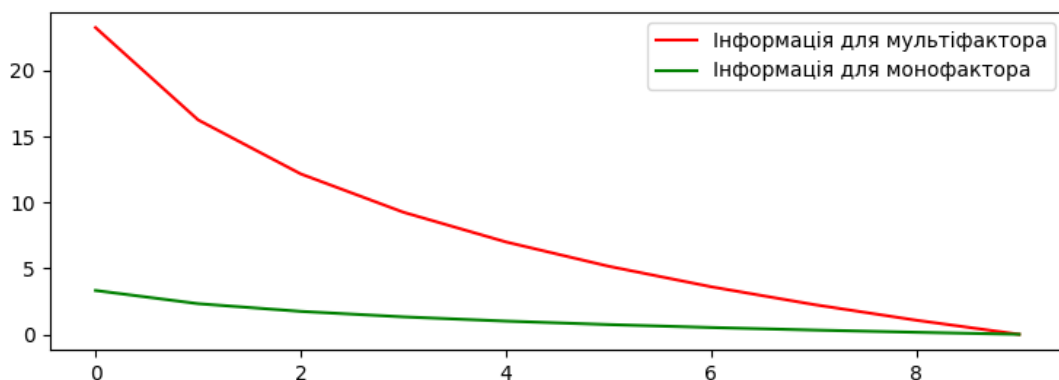
Дослідження зміни параметра (3.55), дають очевидні результати, подані в табл.3.15, 3.16 та на графіку 3.19, 3.20.

Таблиця 3.15

Розрахунок кількості інформації для різної $P(F)$

$P(F)$	0.01	0.1	0.2	0.4	0.6	0.8	0.9	1.0
$I(\hat{F}_{area})$	46.506	23.253	16.253	9.253	5.158	2.253	1.064	0
$I(F)$	6.643	3.321	2.321	1.321	0.736	0.321	0.152	0
Виграш	0.857	0.855	0.857	0.857	0.857	0.857	0.868	0

На графіках рис.3.19 позначено: вісь абсцис – зміна параметра $P(F)$ в цілих числах шкали 1-10; вісь ординат – параметри $I(\hat{F}_{area})$, $I(F)$.

Рис.3.19. Графіки зміни кількості інформації для різної $P(F)$.

Таблиця 3.16

Розрахунок кількості інформації для різної кількості джерел, $P(F)=0.1$

n	1	2	4	6	8	10
$I(\hat{F}_{area})$	3.321	6.643	13.287	19.931	26.575	33.219

На графіках рис.3.19 позначено: вісь абсцис – зміна параметра n ; вісь ординат – параметри $I(\hat{F}_{area})$.

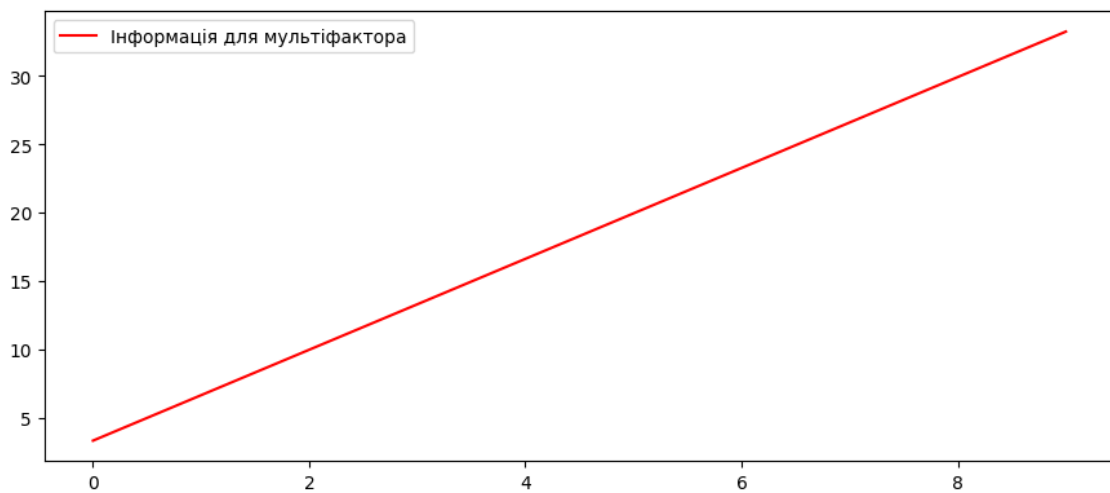


Рис.3.20. Графіки зміни кількості інформації для різної кількості джерел, $P(F)=0.1$

Наведені результати розрахунків доводять очевидний факт позитивного впливу у підвищення повноти за рахунок збільшення кількості факторів, що аналізуються.

Таким чином, цільову настанову досліджень з підвищення достовірності і повноти визначення домінантного інформаційного контенту в глобальних інформаційних мережах досягнуто.

Висновки до 3 розділу

1. У ході досліджень вперше запропоновано математичну модель багатофакторного прогнозування активності контенту в глобальних інформаційних мережах, яка базується на обробці та аналізі часових рядів з використанням методів статистичного та глибинного машинного навчання і відрізняється від відомих рішень: запропонованими підходами до виявлення і очищення вхідних даних від аномальних вимірів; використанням нелінійних за параметрами моделей в методах статистичного навчання; модифікацією згорткових нейронних мереж для обробки часових рядів із впровадженням нелінійної схеми компромісів в моделі нейронної структури та моделюванням навчальної вибірки для екстраполяції за методами статистичного навчання; визначенням інтегрованої оцінки активності інформаційного контенту за нелінійною схемою компромісів як результат багатофакторного аналізу вектору факторів за сферами діяльності соціуму; пошуком прогнозованих біфуркаційних інтервалів як критичної активності інформаційного контенту. Це забезпечує підвищення достовірності (ймовірності виявлення і точності прогнозу) домінантної активності контенту в глобальних інформаційних мережах.

2. Запропоновано декілька алгоритмів виявлення аномальних вимірів. Вони базуються на принципах ковзного вікна з реалізацією: заміни поточних вимірів математичним очікуванням; контролем зміни статистичних характеристик вимірів; аналізом значень похідних для виявлення аномальної динаміки часового ряду. Алгоритми є відносно простими для реалізації, що не обтяжує процес обробки часового ряду. Застосування запропонованих підходів забезпечує зниження похибки вхідних даних із аномаліями до 50%.

В роботі реалізовано розвиток методів статистичного навчання в напрямку застосування нелінійних за параметрами моделей. Це реалізовано з методом диференціально-нетейлрівських перетворень. Результатом є

підвищення прогностичних властивостей навчених за часовими рядами моделей.

Запропоновано модифікацію згорткових нейронних мереж для обробки часових рядів із впровадженням нелінійної схеми компромісів в моделі нейроної структури та моделюванням навчальної вибірки для екстраполяції за методами статистичного навчання. Дослідження довели підвищення точності прогнозування поведінки часових рядів поза межами інтервалу спостереження від 10% до 70%.

Для аналізу і прогнозування динаміки поведінки часових рядів частотної активності інформаційного контенту в різних сферах діяльності соціуму запропоновано порядок обрахунку інтегрованого показника. Це реалізовано за згортокою з нелінійною схемою компромісів як результат багатофакторного аналізу вектору факторів за сферами діяльності соціуму. Аналіз динаміки зміни інтегрованого показника з контролем фрактальної розмірності за індексом Херста дозволив реалізувати пошук прогнозованих біфуркаційних інтервалів як критичної активності інформаційного контенту.

3. Поєднання в єдиному інформаційно-процесному середовищі розробленої математичної моделі запропонованих та відомих рішень забезпечує отримання синергетичного ефекту в підвищення достовірності (точності оцінювання і прогнозу) та повноти (множини врахованих факторів) визначення домінантного інформаційного контенту в глобальних інформаційних мережах в середньому до 50% порівняно з відомими аналогами.

РОЗДІЛ 4.

ТЕХНОЛОГІЯ БАГАТОФАКТОРНОГО ПРОГНОЗУВАННЯ АКТИВНОСТІ КОНТЕНТУ В ГЛОБАЛЬНИХ ІНФОРМАЦІЙНИХ МЕРЕЖАХ

Технологія багатофакторного прогнозування активності контенту в глобальних інформаційних мережах базується і відрізняється впровадженням запропонованих математичних моделей кластеризації інформаційних повідомлень та багатофакторного прогнозування активності контенту на спеціалізованому програмно-апаратному комплексі, побудованому за принципами розподіленої комп'ютерної системи з відкритою архітектурою.

Використання розробленої технології забезпечує підвищення достовірності та повноти визначення домінантного інформаційного контенту в глобальних інформаційних мережах.

4.1. Технологія багатофакторного прогнозування активності контенту в глобальних інформаційних мережах

Суть розробленої технології ілюструється структурною схемою рис.4.1. Запропоноване архітектурне рішення є базовим модулем, який розгортається і реалізується в конкретиці складу компонент залежно від задач і практики його застосування. Наприклад для оперативного визначення соціальних настроїв різних сфер соціуму, фокус-групах, цільових аудиторій, що реалізуються: комерційними структурами в сфері надання послуг, реалізації продажів, тощо; в органах державного управління різної ланки – місцевого самоврядування, міністерств та відомств; в сфері безпеки і оборони.

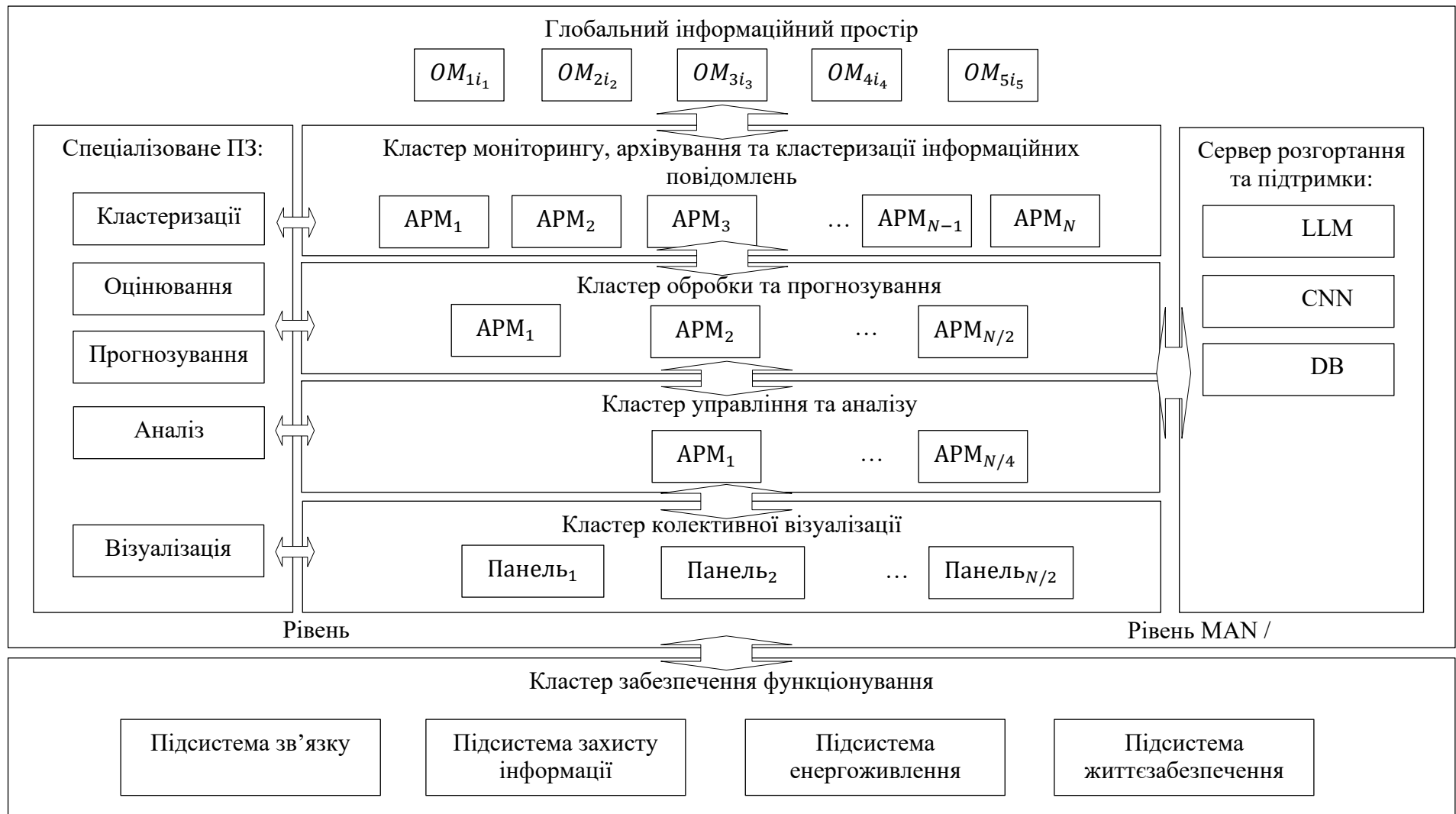


Рисунок 4.1. Структурна схема програмно-апаратного комплексу з реалізацією запропонованої технології

Зазначений спектр прикладних галузей застосування запропонованої технології забезпечується завдяки відкритій архітектурі програмно-апаратного комплексу, що не змінює, а лише масштабує запропоновані рішення.

Практична реалізація технології пропонується за принципами ситуаційних центрів. Варіант фізичної реалізації програмно-апаратного комплексу із впровадженням технології багатофакторного прогнозування активності контенту в глобальних інформаційних мережах представлений схемою рис.4.2.

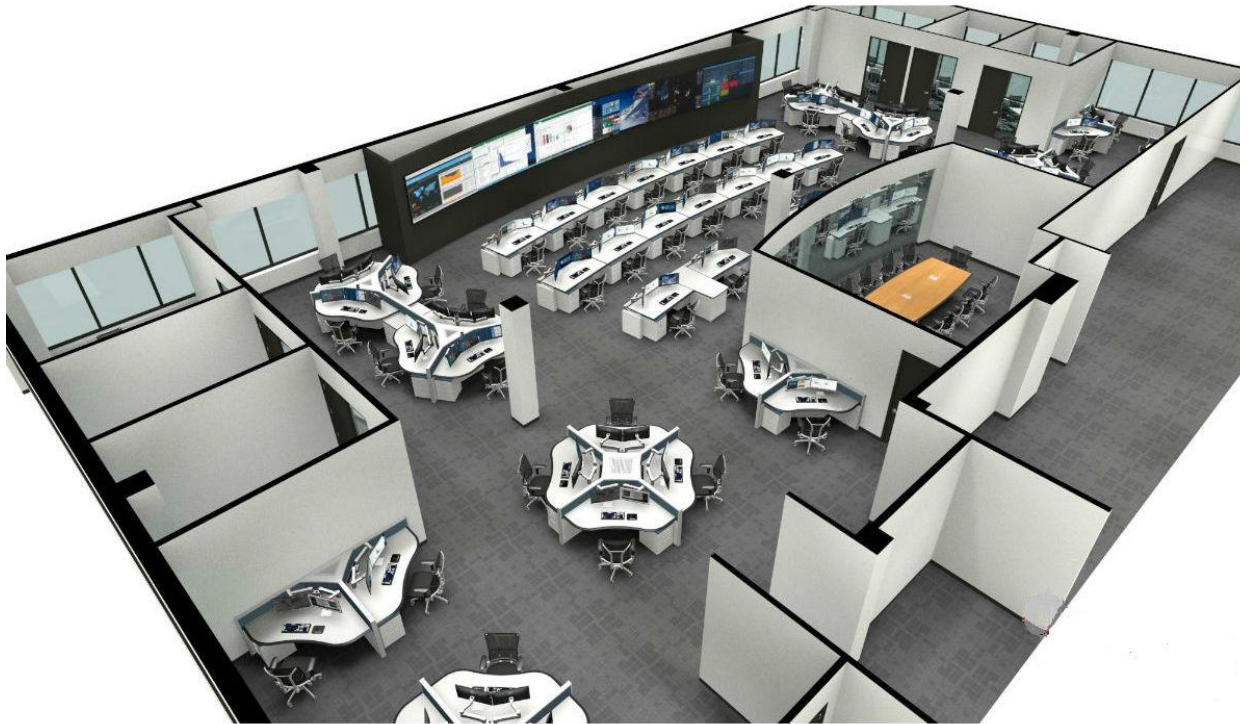


Рис.4.2. Варіант фізичної реалізації програмно-апаратного комплексу із технологією багатофакторного прогнозування активності контенту в глобальних інформаційних мережах

[\[https://constanttech.com/key-command-center-design-considerations/\]](https://constanttech.com/key-command-center-design-considerations/)

Технологія багатofакторного прогнозування активності контенту в глобальних інформаційних мережах складається з практичної реалізації етапів, запропонованих у ході досліджень:

математичної моделі кластеризації інформаційних повідомлень (див. структурну схему рис. 2.4);

математичної моделі багатofакторного прогнозування активності контенту в глобальних інформаційних мережах (див. структурну схему рис. 3.1).

У відповідності до запропонованих математичних моделей реалізовано програмні компоненти, що розгорнуті на спеціалізованому програмно-апаратному комплексі (рис.4.1), побудованому за принципами розподіленої комп'ютерної системи з відкритою архітектурою.

Запропоноване архітектурне рішення є базовим модулем, який розгортається і реалізується в конкретиці складу компонент залежно від задач і практики його застосування. Наприклад для оперативного визначення соціальних настроїв різних сфер соціуму, фокус-групах, цільових аудиторій, що реалізуються: комерційними структурами в сфері надання послуг, реалізації продажів, тощо; в органах державного управління різної ланки – місцевого самоврядування, міністерств та відомств; в сфері безпеки і оборони.

Програмно апаратна компонента інформаційно поєднується в єдину систему, що формується декількома кластерами.

Кластер моніторингу, архівування та кластеризації інформаційних повідомлень – являє собою множину автоматизованих робочих місць (АРМ), де реалізуються процеси збору первинної інформації та їх кластеризація. Кількісний склад кластера N АРМ, залежно від кількості ОМ та оперативності моніторингу і первинної кластеризації повідомлень.

Кластер обробки та прогнозування – представлено $N/2$ складом АРМ на яких реалізується головні процеси обробки інформації.

Кластер управління та аналізу ($N/4$ - АРМ) – реалізує: визначення ОМ, координація дій системи в цілому, глобальний аналіз результатів обробки, формування аналітичної звітності та вироблення рекомендацій з протидії виявленим критичним ситуаціям в інформаційному просторі.

Кластер колективної візуалізації – як атрибут ситуаційного центру являє собою інтерфейс інформаційних потоків у тому числі з візуалізацією результатів роботи системи.

Головними компонентами, що визначають функціонал та ефективність системи є спеціалізоване програмне забезпечення, з імплементацією розроблених у ході досліджень математичними моделями кластеризації інформаційних повідомлень та багатофакторного прогнозування активності контенту в глобальних інформаційних мережах.

Кластер забезпечення функціонування відображає практику організації роботи системи в реальних умовах для стаціонарного і мобільного виконання системи.

Вимоги до апаратної складової АРМ: процесор не гірше: 12-ядерний Intel Core Ultra 7 165U (1.7-4.9 ГГц); відеокарта, інтегрована із сопроцесором. наприклад типу Arc Graphics; накопичувачі даних: Обсяг SSD 2 ТБ.

Сервер розгортання та підтримки – передбачає зовнішнє розгортання на високопродуктивному кластері з можливістю зберігання та оброблення великих даних з технологіями великих мовних моделей.

Окремі АРМ та комунікація із зовнішнім сервером реалізовано з протокольним обміном розподіленої клієнт-серверної архітектури з використанням локальних та глобальних комп'ютерних мереж.

Окремі елементи програмного забезпечення для прискорення обчислень розгорнуті на мікрокомп'ютерному обладнанні, зокрема:

- алгоритми очищення від аномалій;
- визначення порядку поліноміальної моделі;
- метод найменших квадратів;
- алгоритми побудови нелінійних за параметрами моделей.

4.1.1. Апаратна реалізація кластеру обробки та прогнозування

Для апаратної підтримки процесу обробки та прогнозування активності інформаційного контенту та імплементації запропонованої технології для широкого кола споживачів проведені дослідження варіантів апаратної реалізації кластеру обробки.

Дослідження спрямовано на реалізацію моделі обробки інформації (див. структуру рис.3.1) з фокусом на обробку часових рядів з оптимізаційною нейронною мережею за процесами рис. 3.10. Саме імплементації нейронної мережі для обробки Big Data дає пікове навантаження на апаратну частину кластеру обробки та прогнозування.

Метою дослідження є тестування швидкості обробки часових рядів оптимізаційною нейронною мережею на етапі навчання і прогнозування з використанням різних підходів до побудови кластеру обробки та прогнозування, а саме:

- доступність комплектуючих;
- швидкість обробки даних;
- енергоефективність.

Для побудови кластеру обробки та прогнозування на основі доступних комплектуючих перспективним є використання процесорів з x86-64 подібною архітектурою, для максимальної швидкості обробки даних – апаратні прискорювачі, для енергоефективних рішень – процесори з архітектурою ARM.

Для досліджень обрано 3 апаратні реалізації:

- процесор з x86-64 подібною архітектурою – AMD Ryzen 9 9900X;
- графічний прискорювач з підтримкою архітектури CUDA – NVIDIA RTX 5070 Ti;
- процесор з ARM архітектурою - Broadcom BCM2712.

Сучасні процесори AMD з архітектурою Zen 5 широко застосовуються в обчислювальних системах та підходять для використанні у складі кластеру обробки та прогнозування. Процесори з архітектурою Zen 5 базуються на наборі інструкцій x86-64, що дозволяє використати подібний процесор для порівняння швидкості обчислення. Недоліком таких рішень є низька енергоефективність та доволі висока вартість.

Графічні прискорювачі NVIDIA є основними для вирішення задач з підвищення швидкості математичних обчислень. Особливістю прискорювачів NVIDIA з підтримкою програмно-апаратної архітектури CUDA є можливість розподілення великої кількості однакових математичних операцій на велику кількість обчислювальних ядер CUDA, що дозволяє суттєво скоротити час виконання обчислень. Оскільки обчислення необхідні для реалізації нейронної мережі є типовими для технології пришвидшення обчислень CUDA, графічні прискорювачі Nvidia можуть бути використані у складі кластеру обробки та прогнозування. Для порівняння швидкості обчислення було обрано графічний прискорювач NVIDIA RTX 5070 Ti. Недоліками рішень на основі прискорювачів обчислень є їх дуже висока вартість.

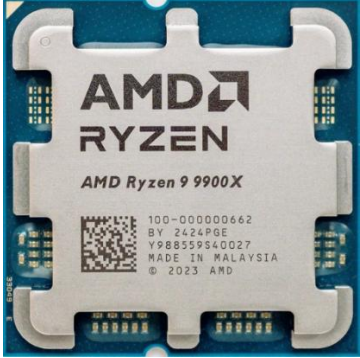
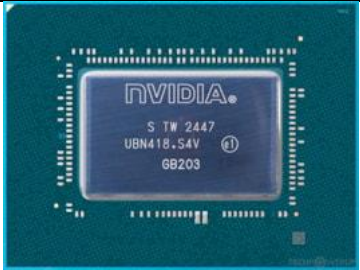
Серед енергоефективних рішень перевагу мають процесори з архітектурою ARM, через свою малу споживану потужність відносно обчислювальної спроможності. Процесори з архітектурою ARM широко застосовуються у компактних пристроях,. Використання подібних процесорів для побудови кластеру обробки та прогнозування є доцільним у випадку відсутності потреби проведення обчислень у реальному часі та використовуючи малі інтервали для обробки та прогнозування. Недолік – мала потужність, відповідно є суттєве обмеження у кількості даних для обробки

Для проведення дослідження використовуючи енергоефективну апаратну частину було обрано одноплатний комп'ютер Raspberry Pi 5 з процесором Broadcom BCM2712.

Технічні характеристики розглянутих пристроїв – варіантів апаратної реалізації кластеру обробки та прогнозування наведено в табл.4.1.

Таблиця 4.1

Характеристики апаратної реалізації кластеру обробки та прогнозування

Апаратна платформа	Характеристики	Зовнішній вигляд
Процесор AMD Ryzen 9 9900X	Архітектура: x86-64 подібна; Кількість ядер: 12; Тактова частота: 4.4 ГГц.	
Графічний прискорювач NVIDIA RTX 5070 Ti	Архітектура: Blackwell 2.0; Підтримує програмно-апаратну архітектуру CUDA; Кількість ядер CUDA: 8960; Тактова частота: 2295 МГц.	
Графічний процесор Broadcom BCM2712	Архітектура: ARM Cortex-A76; Кількість ядер: 4; Тактова частота: 2.4 ГГц.	

Дослідження проводились із реалізацією процесів навчання оптимізаційної нейронної мережі та подальшого прогнозування на часовому ряді з 1520 елементів. Архітектура мережі відповідає характеристикам, зазначеним в розділі. 3.3.

Результати досліджень подано в табл.4.2, 4.3. В таблицях наведено час, що витрачено на тренування та прогнозування значень часових рядів з використанням оптимізаційної нейронної мережі для різних апаратних платформ. Дослідження проводилось шляхом вимірювання часу прогнозування кожного наступного елементу числового ряду з дискретністю 50-ти елементів.

Таблиця 4.2

Результати тестування швидкості тренування нейронної мережі

Апаратна частина	Час тренування, с
Процесор з x86-64 архітектурою	28,09134
Графічний прискорювач	9,51222
Процесор з ARM архітектурою	571,82862

Таблиця 4.3

Результати тестування швидкості прогнозування нейронної мережі

Апаратна частина	Середній час прогнозування, с
Процесор з x86-64 архітектурою	0,096613
Графічний прискорювач	0,0459
Процесор з ARM архітектурою	2,125715

Отримані результати дозволяють сформулювати такі висновки. Рішення на основі x86-64 процесорів не є енергоефективними показуючи середній результат. Рішення на основі прискорювачів обчислень є найшвидшим рішенням, проте найдорожчим. Рішення на основі процесорів з архітектурою ARM є найповільнішими, проте дешеві та енергоефективні.

4.2. Розробка макету програмно-апаратного комплексу прогнозування активності контенту в глобальних інформаційних мережах та оцінювання ефективності запропонованих рішень

Оцінювання ефективності запропонованих рішень здійснювалось в комплексному їх використанні для конкретних задач практики. Для цього розроблено демонстраційний прототип запропонованої технології. Код програмної складової наведено у додатку.

Приклади практичного застосування розробленої технології підтверджено актами впровадження (див. додатки).

В якості апаратної складової застосовувалось два автоматизовані робочих місця. Імітація роботи серверу здійснювалась як стаціонарна комп'ютерна система. АРМ обробки мали локальну бездротову мережу. Доступ до сервера та до глобальних інформаційних мереж здійснювався засобами дротового та бездротового зв'язку. Макет, на якому розгорталось програмне забезпечення представлений для фізичного моделювання та практичної апробації рішень подано на фото рис.4.3.

Джерелами інформаційних повідомлень виступали ОМ, що репрезентують перелік, наведений в табл.2.1. В результаті парсингу ОМ отримано кластер вхідних повідомлень, що міститься в базі даних.

Первинні повідомлення підлягали кластеризації за сферами діяльності соціуму (рис. 2.1) та за тональністю відповідно до етапів запропонованої математичної моделі кластеризації інформаційних повідомлень.

Зазначені процеси реалізовано на кластері моніторингу, архівування та кластеризації інформаційних повідомлень (див. схему рис.4.1 та автоматизовані робочі місця макету, зображені на рис.4.3-4.6).



Рис.4.3. Макет апаратно-програмного комплексу багатфакторного прогнозування активності контенту в глобальних інформаційних мережах

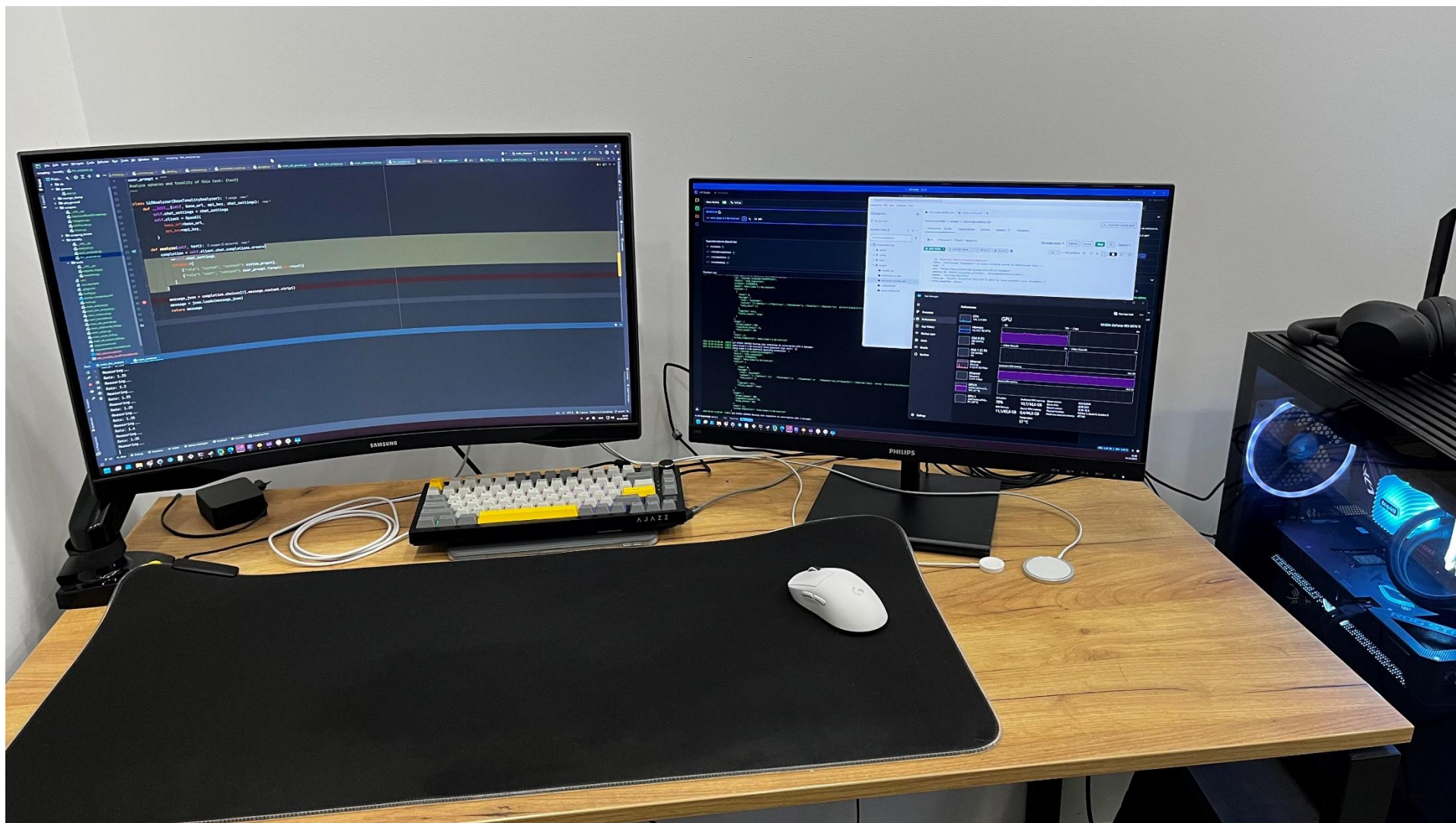
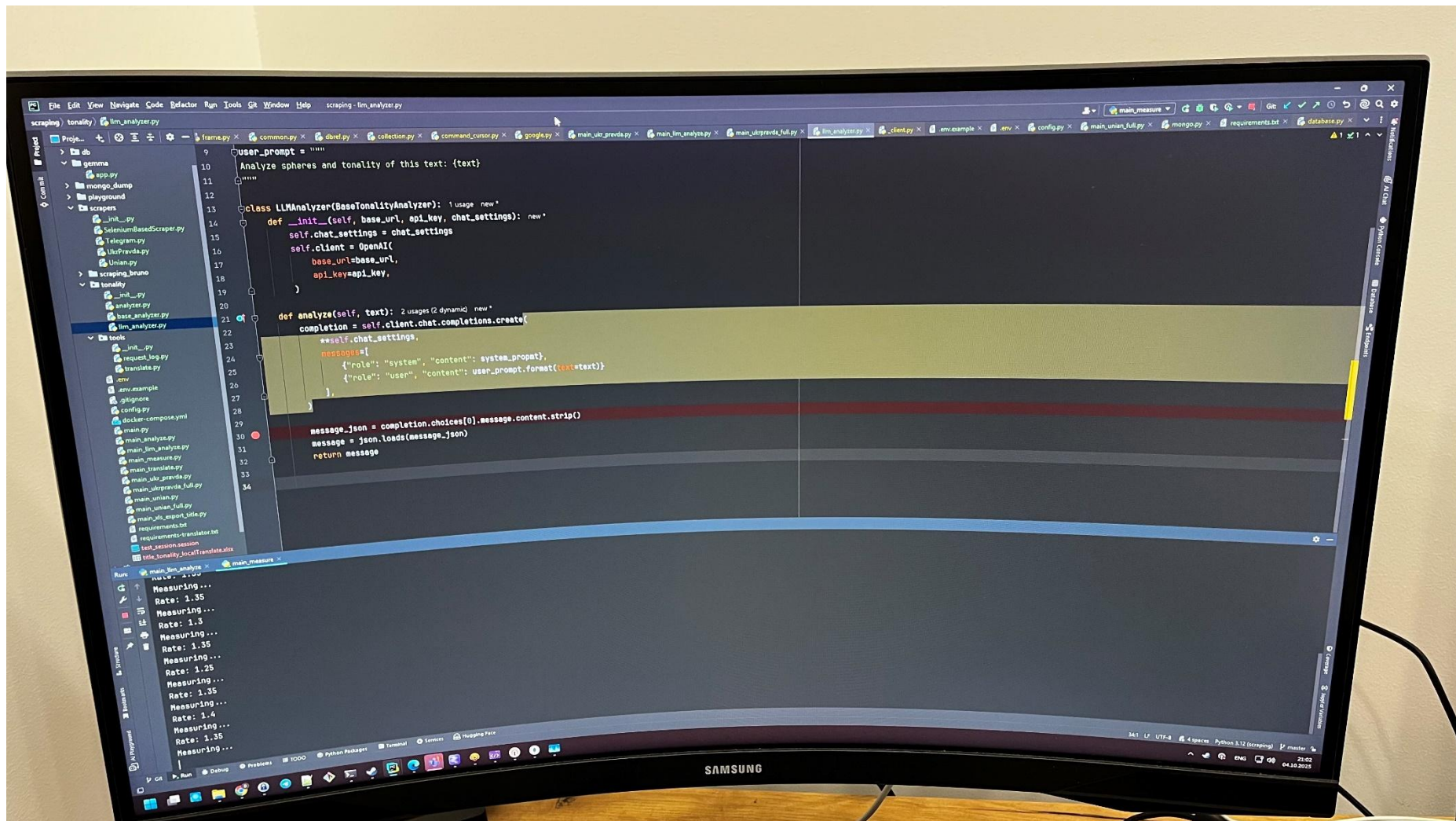
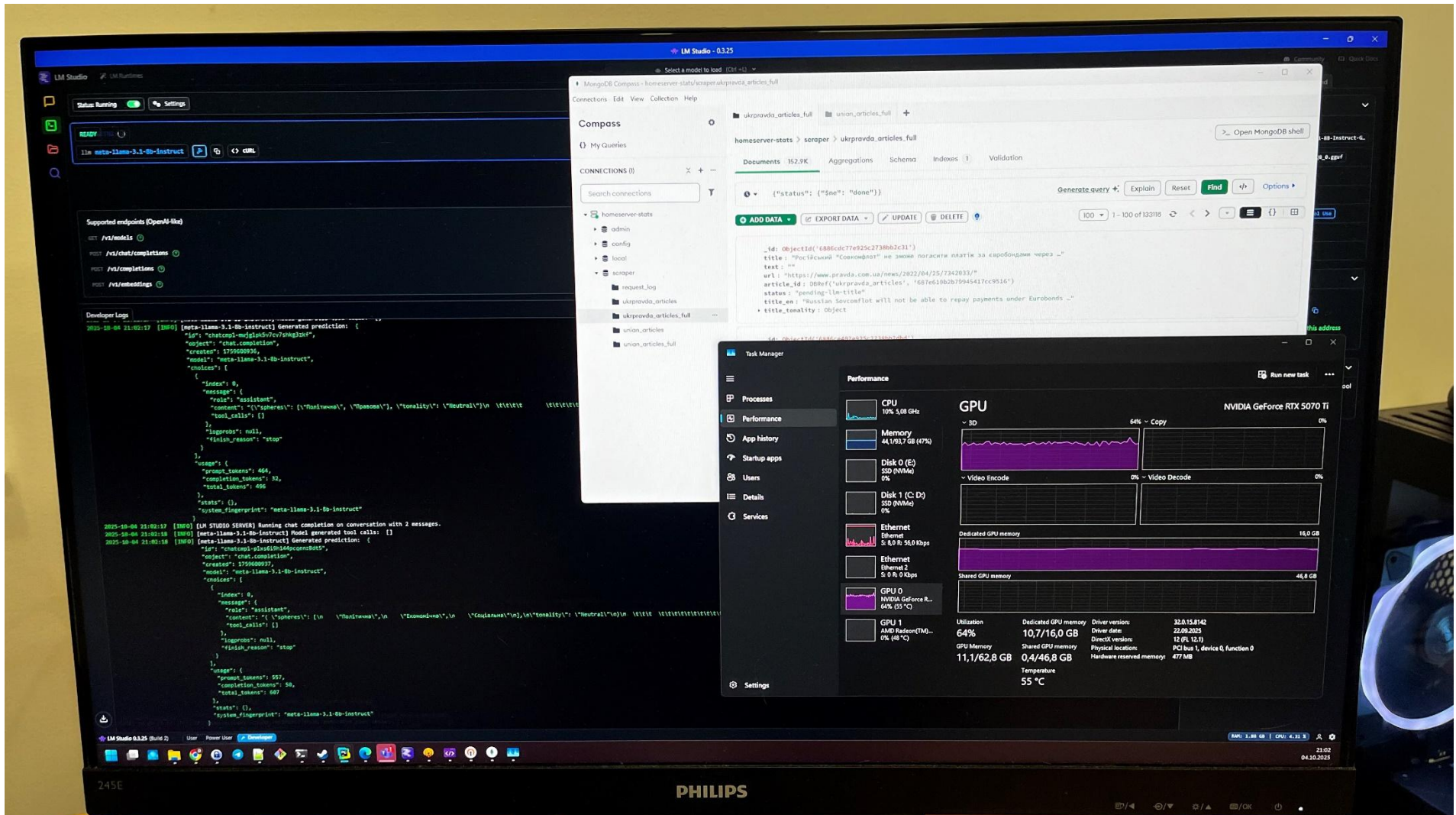


Рис.4.4. Макет серверного оборудования



a)

Рис.4.5. Програмне забезпечення серверного обладнання



б)

Рис.4.5. Програмне забезпечення серверного обладнання

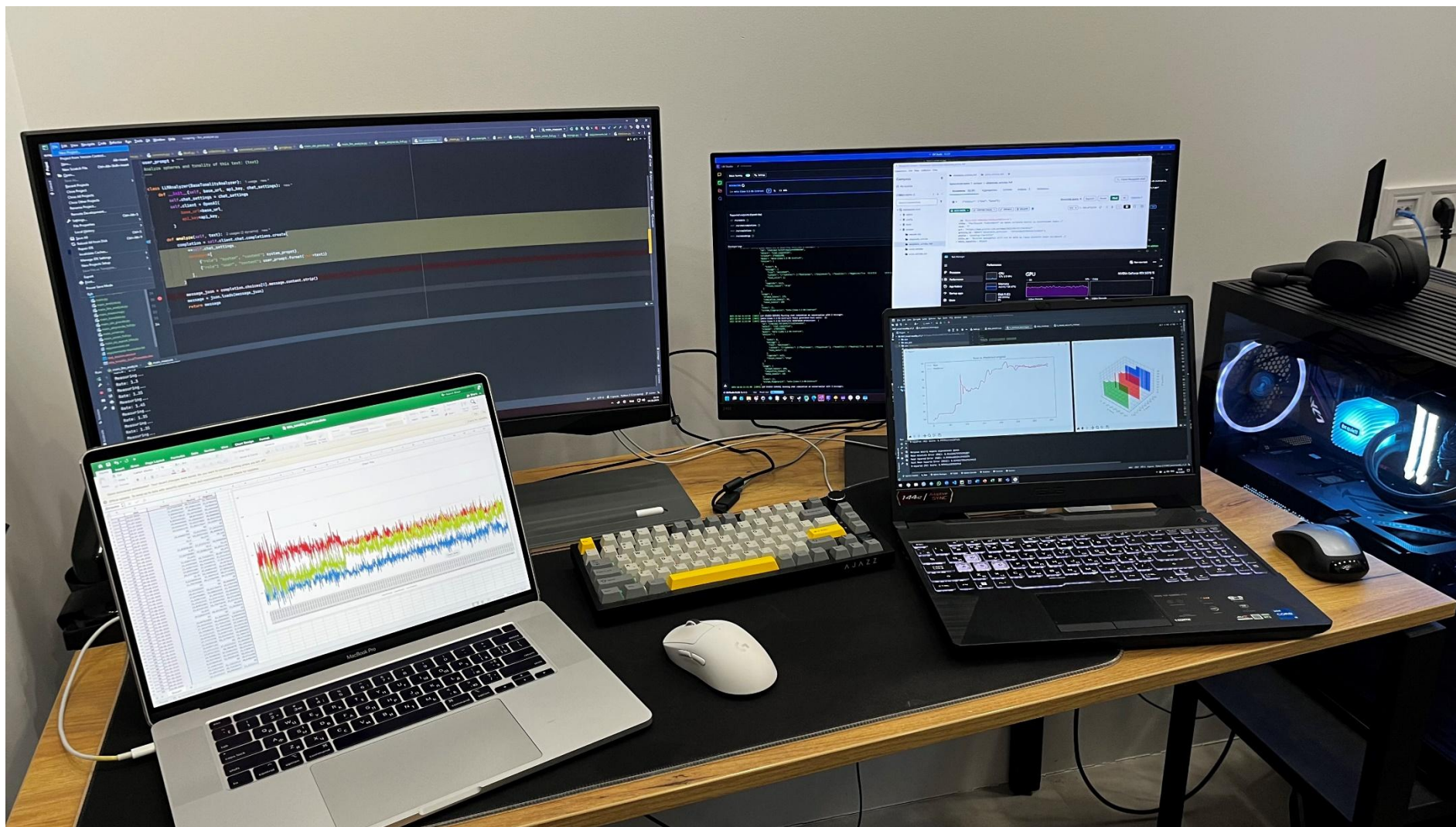


Рис.4.6. Макет автоматизованих робочих місць із програмним забезпеченням

Результатом моніторингу ОМ та кластеризації зібраних інформаційних повідомлень є часові ряди, що утворюють датасет за період 2020-2025 роки. Формат даних відповідає структурі $\hat{F}_{j,area,tonality}$ (2.10). Вхідні дані систематизовані у файлову структуру сферичної тональності (рис.2.2), що подана на рис.4.7.

	A	B	C	D	E
1		date	Positive	Neutral	Negative
2	0	19.07.2020	0	0	100
3	1	20.07.2020	0	100	0
4	2	21.07.2020	0	33,3333333	66,66666667
5	3	22.07.2020			
6	4	23.07.2020			
7	5	24.07.2020			
8	6	27.07.2020			
9	7	29.07.2020			
10	8	30.07.2020			
11	9	31.07.2020			
12	10	01.08.2020	66,66666667		
13	11	03.08.2020	33,3333333		
14	12	04.08.2020	0	33,3333333	66,66666667
15	13	05.08.2020	0	50	50
16	14	06.08.2020	0	100	0
17	15	08.08.2020	0	0	100
18	16	11.08.2020	0	100	0
19	17	12.08.2020	50	0	50
20	18	13.08.2020	0	50	50
21	19	14.08.2020	0	100	0

	A	B	C	D	E
1		date	Positive	Neutral	Negative
2	0	19.07.2020	0	0	100
3	1	20.07.2020	0	100	0
4	2	21.07.2020	0	33,3333333	66,66666667
Безпека.xlsx Духовна.xlsx Економічна.xlsx Політична.xlsx Права і свободи особистості.xlsx Правова.xlsx Соціальна.xlsx 09.10.2025 2:58 09.10.2025 2:58 09.10.2025 2:58 09.10.2025 2:58 09.10.2025 2:58 09.10.2025 2:58 09.10.2025 2:58 Аркуш Microsoft ... Аркуш Microsoft ... Аркуш Microsoft ... Аркуш Microsoft ... Аркуш Microsoft ... Аркуш Microsoft ... Аркуш Microsoft ... 55 КБ 52 КБ 70 КБ 79 КБ 26 КБ 63 КБ 75 КБ					
14	12	04.08.2020	0	33,3333333	66,66666667
15	13	05.08.2020	0	50	50
16	14	06.08.2020	0	100	0
17	15	08.08.2020	0	0	100
18	16	11.08.2020	0	100	0
19	17	12.08.2020	50	0	50
20	18	13.08.2020	0	50	50
21	19	14.08.2020	0	100	0

Рис.4.7. Файлова структура вхідних даних.

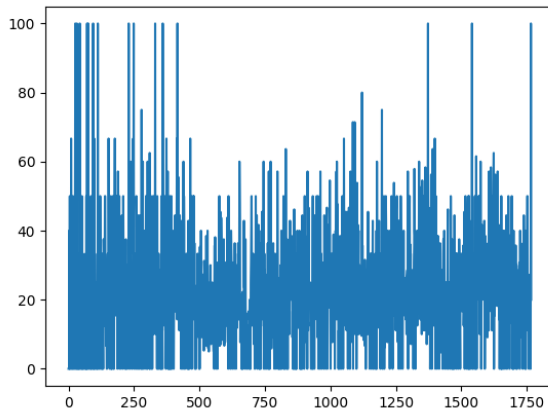
Результати кластеризації, представлені в форматі вхідних даних, що зберігаються у файлах, структура яких подана на рис.4.7 є вхідною інформацією для подальших етапів обробки відповідно до схеми рис.3.1 (етапів математичної моделі багатфакторного прогнозування активності) та технологічних дій структурної схеми рис.4.1. Для цього активізується функціонал кластеру обробки та прогнозування.

1. Підготовчий етап з аналізом та препроцесінгом даних.

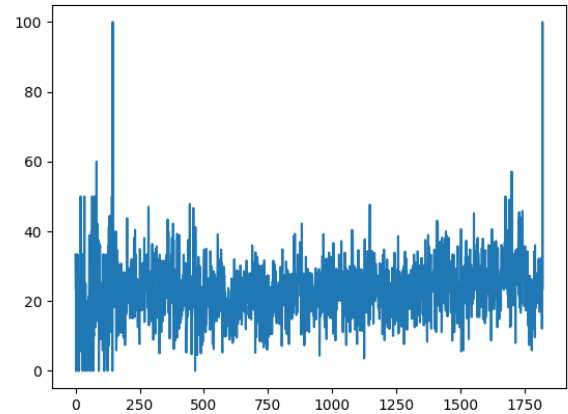
Виявлення властивостей вхідних даних реалізовано методами візуального контролю з розрахунком і аналізом об'єктивних показників – статистичних характеристик часових рядів (див. рис.4.7).

Результати візуалізації та розрахунків подано на графіках рис.4.7 та в табл.4.1.

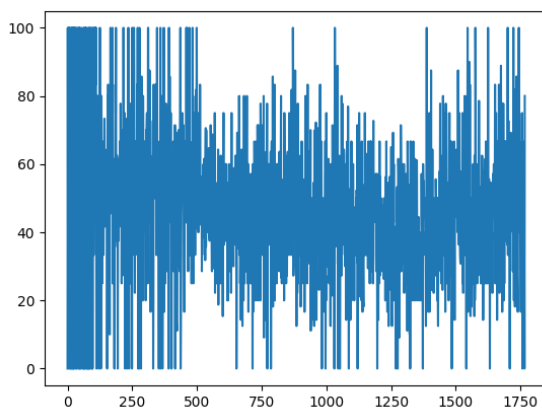
Кластер «Безпека.xlsx» та «Економічна.xlsx» - для трьох тональностей:
Positive, Neutral, Negative відображає типову динаміку часових рядів, рис. 4.8.



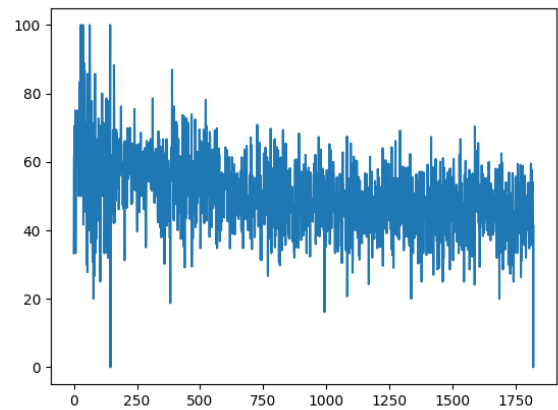
a) Positive



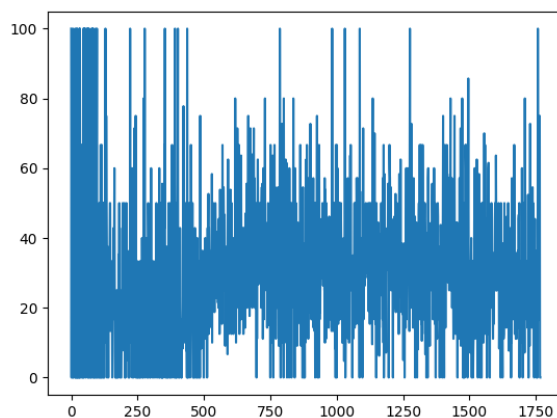
a) Positive



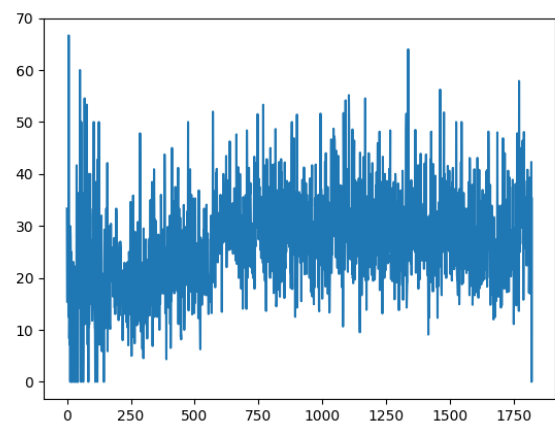
б) Neutral



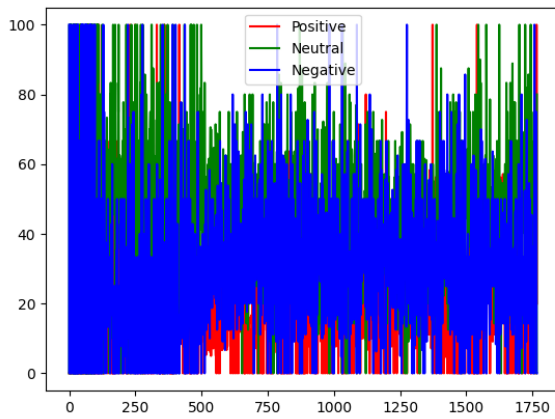
б) Neutral



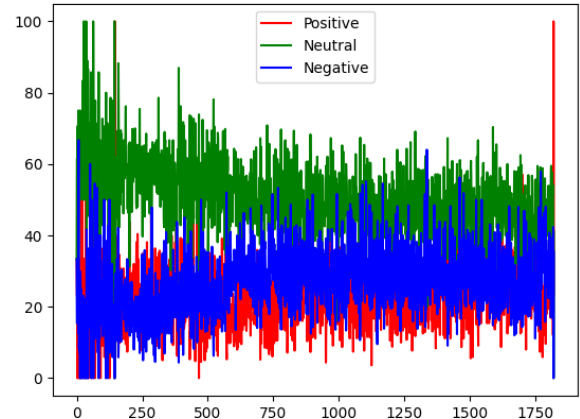
в) Negative



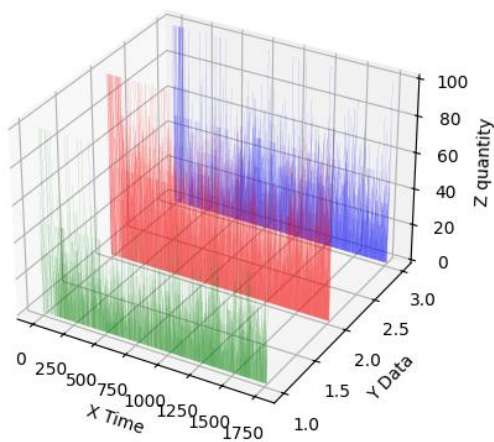
в) Negative



г) Positive, Neutral, Negative

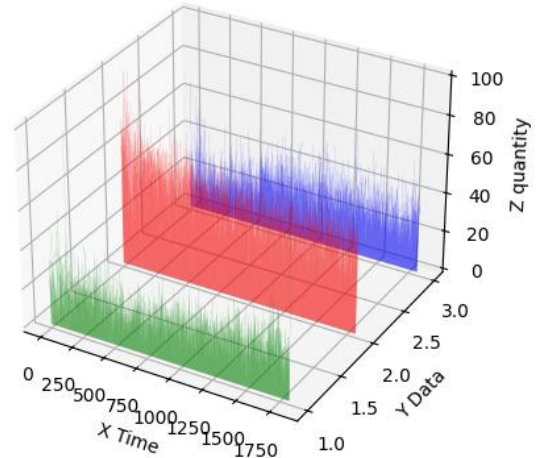


г) Positive, Neutral, Negative



д) Positive, Neutral, Negative

«Безпека.xlsx»



д) Positive, Neutral, Negative

«Економічна.xlsx»

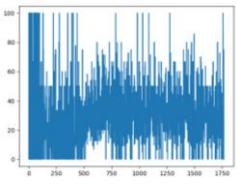
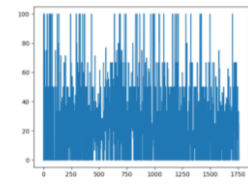
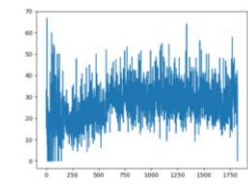
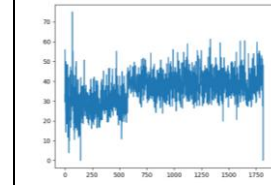
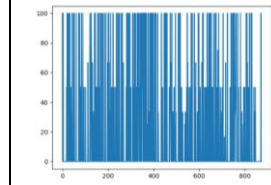
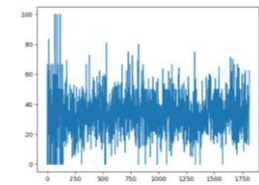
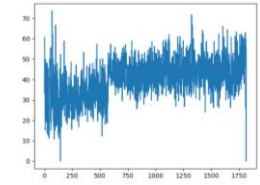
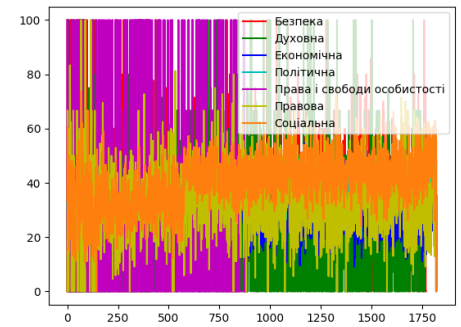
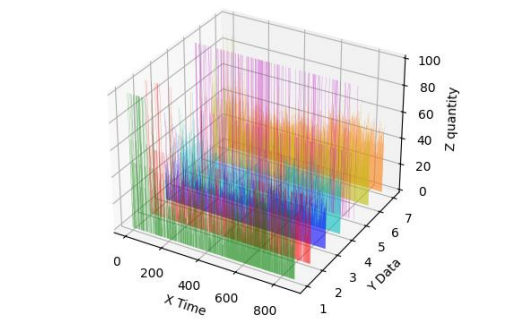
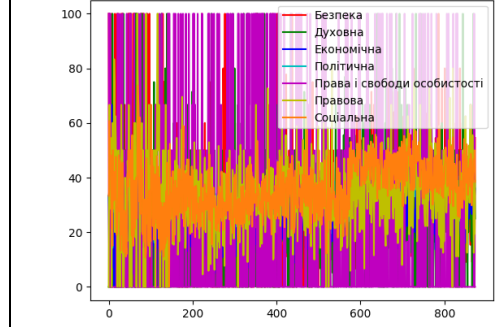
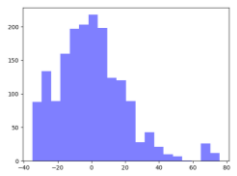
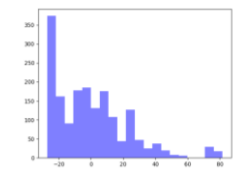
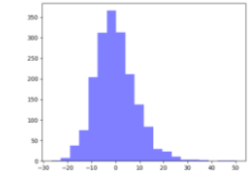
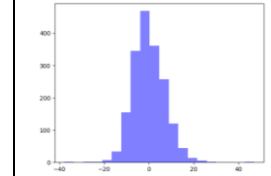
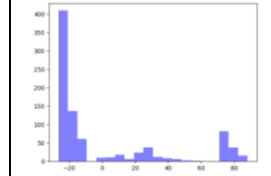
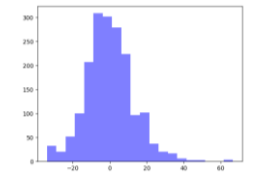
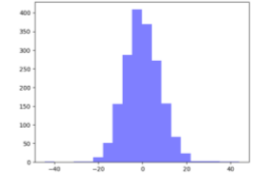
Рис. 4.8. Графіки вхідних даних

Результати статистичного аналізу вхідних даних для усіх сфер за тональністю «Negative» надано в узагальненій табл.4.1. Для даних з іншими тональностями «Positive», «Neutral» властивості часових рядів, що їх репрезентують аналогічним чином.

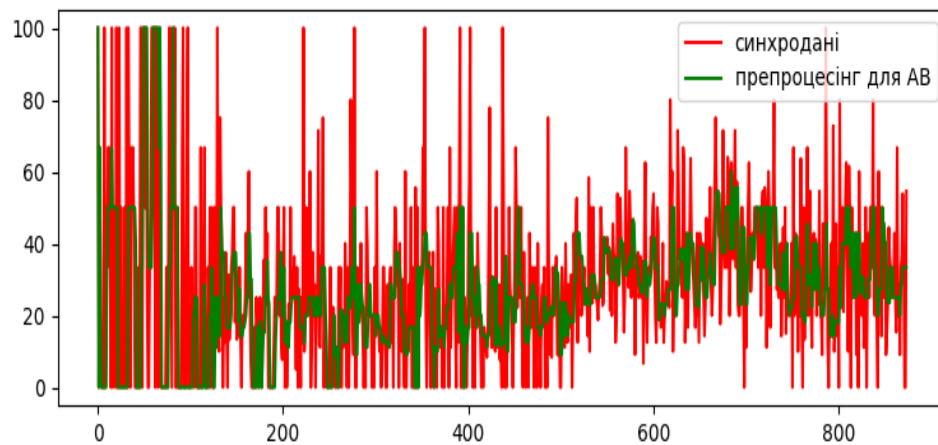
Результати статистичного аналізу вхідних даних для семи сфер діяльності соціуму за трьома тональностями (див. дані табл. 4.1) дозволяють виявити їх наступні властивості. Вхідні дані мають тренд, забруджений випадковим шумом з нормальним та експоненціальним законом розподілу. Вхідні дані містять аномалії типу грубі виміри та пропуски, що позначені нульовими значеннями.

Таблиця 4.1

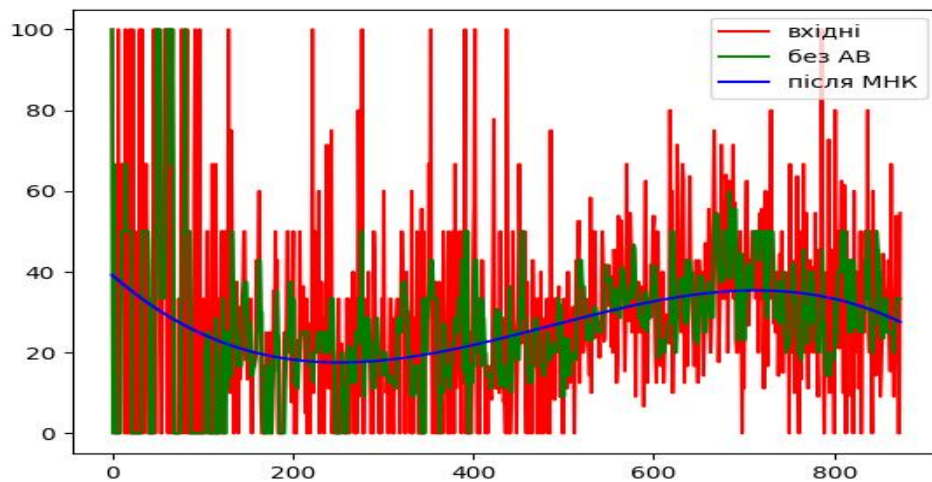
Результати статистичного аналізу вхідних даних для семи сфер діяльності соціуму

	Безпека.xlsx	Духовна.xlsx	Економічна.xlsx	Політична.xlsx	Права і свободи особистості.xlsx	Правова.xlsx	Соціальна.xlsx
Вхідні							
Композитне	 Оригінал 2d		 Синхронізовані 3d		 Синхронізовані 2d		Максимальний розмір: 1820 Мінімальний розмір: 874 Масиви синхронізовано до: 874
Закон повної	 n: 1768 m: 1.1917e-12 D: 417.1047 MSE: 20.4231	 n: 1762 m: 8.0582e-13 D: 527.2204 MSE: 22.9612	 n: 1820 m: -2.2131e-12 D: 71.8377 MSE: 8.4757	 n: 1819 m: -6.474e-13 D: 50.2589 MSE: 7.0893	 n: 874 m: 1.5585e-12 D: 1358.4683 MSE: 36.8574	 n: 1815 m: 4.1325e-14 D: 168.4141 MSE: 12.9774	 n: 1819 m: -6.0726e-13 D: 66.2683 MSE: 8.1405

Препроцесінг даних полягає в очищення від аномалій. Очищення часового ряду від аномалій типу грубі виміри реалізується з алгоритмом, що базуються на реалізації ковзного вікна – із заміною вимірів оцінками математичного очікування. Пропуски даних відновлюються за алгоритмом МНК. Результати подано на рис.4.9 (приклад очищення часового ряду «Безпека») та 3.10 (очищення від АВ типу «грубі виміри» для часових рядів усіх семи груп).



а) очищення від АВ типу «грубі виміри»



б) очищення від АВ типу «пропуски»

Рис.4.9. Препроцесінг вхідних даних для часового ряду «Безпека»

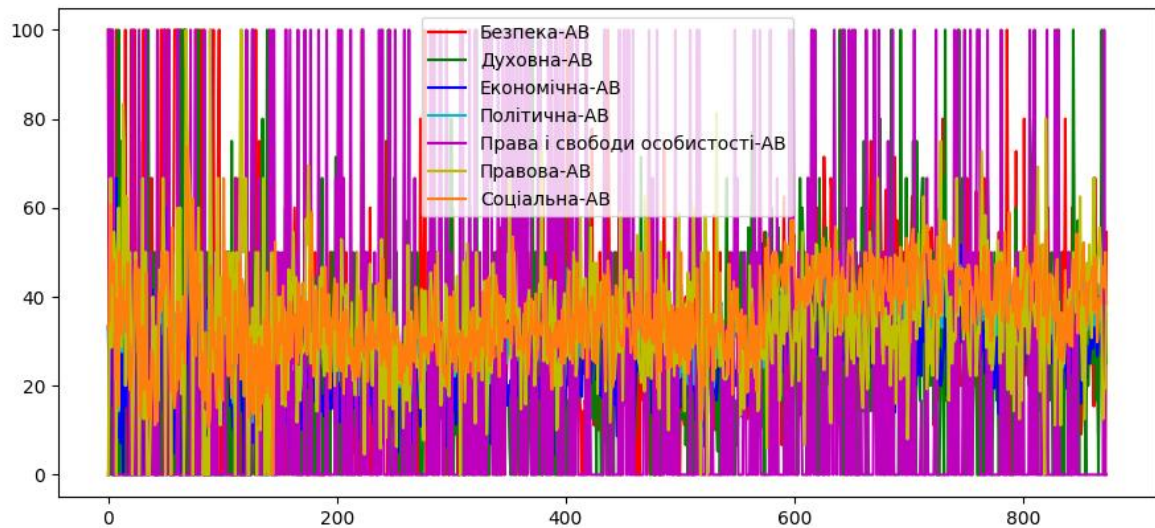


Рис.4.10. Очищення від АВ типу «грубі виміри» для семи груп

Структурна оптимізація моделі за швидким (в межах одного проходу верхівки часового ряду) алгоритмом контролю значень похідних високих порядків визначає третій оптимальний ступінь поліному. Порівняльна характеристика ймовірності апроксимації для оптимального $n=3$ та для $n=5$ порядків полінома наведена в табл.4.2

Таблиця 4.2

Значення ймовірності апроксимація R^2

Безпека	Духовна	Економічна	Політична	Права і свободи особистості	Правова	Соціальна
$n=3$						
0.1493	0.1046	0.4383	0.4317	0.0233	0.0595	0.4305
$n=5$						
0.1502	0.1046	0.4481	0.4519	0.0600	0.0692	0.4579

Дослідження показали незначне збільшення показника R^2 для $n=5$, що підтверджує рішення про $n=3$. Загальний відносно низький рівень параметра R^2 пояснюється специфікою вхідних даних, забруднених доволі великим за дисперсією випадковим шумом та пропусками (див. дані табл.4.1).

2. Обробка та прогнозування часових рядів.

Обробка вхідних даних за сферами діяльності соціуму спрямовано на виділення тренду та реалізовано методами статистичного навчання, відповідно до модифікацій, запропонованих в розділі 3 Результати подано на графіках рис.

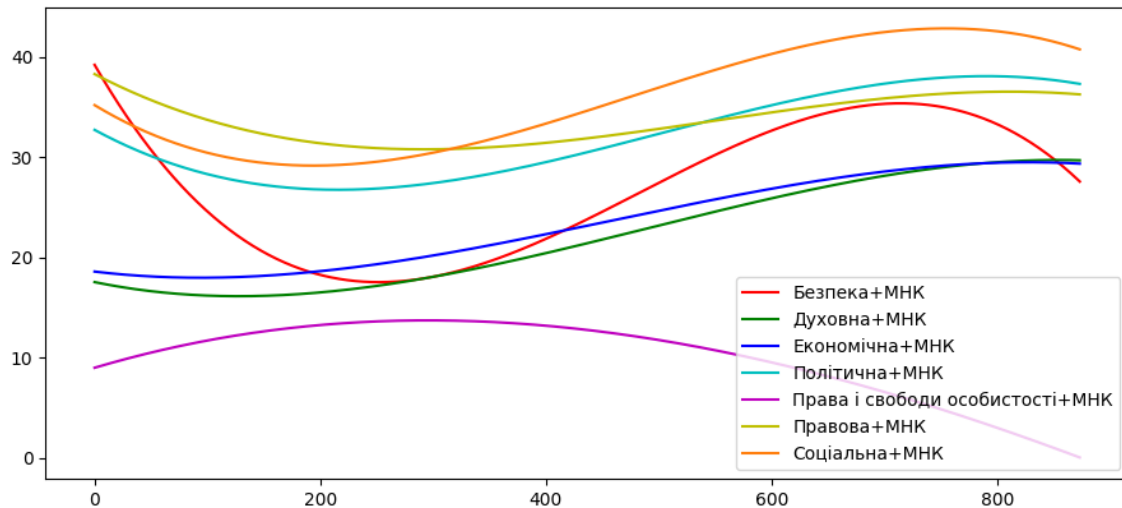


Рис.4.11. Графіки результатів обробки вхідних даних

3. Розрахунок інтегрованої оцінки для багатфакторного прогнозування активності контенту в глобальних інформаційних мережах з пошуком біфуркаційних інтервалів дозволив отримати результати, подані на графіку рис.4.12. На графіку рис.4.13 подано також результат застосування МНК з модифікаціями для виділення тренду з інтегрованого показника для прогнозування активності інформаційного контенту.

На графік рис.4.13 накладені інтервальні ознаки, що відповідають біфуркаційному інтервалу, пікової активності інформаційного контенту в глобальних інформаційних мережах та ділянка зниженні інформаційної активності (див. рис.1.3, 1.5).

Інтегрований показник отримано за реальними даними, що репрезентують датасет за період 2020-2025 роки. Графік інтегрованого показника має динаміку зміни та яскраво виражений пік подібно до класичних процесів зміни активності інформаційного контенту (рис.1.3).

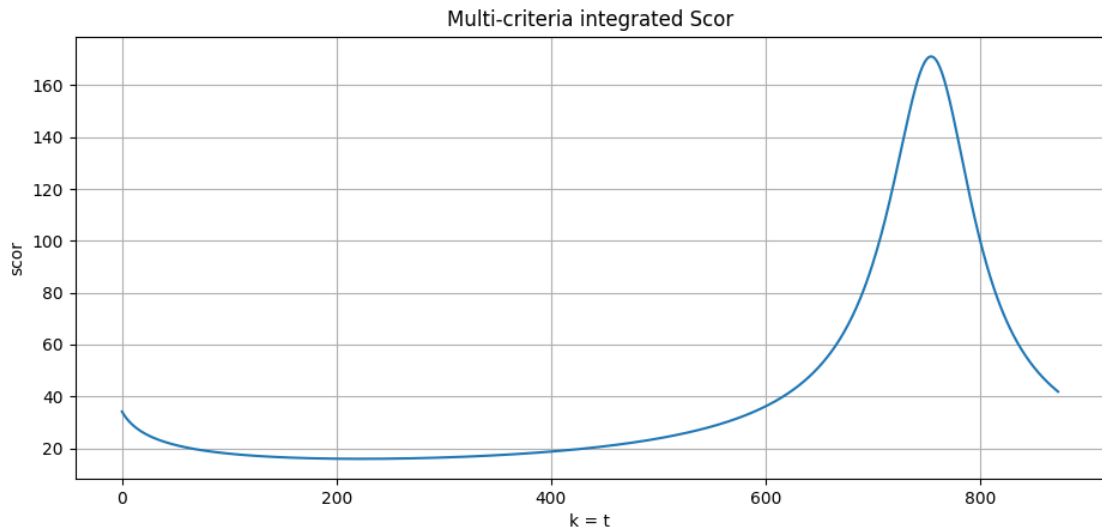


Рис.4.12. Графік зміни інтегрованого показника

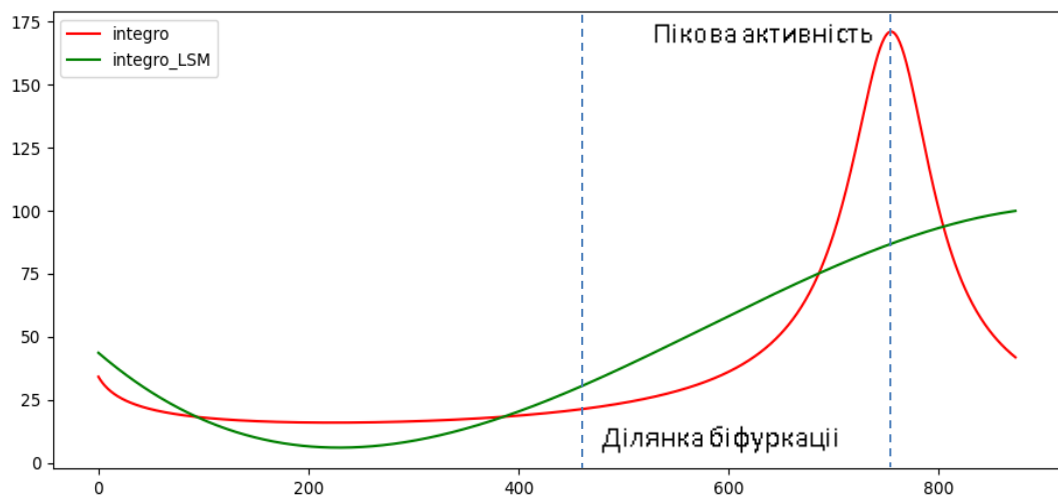


Рис.4.13. Графік зміни інтегрованого показника

Тож точки біфуркації не спостерігається та зміна катастрофічних наслідків інформаційної активності в семи головних сферах діяльності соціуму – не зафіксовано (див. рис.1.5).

Інтерпретація отриманих результатів і прив'язка їх до астрономічного часу дає часові координати: початок біфуркаційного інтервалу / експоненціального зростання – 03.2022; точка пікової активності – 08. 2022. Зазначені дати, з огляду на реальні події дозволяють верифікувати результати розрахунків як такі, що відповідають дійсності.

Результати розрахунку індексу Херста подано в табл.

Таблиця 4.3

Значення індексу Херста

Безпека	Духовна	Економічна	Політична	Права і свободи особистості	Правова	Соціальна	Інтегроскор
0.256	0.447	0.452	0.331	0.701	0.324	0.319	0.425

Індекс Херста фіксує зміну в подальшому тенденцій тренду інтегрованого показника.

Таким чином, результати досліджень та розрахунки на реальних даних, отриманих шляхом практичної реалізації технології багатofакторного прогнозування активності контенту в глобальних інформаційних мережах є дієвою та придатною до використання.

Висновки до 4 розділу

1. У ході досліджень вперше розроблено технологію багатофакторного прогнозування активності контенту в глобальних інформаційних мережах, яка базується на реалізації запропонованих математичних моделей кластеризації інформаційних повідомлень та багатофакторного прогнозування активності контенту на спеціалізованому програмно-апаратному комплексі, побудованому за принципами розподіленої комп'ютерної системи з відкритою архітектурою та відрізняється впровадженням архітектурних рішень програмної компоненти з імплементацією розроблених математичних моделей кластеризації та багатофакторного прогнозування інформаційних повідомлень.

Запропоноване архітектурне рішення є базовим модулем, який розгортається і реалізується в конкретиці складу компонент залежно від задач і практики його застосування. Наприклад для оперативного визначення соціальних настроїв різних сфер соціуму, фокус-групах, цільових аудиторій, що реалізуються: комерційними структурами в сфері надання послуг, реалізації продажів, тощо; в органах державного управління різної ланки – місцевого самоврядування, міністерств та відомств; в сфері безпеки і оборони.

Окремі елементи програмного забезпечення для прискорення обчислень розгорнуті на мікрокомп'ютерному обладнанні, зокрема: алгоритми очищення від аномалій; визначення порядку поліноміальної моделі; метод найменших квадратів; алгоритми побудови нелінійних за параметрами моделей. Зазначене реалізовано на мікрокомп'ютері типу Raspberry Pi 4, Model B, 8GB (RPI4-MODBP-8GB і дозволяє пришвидшення процесів обчислень на до 10% порівняно із повнофункціональними комп'ютерними системами.

2. Оцінювання ефективності запропонованих рішень здійснювалось в комплексному їх використанні для конкретних задач практики. Для цього розроблено демонстраційний прототип запропонованої технології.

Моніторинг інформаційних джерел реалізовано з отриманням інформаційних повідомлень за період 2020-2025рр. Dataset склав понад 1800 вимірів часового ряду для 7 сфер по 3 тональності в кожній. Дослідження показали, що часові ряди мають тренд, забруднений випадковими похибками з нормальним та експоненціальним законом. Також мають місце аномалії типу грубі значення та пропуски.

За результатами розрахунків виявлено ділянку біфуркації та пікову активність: початок біфуркаційного інтервалу березень 2022р.; точка пікової активності серпень 2022р. Індекс Херста $H=0.4$ – інформаційна активність надалі змінює тенденцію. Зазначені дати, з огляду на реальні події, дозволяють верифікувати результати розрахунків як такі, що відповідають дійсності.

Таким чином розроблена технологія є дієвою та придатна для визначення домінантного інформаційного контенту в глобальних інформаційних мережах.

Використання розробленої технології забезпечує підвищення достовірності та повноти визначення домінантного інформаційного контенту в глобальних інформаційних мережах.

ВИСНОВКИ ПО РОБОТІ

У дисертаційній роботі здійснено розв'язок актуальної науково-прикладної задачі розробки і удосконалення математичних моделей багатофакторного аналізу та машинного навчання для обробки часових рядів.

1. Аналіз показав, що протидія інформаційним загрозам потребує організації ефективного моніторингу глобальних інформаційних мереж та прогнозування активності його інформаційного контенту. Своєчасна та адекватна реакція на виявлені ІЗ і недопущення їх переростання в небезпечні ситуації вимагає оперативних, достовірних та точних даних про: рівень їх безпеки та пріоритетність у нейтралізації; тенденцій та динаміці розвитку; критичних часових точок та інтервали зміни динаміки розвитку.

Недоліки відомих підходів до кластеризації інформаційних повідомлень за принципами обробки природньої мови включають: препроцесінг – складність роботи в національному інформаційному просторі; суб'єктивізм в кластеризації – необхідність врахування суб'єктивної спрямованості в кластеризації за тональністю; конкретика цільового спрямування – потребує врахування сфер та об'єктів ІВ; великий обсяг та значна динаміка зміни інформаційного контенту – потребує оперативного аналізу, зберігання та компресії вхідних даних.

Класичні підходи до статистичного навчання потребують модифікації в напрямку застосування нелінійних за параметрами моделей. В той же час використання алгоритмів апроксимації не передбачається за можливе.

Стрімкий розвиток методів глибинного навчання з використанням штучних нейронних мереж для обробки часових рядів робить їх пріоритетними для застосування і розвитку.

2. У ході досліджень отримала подальший розвиток математична модель кластеризації інформаційних повідомлень, яка базується на частотному аналізі їх тональності з використанням методологій обробки природньої мови (Natural Language Processing (NLP)) з підтримкою великих мовних, моделей; OLAP

візуалізації результатів кластеризації та вирізняється встановленою системою індикаторів активності інформаційного контенту за сферами діяльності соціуму з ієрархічною компресією вхідних Big Data масивів, що обумовлює модель бази даних для їх зберігання. Це забезпечує підвищення повноти аналізу інформаційних повідомлень в глобальних інформаційних мережах шляхом врахування множини факторів за сферами діяльності соціуму.

Для визначення функціональної придатності до практичного застосування, верифікації та оцінювання ефективності запропонованої математичної моделі кластеризації інформаційних повідомлень розроблено програмний модуль моніторингу інформаційних джерел та кластеризації інформаційних повідомлень.

Розрахунки показали наступні показники ефективності кластеризації за тональністю інформаційних повідомлень: ймовірність вірної ідентифікації / ефективність $P = 0,87$; ймовірність промаху / втрати $F = 0,2$. Зазначені показники ефективності є прийнятними для практичного використання результатів кластеризації.

Результати OLAP аналізу довели, що наразі в інформаційному просторі кількість негативних повідомлень привалює над позитивними. Це спостерігається увесь період моніторингу. Співвідношення позитив / негатив / нейтральні повідомлення протягом доби мають відносно стабільний баланс. Протягом часу зміна активності контенту за тональністю має коливальний монотонний тренд з випадковим шумом. Пікова (аномальна) кількість негативних повідомлень спостерігаються на фактах підтверджених резонансних подій. Таким чином, сформовані індикатори активності інформаційного контенту як частота тональності інформаційних повідомлень відображають реальні події і динаміку інформаційних повідомлень, а відтак є придатними для практичного застосування.

Таким чином, запропоновані рішення з кластеризації інформаційних повідомлень є ефективними та придатними до практичного застосування.

1. У ході досліджень вперше запропоновано математичну модель багатфакторного прогнозування активності контенту в глобальних

інформаційних мережах, яка базується на обробці та аналізі часових рядів з використанням методів статистичного та глибинного машинного навчання і відрізняється від відомих рішень: запропонованими підходами до виявлення і очищення вхідних даних від аномальних вимірів; використанням нелінійних за параметрами моделей в методах статистичного навчання; модифікацією згорткових нейронних мереж для обробки часових рядів із впровадженням нелінійної схеми компромісів в моделі нейронної структури та моделюванням навчальної вибірки для екстраполяції за методами статистичного навчання; визначенням інтегрованої оцінки активності інформаційного контенту за нелінійною схемою компромісів як результат багатofакторного аналізу вектору факторів за сферами діяльності соціуму; пошуком прогнозованих біфуркаційних інтервалів як критичної активності інформаційного контенту.

Поєднання в єдиному інформаційно-процесному середовищі розробленої математичної моделі запропонованих та відомих рішень забезпечує отримання синергетичного ефекту в підвищення достовірності (точності оцінювання і прогнозу) та повноти (множини врахованих факторів) визначення домінантного інформаційного контенту в глобальних інформаційних мережах в середньому до 50% порівняно з відомими аналогами.

4. Вперше розроблено технологію багатofакторного прогнозування активності контенту в глобальних інформаційних мережах, яка базується на реалізації запропонованих математичних моделей кластеризації інформаційних повідомлень та багатofакторного прогнозування активності контенту на спеціалізованому програмно-апаратному комплексі, побудованому за принципами розподіленої комп'ютерної системи з відкритою архітектурою та відрізняється впровадженням:

багатокритеріального архітектурного синтезу апаратної компоненти;

впровадженням принципів ситуаційного управління;

архітектурних рішень програмної компоненти з імплементацією розроблених математичних моделей кластеризації та багатofакторного прогнозування інформаційних повідомлень.

Використання розробленої технології забезпечує підвищення достовірності та повноти визначення домінантного інформаційного контенту в глобальних інформаційних мережах.

5. Дослідження ефективності запропонованих рішень здійснювалось методами математичного моделювання та шляхом обробки реальних даних. Об'єкт дослідження ефективності – технологія багатофакторного прогнозування активності контенту в глобальних інформаційних мережах, яка базується на реалізації запропонованих математичних моделей кластеризації інформаційних повідомлень та багатофакторного прогнозування активності контенту на спеціалізованому програмно-апаратному комплексі. Розглядалась конкретна прикладна задача прогнозування активності інформаційного контенту за результатами моніторингу глобальних інформаційних мереж та виявлення біфуркаційних інтервалів у поведінці трендів контрольованих параметрів – кластерів частоти ІІ, за тональністю та за сферами діяльності соціуму.

Результати досліджень доводять достовірність результатів обробки та ефективність запропонованих рішень за показниками точності, оперативності та повноти від 10 до 50 відсотків, порівняно із відомими однофакторними підходами.

Запропонована технологія доведена до демонстраційного прототипу, а практика її застосування для практичних задач органів місцевого самоврядування довели доцільність практичного застосування запропонованих рішень (див. акти впровадження Додаток А).

Таким чином, цільову настанову досліджень з підвищення достовірності і повноти визначення домінантного інформаційного контенту в глобальних інформаційних мережах досягнуто. Усі часткові завдання виконані у повному обсязі.

ЛІТЕРАТУРА

1. O. Pysarchuk, A. Gizun, A. Dudnik, T. V. Griga, Domkiv, S. Gnatyuk. "Bifurcation prediction method for the emergence and development dynamics of information conflicts in cybernetic space." CEUR Workshop Proceedings, 2020, 2654, pp. 692–709. <http://ceur-ws.org/Vol2654/paper54.pdf>.
2. Левченко О.В. Класифікація інформаційної зброї за засобами ведення інформаційної боротьби / О.В. Левченко // Сучасні інформаційні технології у сфері безпеки та оборони, №2 (20), 2014. – С. 142 – 146.
3. Жарков Я.М. Інформаційно-психологічне протиборство (еволюція та сучасність): Монографія. / Я.М. Жарков, В.М. Петрик, М.М. Присяжнюк та ін.. – К.: ПАТ «Віполь», 2013. – 248 с.
4. Петрик В.М. Сугестивні технології маніпулятивного впливу: навч. посіб. / В.М. Петрик, М.М. Присяжнюк, Л.Ф. Компанцева, Є.Д. Скулиш, О.Д. Бойко, В.В. Остроухов, за аг.ред Є.Д. Скулиша. – 2-ге вид. – К.: ЗАТ «ВІПОЛЬ», 2011. – 248 с. НУОУ, 2012.-264 с.
5. Петрик М.В. Сутність і особливості проведення спеціальних інформаційних операцій та акцій інформаційного впливу / М.В. Петрик // Сучасні інформаційні технології у сфері безпеки і оборони. – №3(6), 2009. – С. 71-75.
6. Поляруш О.О. Парадигма інформаційно-психологічних впливів на соціальні системи. / О.О. Поляруш / Інформаційна безпека людини, суспільства, держави. – №1(1), 2009. – С. 21-26.
7. Hryshchuk, R., & Molodetska, K. (2017). Synergetic control of social networking services actors' interactions. In Recent Advances in Systems. Control and Information Technology: Proceedings of the International Conference SCIT 2016, May 20-21, 2016, Warsaw, Poland, pp. 34-42.

8. Hryshchuk, R., & Molodetska-Hrynhchuk, K. (2018). Methodological foundation of state's information security in social networking services in conditions of hybrid war. *Information & Security*, 41.
9. Hryshchuk, R., Molodetska, K., & Syerov, Y. (2019). Method of improving the information security of virtual communities in social networking services.
10. Hryshchuk, R., Zhovnovatiuk, R., & Nosova, H. (2019). Гібридні загрози у кіберпросторі: фактори впливу на природу виникнення. Сучасні інформаційні технології у сфері безпеки та оборони, 36(3), С. 53-58.
11. Hryshchuk, R.; Molodetska, K.; & Syerov, Y (2020) Method of Improving the Information Security of Virtual Communities in Social Networking Services. *Proceedings of the CEUR Workshop*, pp. 1754-1764.
12. Hryshchuk, R.; Molodetska, K.; & Syerov, Y (2020) Method of Improving the Information Security of Virtual Communities in Social Networking Services. *Proceedings of the CEUR Workshop*.
13. Pitafi, S.; Anwar, T.; Sharif, Z. A Taxonomy of Machine Learning Clustering Algorithms, Challenges, and Future Realms. *Appl. Sci.* 2023, 13, 3529. <https://doi.org/10.3390/app13063529>
14. O. Puchkov, D. Lande, I. Subach Methods of Creating, Clustering and Visualization of Correlation Networks Determined by the Dynamics of Thematic Information Flows // Collection “Information Technology and Security”, 3 N 1 - pp., 6-16. DOI: 10.20535/2411-1031.2025.13.1.328753, 2025, <http://dwl.kiev.ua/art/its2025-1/its2025-1.pdf>
15. Lande D., Strashnoy L. Advanced Semantic Networking based on the large language models: Monograph. - Kyiv: Engineering - 258 p. ISBN 978-617-8180-02-7, 2025.
16. Al-Yazidi, S., Berri, J., Al-Qurishi, M., & Al-Alrubaian, M. Measuring reputation and influence in online social networks: a systematic literature review. *IEEE Access*, 8, pp. 105824-105851, 2020, <https://doi.org/10.1109/ACCESS.2020.2999033>

17. Ananthaswamy, V., & Seethalakshmi, B. (2015). Mathematical analysis of information dissemination model for social networking services. *American Journal of Modeling and Optimization*, 3(1), pp. 26-34, 2015, <https://pubs.sciepub.com/ajmo/3/1/4/>
18. Dong, J., Chen, B., Liu, L., Ai, C., & Zhang, F. The analysis of influencing factors of information dissemination on cascade size distribution in social networks. *IEEE Access*, 6, 54185-54194, 2018, <https://ieeexplore.ieee.org/document/8472147>
19. Cui, P., Yin, B., & Xu, B. The application of social recommendation algorithm integrating attention model in movie recommendation. *Scientific Reports*, 13(1), pp. 169-188, 2023, <https://www.nature.com/articles/s41598-023-43511-1>
20. Lande, D., & Feher, A. (2023). OSINT Time Series Forecasting Methods Analysis. *Theoretical and Applied Cybersecurity*, 5(1).
21. Линник, Р., & Висоцька, В. (2025). Аналіз семантичних кластерів у україномовних дописах: методи NLP та візуалізації. *Herald of Khmelnytskyi National University. Technical Sciences*, 355(4), 349-356. <https://doi.org/10.31891/2307-5732-2025-355-50>
22. Певцов Г.В. Методичний підхід до кластеризації інформаційних повідомлень у ході протидії інформаційно-психологічним впливам противника /Г.В. Певцов, С.В. Залкін, С.О. Сідченко, К.І. Хударковський // Наука і техніка Повітряних Сил Збройних Сил України, 2019, №2(35) С.39-46.DOI:10.30748/nitps.2019.35.05.
23. Батюк, Т., та Досин, Д. (2023). Реалізація інтелектуальної системи аналізу настроїв та кластеризації публікацій у соціальній мережі twitter. *Інноваційні технології та наукові рішення для галузей*, (1 (23), 25–44. <https://doi.org/10.30837/ITSSI.2023.23.025>
24. Писарчук О.О. Багатокритеріальні математичні моделі ситуаційного управління та самоорганізації у складних інформаційних системах / Писарчук О.О., Даник Ю. Г., Шестаков В.І., Соколов К.О., та інші. – Монографія. – Житомир: ПП «Рута», 2016. – 232 с.

25. Kumar J. Big Data and Analytics: The key concepts and practical applications of big data analytics / J. Kumar, A. Kumar, R. Kumar. – BPB Publications, 2024. – 232 p.
26. Dietrich D. Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data / D. Dietrich, B. Heller, B. Yang. – Indianapolis, Indiana: John Wiley & Sons, 2015. – 420 p.
27. Provost F. Data Science for Business / F. Provost, T. Fawcett. – New York: O'Reilly Media, Inc, 2013. – 409 p.
28. Уманець Т.В. Статистика : навч. посібник / Т.В.Уманець, Ю.В.Пігарев. – К.: Вікар, 2013. – 429 с
29. Brockwell P. J. Introduction to Time Series and Forecasting / P. J. Brockwell, R. A. Davis. – New York: Springer, 2016. – 439 p.
30. Machine Learning for Anomaly Detection: A Systematic Review / [A. B. Nassif, M. A. Talib, Q. Nasir, F. M. Dakalbab] // IEEE Access. – 2021. – 9. – P: 78658–78700. DOI: 10.1145/3439950.
31. Deep Learning for Anomaly Detection: / [G. Pang, C. Shen, L. Cao, A. Hengel] // ACM Computing Surveys. – 2021. – 54(2). – P. 1–38. DOI: 10.1145/3439950.
32. Conditional Anomaly Detection [X. Song, M. Wu, C. Jermaine, S Ranka] // IEEE Transactions on Knowledge and Data Engineering. – 2007. – 19. – P. 631–645. DOI: 10.1109/TKDE.2007.1009.
33. Mehrotra K. G. Anomaly Detection Principles and Algorithms / K. G. Mehrotra, C. K. Mohan, H. Huang. – Switzerland: Springer, 2017. – 229 p.
34. McKinney W. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython / McKinney W. – O'Reilly Media, 2017. – 550 p.
35. Nelli F. Python Data Analytics: With Pandas, NumPy, and Matplotlib, 2nd ed. Edition / F. Nelli. – Apress, 2018. – 588p.
36. Raschka S. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, Second Edition / S. Raschka, V. Mirjalili. – Packt Publishing, 2017. – 622 p.

37. Joshi P. Artificial Intelligence with Python: A Comprehensive Guide to Building Intelligent Apps for Python Beginners and Developers / P. Joshi. – Packt Publishing, 2017. – 466 p.
38. An introduction to statistical learning: With applications in python. / [G. James, D. Witten, T. Hastie, R. Tibshirani, J. Taylor]. – Springer Nature, 2023. – 613 p.
39. Nguyen M. A. Guide on Data Analysis. – Bookdown, 2020.
https://bookdown.org/mike/data_analysis/
40. Chapra S. C. Numerical Methods for Engineers, Seventh Edition. / S. C. Chapra, R. P. Canale. – New York: McGraw-Hill Education, 2015. – 987p.
41. Hastie T. The elements of statistical learning: data mining, inference, and prediction. / T. Hastie, R. Tibshirani, J. Friedman // Springer Series in Statistics. – Springer, USA, 2009 – 745p.
42. Provost F. Data Science for Business / F. Provost, T. Fawcett. – O'Reilly Media, Inc, USA, 2013. – 409p.
43. Guang-yong C. Generalized exponential autoregressive models for nonlinear time series: stationarity, estimation and applications. / C. Guang-yong, M. Gan, G. Chen. // Information Sciences, vol. 438, 2018. – p. 46-57. DOI: <https://doi.org/10.1016/j.ins.2018.01.029>
44. Method for Statistical Evaluation of Nonlinear Model Parameters in Statistical Learning Algorithms. / [O. Tuhanskyh, D. Baran, O. Pysarchuk], Yang, XS., Sherratt, S., Dey, N., Joshi, A. (eds) // Proceedings of Ninth International Congress on Information and Communication Technology. ICICT 2024. Lecture Notes in Networks and Systems. – Springer, Singapore. – 2024. – Vol 1013 – p. 265-274. DOI: https://doi.org/10.1007/978-981-97-3559-4_21.
45. Dietrich D. Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data / D. Dietrich, B. Heller, B. Yang. – John Wiley & Sons, Inc., Indianapolis, Indiana, 2015. – 420p.
46. Pronzato L. Design of experiments in nonlinear models. L. Pronzato, A. Pázman. // Lecture notes in statistics. vol. 212.1. – Springer New York, 2013. – 399p.

47. S. Konishi. Introduction to multivariate analysis: linear and nonlinear modeling. // CRC Press. – Taylor & Francis Group, Tokyo, 2014. – 312p.
48. Grafarend. E.W. Applications of linear and nonlinear models. / E.W. Grafarend, J.L. Awange. S. Zwanzig // Springer Geophysics, Fixed Effects. – Springer Nature, Switzerland, 2012. – 1113p.
49. Raymond J.C. Measurement error in nonlinear models: a modern perspective. / [C.J. Raymond, D. Ruppert, L.A. Stefanski, C.M. Crainiceanu] – Chapman and Hall, CRC, 2006. – 242p.
50. Єріна А., Мазуренко О. Є71 Статистичний аналіз часових рядів: навчальний посібник. Київ. КНУІШ, 2022. 164 с.
51. Андрієнко В.М., Семенов А.С. (2018). Методика статистичного аналізу економічних часових рядів. Науковий вісник УжНУ. Серія Міжнар. Економ. відносини та світове гос.-во., Вип. 21. Ч. 1. С. 5-13. http://www.visnyk-econom.uzhnu.uz.ua/archive/21_1_2018ua/3.pdf.
52. Андрусенко Ю.О. (2020). Аналіз основних моделей прогнозування часових рядів. Зб. наук. праць ХНУПС, № 3(65). С. 91-96. DOI: <https://doi.org/10.30748/zhups.2020.65.14>.
53. Денисюк Р. С., Буркіна Н. В. (2019). Використання методів статистичного моделювання при аналізі зміни динаміки експорту товарів України, С. 190-195. <https://jvestnik-sss.donnu.edu.ua/article/view/6707/6739>
54. Харченко В. П., Писарчук О. О. Нелінійне та багатокритеріальне моделювання процесів у системах керування рухом : монографія. Київ: Ін-т обдарованої дитини, 2015. 248 с.
55. E. Egrioglu, M. Khashei, C. H. Aladag, I. B. Turksen, and U. Yolcu, “Advanced Time Series Forecasting Methods,” Mathematical Problems in Engineering, vol. 2015, pp. 1–2, 2015, doi: 10.1155/2015/918045.
56. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. Springer New York, 2009. doi: 10.1007/978-0-387-84858-7.

57. G. P. Zhang, Time series forecasting using a hybrid ARIMA and neural network model, *Neurocomputing*, vol. 50, pp. 159–175, Jan. 2003, doi: 10.1016/s0925-2312(01)00702-0.
58. M. Khashei and M. Bijari, A new class of hybrid models for time series forecasting, *Expert Systems with Applications*, vol. 39, no. 4, pp. 4344–4357, Mar. 2012, doi: 10.1016/j.eswa.2011.09.157.
- A. Tealab, Time series forecasting using artificial neural networks methodologies: A systematic review, *Future Computing and Informatics Journal*, vol. 3, no. 2, pp. 334–340, Dec. 2018, doi: 10.1016/j.fcij.2018.10.003.
59. P. Singh, M. K. Singh, R. Singh, and N. Singh, *Federated Learning: Challenges, Methods, and Future Directions*, EAI/Springer Innovations in Communication and Computing. Springer International Publishing, pp. 199–214, 2022. doi: 10.1007/978-3-030-85559-8_13.
60. L. Su, X. Zuo, R. Li, X. Wang, H. Zhao, and B. Huang, A systematic review for transformer-based long-term series forecasting, *Artif Intell Rev*, vol. 58, no. 3, Jan. 2025, doi: 10.1007/s10462-024-11044-2.
61. R. Velastegui, L. Zhinin-Vera, G. E. Pilliza, and O. Chang, Time Series Prediction by Using Convolutional Neural Networks, *Advances in Intelligent Systems and Computing*. Springer International Publishing, pp. 499–511, Oct. 31, 2020. doi: 10.1007/978-3-030-63128-4_38.
62. D. Baran and O. Pysarchuk, Mathematical Model of Clustering of Informational Messages with Indicators of Activity for the Information Content by Tone and Areas of Society Activity, *Inf. Comput. and Intell. syst. j.*, no. 6, pp. 230–241, Sept. 2025, doi: 10.20535/2786-8729.6.2025.339127.
63. N. Voronin, A method of multicriteria evaluation and optimization of hierarchical systems,” *Cybern Syst Anal*, vol. 43, no. 3, pp. 384–390, May 2007, doi: 10.1007/s10559-007-0060-8.
64. Gerrit Muller. *System Modeling and Analysis: a Practical Approach*. – University of South-Eastern Norway-NISE, 2023. – 128p.<https://www.gaudisite.nl/SystemModelingAndAnalysisBook.pdf>

65. Nolberto Munier. Mathematical Modelling of Decision Problems. Using the SIMUS Method for Complex Scenarios. – Springer, 2021. – 196p.
66. Albert Voronin. Multi-Criteria Decision Making for the Management of Complex Systems. – IGI Global, 2017. – 220p.
67. Goran Cirović. Multiple-Criteria Decision Making / Goran Cirović ' Dragan Pamučar. – MDPI. – 2022. – 313p.
68. <https://fakty.com.ua/ua/proisshestvija/20250405-vybuhy-u-kryvomu-rozi-4-kvitnya-2025-roku-shho-vidomo/>
69. <https://fakty.com.ua/ua/proisshestvija/20250406-vybuhy-u-kyievi-u-troh-rajonah-stolyczi-stalysya-pozhezhi-cherez-raketnu-ataku-rf/>
70. Algorithms of Statistical Anomalies Clearing for Data Science Applications - Oleksii Pysarchuk, Yurii Mironov, Illya Pysarchuk, Danylo Baran. // System research and information technologies. – No. 1 (2023), P.78-84. <https://doi.org/10.20535/SRIT.2308-8893.2023.1.06>, <http://journal.iasa.kpi.ua/article/view/260175>.
71. O. Pysarchuk, T. Andreieva, O. Grinenko, D. Baran. Multi-factor forecasting of statistical trends for data science problems // System Research & Information Technologies, 2024, No 2, P.21-34. <https://doi.org/10.20535/SRIT.2308-8893.2024.2.02> <http://journal.iasa.kpi.ua/article/view/287739>.
72. Pysarchuk, O., Tuhanskykh, O., Baran, D. (2026). Structural Adaptation of Nonlinear Models in Time Series Processing Problems. In: Zgurovsky, M., Pankratova, N. (eds) System Analysis and Data Mining. Studies in Systems, Decision and Control, vol 609. Springer, Cham. ISBN 978-3-031-97529-5 https://doi.org/10.1007/978-3-031-97529-5_3/
73. O. O Pysarchuk, O. A. Tuhanskykh, D. R. Baran (2025). Method for determining the structure of nonlinear models for time series processing. Radio Electronics, Computer Science, Control, 1, pp. 53-62 DOI: 10.15588/1607-3274-2025-1-5, <https://ric.zp.edu.ua/article/view/324253>.
74. Baran D., Pysarchuk O., Pavlova S (2025). The method of adaptation of the parameters of algorithms for the detection and cleaning of a statistical sample

- from anomalies for data science problems. *Radio Electronics, Computer Science, Control*, 3, pp. 37-44 DOI: 10.15588/1607-3274-2025-3-4, <https://ric.zp.edu.ua/article/view/339011>.
75. Baran D. Pysarchuk O., Pysarchuk I., Vasylieva M. (2025). Multi-criteria mathematical model of credit scoring in data science problems. *System research and information technologies*, No. 3 (2025), pp. 99-112. <http://journal.iasa.kpi.ua/article/view/308786>.
76. Баран Д.Р., Писарчук О.О., Корочкін О.В., Визначення порядку поліноміальної моделі для побудови лінії тренду в задачах Data Science // Проблеми інформатизації та управління, 3(71), 2022/ – С.35-40. <https://doi.org/10.18372/2073-4751.71.17001> <https://jrnl.nau.edu.ua/index.php/PIU/article/view/17001>.
77. Писарчук, О., Баран, Д., & Туганських, О. (2023). Методика статистичного навчання параметрів експоненційної математичної моделі. *Measuring and computing devices in technological processes*, (3), 224–227. <https://doi.org/10.31891/2219-9365-2023-75-26> <https://vottp.khmnu.edu.ua/index.php/vottp/article/view/177>.
78. Baran D., Pysarchuk O. (2025). Mathematical Model of Clustering of Informational Messages with Indicators of Activity for the Information Content by Tone and Areas of Society Activity. *Information, Computing and Intelligent systems*, 7, pp 39-48 DOI: <https://doi.org/10.20535/2786-8729.6.2025.339127>, <https://itvisnyk.kpi.ua/article/view/339127>.
79. Pysarchuk O., Baran D. (2025). Optimization neural network for time series processing. *Information, Computing and Intelligent systems*, 6, pp 230-241 DOI: <https://doi.org/10.20535/2786-8729.7.2025.341480>, <https://itvisnyk.kpi.ua/article/view/341480>.
80. Konrad, T., Pysarchuk, O., Noga, H., Gajewska, A., Volkogon, V., Baran, D. (2024). Method of Multifactor Analysis of the Alternatives Using the Example of Multi-criterial Selection of the Optimal Cargo Transportation Route. In: Yang, XS., Sherratt, S., Dey, N., Joshi, A. (eds) *Proceedings of Ninth*

- International Congress on Information and Communication Technology. ICICT 2024 2024. Lecture Notes in Networks and Systems, vol 1002. Springer, Singapore. https://doi.org/10.1007/978-981-97-3299-9_6
https://link.springer.com/chapter/10.1007/978-981-97-3299-9_6#citeas
81. Tuhanskyk, O., Baran, D., Pysarchuk, O. (2024). Method for Statistical Evaluation of Nonlinear Model Parameters in Statistical Learning Algorithms. In: Yang, XS., Sherratt, S., Dey, N., Joshi, A. (eds) Proceedings of Ninth International Congress on Information and Communication Technology. ICICT 2024 2024. Lecture Notes in Networks and Systems, vol 1013. Springer, Singapore. https://doi.org/10.1007/978-981-97-3559-4_21
https://link.springer.com/chapter/10.1007/978-981-97-3559-4_21
 82. Метод аналітичного визначення параметрів нелінійних моделей та статистичною вибіркою вимірів – Баран Д., Туганських О., Писарчук О. // Міжнародна науково-практична конференція молодих вчених та студентів «Інженерія програмного забезпечення і передові інформаційні технології» (SoftTech-2023) Кафедра ІІІ ФІОТ, КІІ ім. Ігоря Сікорського, 2023, <https://ela.kpi.ua/handle/123456789/59987>
<https://ela.kpi.ua/items/f02db5cb-3489-4c68-941e-179fdca13cbe/full>
 83. Algorithms for purification of statistical data samples from anomalies for Data Science tasks – Baran D., Pysarchuk O. // The International Conference on Security, Fault Tolerance, Intelligence” (ICSFTI2022 Online).
 84. Oleksii Pysarchuk, Oleksandr Tuhanskykh, and Danylo Baran Structural Adaptation of Nonlinear Models in Time Series Processing Problems // SAIC'2024 (IEEE International Conference on System Analysis & Intelligent Computing), http://saic.ieee.org.ua/?page_id=96
 85. Oleksii Pysarchuk, Oleksandr Tuhanskykh, and Danylo Baran Methodology for statistical learning of nonlinear model parameters with the integral form of ols in the scheme of differential-nontaylor transformations // SAIC'2024 (IEEE International Conference on System Analysis & Intelligent Computing), http://saic.ieee.org.ua/?page_id=96

86. Oleksii Pysarchuk, Oleksandr Tuhanskykh, and Danylo Baran Methodology for statistical learning of nonlinear model parameters with the integral form of ols in the scheme of differential-nontaylor transformations // Security, Fault Tolerance, Intelligence: proceedings of the VIII International Conference ICSFTI2025, Kyiv, Ukraine, may 14, june 30, 2025. – Kyiv : Igor Sikorsky Kyiv Polytechnic Institute, publishing house “Polytechnica”, 2025. <https://icsfti-proc.kpi.ua/article/view/338732>
87. Теорія інформації та кодування: Навчальний посібник для підготовки до практичних занять / Уклад. Гайдур Г.І, Бондаренко З.З. – К.: ДУІКТ, 2024 – 43 с.

ДОДАТОК А

Акти впровадження результатів дисертаційних досліджень

Данили БАРАНА

ЗАТВЕРДЖУЮ

Начальник Житомирського військового
інституту імені С. П. Корольова
доктор історичних наук, професор
полковник



Андрій СЛЮСАРЕНКО

2025 року

АКТ ВПРОВАДЖЕННЯ

результатів дисертаційних досліджень аспіранта кафедри
обчислювальної техніки факультету інформатики та
обчислювальної техніки Національного технічного університету
України «Київський політехнічний інститут імені Ігоря
Сікорського» Барана Данила Романовича за темою «Математичні
моделі багатофакторного прогнозування активності контенту в
глобальних інформаційних мережах»

Комісія у складі –

голови комісії, начальника наукового центру кандидата технічних наук
полковника Перегуди Олександра;

членів комісії:

заступника начальника наукового центру кандидата технічних наук
підполковника Романчука Миколи;

заступника начальника науково-дослідного відділу інформаційної та
кібернетичної безпеки наукового центру доктора філософії підполковника
Ахтирцевої Владислави;

старшого наукового співробітника науково-дослідного відділу
інформаційної та кібернетичної безпеки наукового центру підполковника
Павленка Михайла –

склали цей Акт про те, що результати дисертаційних досліджень Барана
Данила Романовича, а саме:

математична модель кластеризації інформаційних повідомлень, яка
базується на частотному аналізі їх тональності з використанням методологій
обробки природної мови;

математична модель багатофакторного прогнозування активності
контенту в глобальних інформаційних мережах з використанням нелінійних
математичних моделей в методах статистичного навчання;

технологія багатофакторного прогнозування активності контенту в
глобальних інформаційних мережах, яка базується на реалізації
запропонованих математичних моделей кластеризації інформаційних
повідомлень та багатофакторного прогнозування –

впроваджено в оперативному завданні, шифр “Простір” у Житомирському

військовому інституті імені С. П. Корольова, що присвячено розробленню спеціального програмного забезпечення для моніторингу інформаційного простору в інтересах Збройних Сил України.

Комісія встановила, що матеріал, викладений у вищевказаних результатах дисертаційних досліджень буде використаний при створенні аналітичної складової спеціального програмного забезпечення для моніторингу інформаційного простору.

ВИСНОВОК:

Зазначені результати дисертаційних досліджень дозволять:

підвищити ефективність моніторингу інформаційного простору завдяки використанню сучасних технологій кластеризації та аналізу тональності повідомлень;

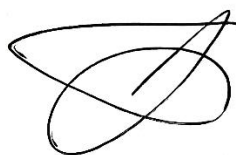
забезпечити можливість прогнозування активності інформаційного контенту, що сприяє оперативному реагуванню на інформаційні загрози;

покращити інструменти аналізу великих масивів даних у глобальних інформаційних мережах завдяки впровадженню багатофакторних моделей прогнозування.

Впровадження цих моделей і технологій в рамках оперативного завдання сприятиме підвищенню інформаційної безпеки, забезпеченню оперативності прийняття рішень у сфері моніторингу та управління інформаційним середовищем в інтересах Збройних Сил України.

Голова комісії –

полковник



Олександр ПЕРЕГУДА

Члени комісії:

підполковник



Микола РОМАНЧУК

підполковник



Владислава АХТИРЦЕВА

підполковник



Михайло ПАВЛЕНКО



**ЗВЯГЕЛЬСЬКА РАЙОННА ДЕРЖАВНА АДМІНІСТРАЦІЯ
ЖИТОМИРСЬКОЇ ОБЛАСТІ
ЗВЯГЕЛЬСЬКА РАЙОННА ВІЙСЬКОВА АДМІНІСТРАЦІЯ
ЖИТОМИРСЬКОЇ ОБЛАСТІ**

вул. Шевченка, буд. 16, м. Звягель, Звягельський р-н, Житомирська обл., 11701,
тел.(04141) 3-53-91, <https://zvrda.gov.ua>, e-mail: 04053654@nvrda.zht.gov.ua; код ЄДРПОУ 04053654

20.02.2025 № 01-35/28

На № _____ від _____

ДОВІДКА

**про впровадження результатів дисертаційних досліджень
Данили БАРАНА аспіранта кафедри обчислювальної техніки,
факультету інформатики та обчислювальної техніки Національного
технічного університету України «Київський політехнічний інститут
імені Ігоря Сікорського»**

Дійсним зазначаємо, що програмна система «Багатофакторного прогнозування активності контенту в глобальних інформаційних мережах» де використані результати дисертаційних досліджень Данили БАРАНА впроваджено для забезпечення автоматизації процесу щоденного моніторингу інформаційного простору в оточенні Звягельського району.

Результати моніторингу дозволяють оцінювати стан настроїв населення та превентивно реагувати на виклики і загрози в інформаційній сфері.



Заступник начальника

Олег ЗУБОВСЬКИЙ



**ВИКОНАВЧИЙ КОМІТЕТ КОЦЮБИНСЬКОЇ СЕЛИЩНОЇ РАДИ
БУЧАНСЬКОГО РАЙОНУ КИЇВСЬКОЇ ОБЛАСТІ**

вул. Доківська, 2, сел. Коцюбинське, Бучанський р-н, Київська обл., 08298
тел. (04597) 95-2-37, E-mail: radakocubinske@gmail.com, Код ЄДРПОУ 44474482

від 30.01. 2025 року № 151/01.28

АКТ ВПРОВАДЖЕННЯ

результатів дисертаційних досліджень аспіранта кафедри обчислювальної техніки, факультету інформатики та обчислювальної техніки Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» Барана Данила Романовича за темою «Математичні моделі багатфакторного прогнозування активності контенту в глобальних інформаційних мережах» на здобуття наукового ступеня доктора філософії за спеціальності 123 «Комп'ютерна інженерія»

Встановлено, що отримані особисто Бараном Данилом Романовичем результати у ході досліджень за темою дисертації «Математичні моделі багатфакторного прогнозування активності контенту в глобальних інформаційних мережах», а саме – програмна система багатфакторного прогнозування активності контенту в глобальних інформаційних мережах, яка побудована і акумулює наукові результати:

математичну модель кластеризації інформаційних повідомлень, яка базується на частотному аналізі їх тональності з використанням методологій обробки природньої мови (NaturalLanguage Processing (NLP));

математичну модель багатфакторного прогнозування активності контенту в глобальних інформаційних мережах, яка базується на обробці та аналізі часових рядів з використанням методів нелінійного статистичного навчання;

технологію багатфакторного прогнозування активності контенту в глобальних інформаційних мережах, яка базується на реалізації запропонованих математичних моделей кластеризації інформаційних повідомлень та багатфакторного прогнозування активності

впроваджено в практику діяльності Виконавчого комітету Коцюбинської селищної ради для моніторингу настроїв населення. Зазначене дозволяє мати оперативну ситуаційну обізнаність та завчасно організовувати комплекс ефективних заходів з недопущення негативних настроїв і протидії негативним інформаційним впливам та уникнення небажаних наслідків.

**Керуюча справами
Коцюбинської селищної ради**



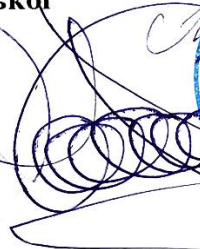
Світлана ЧЕЧКО

**Секретар Коцюбинської
селищної ради**



Юлія ГЛАВАЦЬКА

Селищний голова



Сергій ДАНИШ

ДОДАТОК 5

Програмна компонента реалізації запропонованих рішень

```

'''
Програмний модуль моніторингу інформаційних джерел та кластеризації інформаційних
повідомлень
'''

from data_parser import Parser_URL_rdk, Parser_URL_pressorg24
from data_frequency_analysis import frequency_analysis
from data_analysis_tonality import tonality_detection, parsing
import warnings
warnings.filterwarnings('ignore')

# ----- БЛОК ГОЛОВНИХ ВИКЛИКІВ -----
if __name__ == '__main__':

    path = 'data_plot/'

    print('Оберіть інформаційне джерело:')
    # ----- Головні виклики парсера для отримання даних text mining -----
    print('1 - https://www.rbc.ua/rus/news', 'text mining')
    print('2 - http://pressorg24.com/news', 'text mining')
    mode = int(input('mode:'))

    if (mode == 1):
        print('Обрано інформаційне джерело: https://www.rbc.ua/rus/news')
        url = 'https://www.rbc.ua/rus/news'
        filename = path + 'test_2.txt'
        Parser_URL_rdk(url, filename)

    if (mode == 2):
        print('Обрано інформаційне джерело: http://pressorg24.com/news')
        url = 'http://pressorg24.com/news'
        filename = path + 'test_2.txt'
        Parser_URL_pressorg24(url, filename)

    frequency_analysis(filename)

    with open(filename, encoding="utf-8") as file:
        posts = file.read()

    posts = parsing()
    tonality_detection(posts)

```

```

'''
Парсери сайтів новин
'''

import requests
from bs4 import BeautifulSoup

# ----- Парсер САЙТУ для отримання html структури і вилучення з неї стрічки
# новин -----
def Parser_URL_rdk (url, file_name):

    headers = {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/116.0.0.0 Safari/537.36'
    }
    response = requests.get(url, headers=headers)

    soup = BeautifulSoup(response.text, 'lxml')          # аналіз структури html
документу
    print(soup)
    quotes = soup.find_all('div', class_='newsline')      # вилучення із html
документу ленти новин

    with open(file_name, "w", encoding="utf-8") as output_file:
        print('----- Стрічка новин', url, '-----')
        for quote in quotes:
            print(quote.text)
            output_file.write(quote.text)    # запис ленти новин до текстового
файлу
        print('-----')

    return

# ----- Парсер САЙТУ для отримання html структури і вилучення з неї стрічки
# новин -----
def Parser_URL_pressorg24 (url, file_name):

    headers = {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/116.0.0.0 Safari/537.36'
    }
    response = requests.get(url, headers=headers)

    soup = BeautifulSoup(response.text, 'lxml')          # аналіз структури html
документу
    print(soup)
    quotes = soup.find_all('div', class_='event')        # вилучення із html документу
ленти новин

    with open(file_name, "w", encoding="utf-8") as output_file:
        print('----- Лента новин', url, '-----')
        for quote in quotes:
            print(quote.text)
            output_file.write(quote.text)    # запис ленти новин до текстового
файлу
        print('-----')

    return

```

```

'''
Фільтрація та частотний аналіз стрічки новин
'''

from nltk.draw.dispersion import dispersion_plot
from nltk import FreqDist
import matplotlib.pyplot as plt
import re
import spacy
import pandas as pd
from collections import Counter
import seaborn as sns

# ----- підключення словника - моделі -----
nlp = spacy.load("uk_core_news_sm")

# ----- фільтрація, нормалізація та частотний аналіз -----
def text_filter_ukr(filename):
    # завантаження тексту із файла
    with open(filename, encoding="utf-8") as file:
        text = file.read()
    # фільтрація від шуму - зайвих символів
    text = text.replace("\n", " ")
    text = re.sub(r'[.,?!\\-:]', '', text)

    text = re.sub(r'[0-9]+', '', text)
    modified_text = text.lower()
    words = modified_text.split()
    print('modified_text =', modified_text)
    words.sort()
    words_dict = dict()
    for word in words:
        words_dict[word] = words_dict.get(word, 0) + 1

    print("Кількість слів: %d" % len(words))
    print(f"Кількість унікальних слів: {len(words_dict)}")
    print("Усі використані слова:")
    for word, count in words_dict.items():
        print(word.ljust(20), count)
    return nlp(modified_text)

#----- дисперсійний аналіз токенів -----
def dispersion_plot_words(words, targets):
    # plt.figure(figsize=(12, 8))
    dispersion_plot(words, targets, ignore_case=True, title='Lexical Dispersion Plot')
    # dispersion_plot(words, targets)
    plt.savefig('data_plot/variance_analysis_tokens.jpg')
    plt.show()

    return

#----- частотний аналіз токенів/лем -----
def frequency_words(token_lemma_list):

    n = 10

    plt.ion()
    frequency_distribution = FreqDist(token_lemma_list)
    plt.figure(figsize=(12, 8))
    frequency_distribution.plot(n, cumulative=True, title='TF-IDF Top 10 term

```

```

frequency')
plt.savefig('data_plot/TF-IDF_frequency_distribution_10.jpg')
plt.ylabel('frequency')
plt.ioff()
plt.show()

plt.ion()
frequency_distribution = FreqDist(token_lemma_list)
plt.figure(figsize=(12, 8))
frequency_distribution.plot(n, title='Top 10 Most Common Words')
plt.savefig('data_plot/frequency_distribution_10.jpg')
plt.ylabel('frequency')
plt.ioff()
plt.show()

all_fdist = FreqDist(token_lemma_list).most_common(n)
all_fdist = pd.Series(dict(all_fdist))
fig, ax = plt.subplots(figsize=(12, 8))
sns.barplot(x=all_fdist.index, y=all_fdist.values, ax=ax)
plt.title('Top 10 Most Common Words')
plt.ylabel('frequency')
plt.xticks(rotation=30)
plt.savefig('data_plot/frequency_hist_10.jpg')
plt.show()

n = len(token_lemma_list)

plt.ion()
frequency_distribution = FreqDist(token_lemma_list)
plt.figure(figsize=(12, 8))
frequency_distribution.plot(n, cumulative=True, title='TF-IDF Top n term
frequency')
plt.savefig('data_plot/TF-IDF_frequency_distribution_n.jpg')
plt.ylabel('frequency')
plt.ioff()
plt.show()

plt.ion()
frequency_distribution = FreqDist(token_lemma_list)
plt.figure(figsize=(12, 8))
frequency_distribution.plot(n, title='Top n Most Common Words')
plt.savefig('data_plot/frequency_distribution_n.jpg')
plt.ylabel('frequency')
plt.ioff()
plt.show()

all_fdist = FreqDist(token_lemma_list).most_common(n)
all_fdist = pd.Series(dict(all_fdist))
fig, ax = plt.subplots(figsize=(12, 8))
sns.barplot(x=all_fdist.index, y=all_fdist.values, ax=ax)
plt.title('Top n Most Common Words')
plt.ylabel('frequency')
plt.xticks(rotation=30)
plt.savefig('data_plot/frequency_hist_n.jpg')
plt.show()

return

def frequency_analysis(file_name):

    # ----- фільтрація, нормалізація та частотний аналіз -----
    introduction_doc = text_filter_ukr(file_name)

```

```

# ----- токенизація -----
token_word = [token.text for token in introduction_doc] # особливості
обробки даних
print(type(token_word))
print(token_word)

# ----- фільтрація стоп-слів -----
token_word_is_stop = [token for token in introduction_doc if not
token.is_stop]
print(token_word_is_stop)

# ----- лематизація -----
token_lemma_list = []
for token in token_word_is_stop:
    if str(token) != str(token.lemma_):
        print(f"{str(token):>20} : {str(token.lemma_)}")
        token_lemma_list.append(str(token.lemma_))
print('token_lemma_list: ', token_lemma_list)

# ----- частотний аналіз лем -----
complete_doc = nlp(str(token_lemma_list))
words = [token.text for token in complete_doc if token.is_stop != True and
token.is_punct != True]
word_freq = Counter(words)
common_words = word_freq.most_common(len(word_freq))
print(common_words)

frequency_words(token_lemma_list)

words = token_lemma_list

targets = [common_words[0][0], common_words[1][0], common_words[2][0],
common_words[3][0]]
dispersion_plot_words(words, targets)

return

```

```

'''
Визначення тональності повідомлень стрічки новин
'''

import spacy.cli
from spacytextblob.spacytextblob import SpacyTextBlob
import spacy
from textblob import TextBlob
from deep_translator import GoogleTranslator
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
import nltk
import time
import feedparser
from nltk.sentiment.vader import SentimentIntensityAnalyzer
import re
import pandas as pd
from datetime import datetime
import matplotlib.pyplot as plt
import numpy as np

#----- завантаження сервісних атрибутів для бібліотеки nltk -----
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('vader_lexicon')

# ----- підключення словника - моделі -----
# nlp = spacy.load("en_core_web_sm")
spacy.cli.download("en_core_web_sm")
nlp = spacy.load("uk_core_news_sm")
nlp.add_pipe('spacytextblob')

# ----- парсинг -----
def parsing(rss_feed_url = 'https://www.pravda.com.ua/ukr/rss/view_news/'):
    # print('----- парсинг стрічки новин -----')
    posts = []
    seen_titles = set()
    response = feedparser.parse(rss_feed_url)
    for entry in response.get('entries', []):
        title = entry.get('title')
        if title in seen_titles:
            continue

        seen_titles.add(title)
        posts.append({
            "title": title,
            "link": entry.get('links', [{}])[0].get('href'),
            "tags": [t.get('term') for t in entry.get('tags', [])],
            "date": time.strftime('%Y-%m-%d', entry.get('published_parsed'))
        })
    for i, post in enumerate(posts):
        print(i, post["title"])
    return posts

#----- токенизація -----
def preprocess_text(text):
    # Токенизація
    doc = nlp(text)
    # Видалення стопслів
    filtered_tokens = [token.lemma_ for token in doc if not token.is_stop and token.text.isalpha()]
    # Перетворення токенів до рядка
    preprocessed_text = " ".join(filtered_tokens)

```

```

    return preprocessed_text

#----- переклад -----
def translator_text(text_uk):
    text_en = GoogleTranslator(source='auto', target='en').translate(text_uk)
    print('text_uk: ', text_uk)
    print('text_en: ', text_en, '\n')

    return text_en

#----- визначення тональності текста 1 -----
def analysis_tonality_TextBlob(text, Positive, Negative, Neutral):
    # Очищення тексту
    preprocessed_text = preprocess_text(text)
    blob = TextBlob(preprocessed_text)
    polarity = blob.sentiment.polarity
    if polarity > Positive:
        tonality = "Positive"
        print('TextBlob: polarity = ', polarity, "Positive")
    elif polarity < Negative:
        tonality = "Negative"
        print('TextBlob: polarity = ', polarity, "Negative")
    elif polarity == Neutral:
        tonality = "Neutral"
        print('TextBlob: polarity = ', polarity, "Neutral")

    return tonality

#----- визначення тональності текста 2 -----
def analysis_tonality_vaderSentiment(text , Positive, Negative, Neutral):
    preprocessed_text = preprocess_text(text)
    analyzer = SentimentIntensityAnalyzer()
    vs = analyzer.polarity_scores(preprocessed_text)
    # print('vaderSentiment: ', "{:-<65} {}".format(text, str(vs)), '\n')
    print('vaderSentiment: ', vs, '\n')
    # print('vaderSentiment: ', vs['compound'], '\n')
    polarity = vs['compound']
    if polarity > Positive:
        tonality = "Positive"
        print('TextBlob: polarity = ', polarity, "Positive")
    elif polarity < Negative:
        tonality = "Negative"
        print('TextBlob: polarity = ', polarity, "Negative")
    elif polarity == Neutral:
        tonality = "Neutral"
        print('TextBlob: polarity = ', polarity, "Neutral")

    return tonality

#----- визначення тональності текста 3 -----
def analysis_tonality_nltk(text , Positive, Negative, Neutral):
    tonalit_score = SentimentIntensityAnalyzer().polarity_scores(text)
    print('analysis_tonality_nltk: ', tonalit_score, '\n')
    # print('analysis_tonality_nltk: ', tonalit_score['compound'], '\n')
    polarity = tonalit_score['compound']
    if polarity > Positive:
        tonality = "Positive"
        print('TextBlob: polarity = ', polarity, "Positive")
    elif polarity < Negative:
        tonality = "Negative"
        print('TextBlob: polarity = ', polarity, "Negative")
    elif polarity == Neutral:
        tonality = "Neutral"
        print('TextBlob: polarity = ', polarity, "Neutral")

```

```

    return tonality

#----- визначення тональності текста 4 -----
def analysis_tonality_spacy(text , Positive, Negative, Neutral):
    doc = nlp(text)
    print('polarity =', doc._.blob.polarity, 'subjectivity: ',
doc._.blob.subjectivity)
    # print('polarity =' , doc._.blob.polarity)
    polarity = doc._.blob.polarity
    if polarity > Positive:
        tonality = "Positive"
        print('TextBlob: polarity = ', polarity, "Positive")
    elif polarity < Negative:
        tonality = "Negative"
        print('TextBlob: polarity = ', polarity, "Negative",)
    elif polarity == Neutral:
        tonality = "Neutral"
        print('TextBlob: polarity = ', polarity, "Neutral")

    return tonality

def tonality_detection(posts):

    def pattern_matches(pattern, tonality_rezalt):
        return len(re.findall(pattern, tonality_rezalt))

    def tonality_rezalt(number_Positive, number_Negative, number_Neutral):
        if number_Negative >= 2: tonality_return = 'Negative'
        if (number_Positive >= 2) and (number_Negative <= 1): tonality_return =
'Positive'
        if (number_Neutral >= 2) and (number_Negative <= 1): tonality_return =
'Neutral'
        return tonality_return

    def plot_frequency_tonality(result_list):

        result_list_tonalit = []
        for i, post in enumerate(p['about'] for p in result_list):
            result_list_tonalit.append(post)

        # print('result_list_tonalit !!!!!', type(str(result_list_tonalit)),
'\n')

        number_Positive = pattern_matches('Positive', str(result_list_tonalit))
        number_Negative = pattern_matches('Negative', str(result_list_tonalit))
        number_Neutral = pattern_matches('Neutral', str(result_list_tonalit))
        result_list_tonalit = np.array([number_Positive, number_Negative,
number_Neutral])
        mylabels = ["Positive", "Negative", "Neutral"]

        # Формат dd/mm/YY H:M:S
        now = datetime.now()
        dt_string = now.strftime("%d/%m/%Y %H:%M:%S")

        print("date and time =", dt_string)
        print('result_list_tonalit =', result_list_tonalit)
        print('mylabels =', mylabels)

        plt.figure(figsize=(8, 8))
        plt.title(dt_string)
        plt.pie(result_list_tonalit, labels=mylabels, autopct='%1.1f%%')
        plt.savefig('data_plot/frequency_tonality.jpg')

```

```

plt.show()

filename = 'data_plot/frequency_tonality.txt'
with open(filename, 'a', encoding='utf-8') as file:
    file.write('-----'+ "\n")
    file.write(dt_string+ "\n")
    file.write(str(mylabels)+ "\n")
    file.write(str(result_list_tonalit)+ "\n")
    file.close()
    print("Файл тональності успішно сформовано, див.:", filename)

    return

Positive = 0
Negative = 0
Neutral = 0

# posts = parsing()

print('-----posts-----')
print(posts)
print('-----')
result_list = []

for i, post in enumerate(p['title'] for p in posts):
    print('-----', i, '-----')
    text_uk = (post)
    text_en = translator_text(text_uk)

    tonality_TextBlob = analysis_tonality_TextBlob(text_en, Positive,
Negative, Neutral)
    tonality_vaderSentiment = analysis_tonality_vaderSentiment(text_en,
Positive, Negative, Neutral)
    tonality_nltk = analysis_tonality_nltk(text_en, Positive, Negative,
Neutral)
    tonality_spacy = analysis_tonality_spacy(text_en, Positive, Negative,
Neutral)

    print('tonality_TextBlob = ', tonality_TextBlob)
    print('tonality_vaderSentiment = ', tonality_vaderSentiment)
    print('tonality_nltk = ', tonality_nltk)
    print('tonality_spacy = ', tonality_spacy)

    tonality_rezalt_local = ' '.join([tonality_TextBlob,
tonality_vaderSentiment, tonality_nltk, tonality_spacy])

    print('tonality_rezalt', tonality_rezalt_local, '\n')

    print('tonality_rezalt !!!!!', type(tonality_rezalt_local), '\n')

    print('number_Positive =', pattern_matches('Positive',
tonality_rezalt_local))
    print('number_Negative =', pattern_matches('Negative',
tonality_rezalt_local))
    print('number_Neutral =', pattern_matches('Neutral',
tonality_rezalt_local))

    number_Positive = pattern_matches('Positive', tonality_rezalt_local)
    number_Negative = pattern_matches('Negative', tonality_rezalt_local)
    number_Neutral = pattern_matches('Neutral', tonality_rezalt_local)
    tonality_return = tonality_rezalt(number_Positive, number_Negative,

```

```

number_Neutral)
    print('post ', i, ' tonality_return =', tonality_return)
    print('-----')

    result_list.append({
        'title': i,
        'href': post,
        'about': tonality_return
    })

    # print('result_list =', result_list)

print('result_list =', result_list)
df = pd.DataFrame(data=result_list)
print('df = ', df)
df.to_csv("data_plot/output.csv")
df.to_excel("data_plot/output.xlsx")

plot_frequency_tonality(result_list)

return

```

```

'''
Модуль статистичного навчання
'''

import numpy as np
import math as mt
from scipy.optimize import curve_fit
import warnings
warnings.filterwarnings('ignore')

# Коефіцієнт детермінації - оцінювання якості полінома
def r2_score(SL, Y, Text):

    # статистичні характеристики вибірки з урахуванням тренду
    numerator = 0
    denominator_1 = 0
    Y_length = len(Y)

    for i in range(Y_length):
        numerator += (SL[i] - Y[i, 0]) ** 2
        denominator_1 = denominator_1 + SL[i]

    denominator_2 = 0

    for i in range(Y_length):
        denominator_2 = denominator_2 + (SL[i] - (denominator_1 / Y_length)) ** 2

    R2_score_our = 1 - (numerator / denominator_2)
    print('-----', Text, '-----')
    print('кількість елементів вбірки=', Y_length)
    print('Коефіцієнт детермінації (ймовірність апроксимації)=', R2_score_our)

    return R2_score_our

# Коефіцієнт детермінації - оцінювання якості експоненти
def r2_score_expo(SL, Yout, Text):

    # статистичні характеристики вибірки з урахуванням тренду
    iter = len(Yout)
    numerator = 0
    denominator_1 = 0

    for i in range(iter):
        numerator = numerator + (SL[i] - Yout[i]) ** 2
        denominator_1 = denominator_1 + SL[i]

    denominator_2 = 0

    for i in range(iter):
        denominator_2 = denominator_2 + (SL[i] - (denominator_1 / iter)) ** 2

    R2_score_our = 1 - (numerator / denominator_2)
    print('-----', Text, '-----')
    print('кількість елементів вбірки=', iter)
    print('Коефіцієнт детермінації (ймовірність апроксимації)=', R2_score_our)

    return R2_score_our

```

```

# МНК згладжування
def LSM(S0):

    S0_len = len(S0)
    Yin, F = np.zeros((S0_len, 1)), np.ones((S0_len, 6))

    for i in range(S0_len): # формування структури вхідних матриць МНК
        S_value = S0[i]

        if type(S_value) == np.ndarray:
            S_value = S_value[0]
        Yin[i, 0] = float(S_value) # формування матриці вхідних даних

        for p in range(1, 6):
            F[i, p] = float(i ** p)

    C, *_ = np.linalg.lstsq(F, Yin, rcond=None)
    Yout = F @ C

    print('Регресійна модель:')
    print(f'y(t) = {C[0,0]} + {C[1,0]} * t + {C[2,0]} * t^2 + {C[3,0]} * t^3 + {C[4,0]} * t^4 + {C[5,0]} * t^5')

    return Yout

# МНК детекція та очищення АВ
def LSM_AV_Detect(S0):

    S0_len = len(S0)
    Yin, F = np.zeros((S0_len, 1)), np.ones((S0_len, 3))

    for i in range(S0_len): # формування структури вхідних матриць МНК
        S_value = S0[i]
        if type(S_value) == np.ndarray:
            S_value = S_value[0]
        Yin[i, 0] = float(S_value) # формування матриці вхідних даних
        for p in range(1, 2):
            F[i, p] = float(i ** p)

    C, *_ = np.linalg.lstsq(F, Yin, rcond=None)

    return C[1, 0]

# МНК прогнозування
def LSM_Extrapol(S0, koef):

    S0_len = len(S0)
    Yout_Extrapol = np.zeros((S0_len + koef, 1))
    Yin, F = np.zeros((S0_len, 1)), np.ones((S0_len, 3))

    for i in range(S0_len): # формування структури вхідних матриць МНК
        S_value = S0[i]
        if type(S_value) == np.ndarray:
            S_value = S_value[0]
        Yin[i, 0] = float(S_value) # формування матриці вхідних даних

```

```

        for p in range(1, 2):
            F[i, p] = float(i ** p)

C, *_ = np.linalg.lstsq(F, Yin, rcond=None)

print('Регресійна модель:')
print(f'y(t) = {C[0, 0]} + {C[1, 0]} * t + {C[2, 0]} * t^2')

for i in range(S0_len + koef):
    Yout_Extrapol[i, 0] = C[0, 0] + C[1, 0] * i + (C[2, 0]*i*i)

return Yout_Extrapol

# МНК експонента
def LSM_exponent(S0):

    S0_len = len(S0)
    Yout = np.zeros((S0_len, 1))
    Yin, F = np.zeros((S0_len, 1)), np.ones((S0_len, 4))

    for i in range(S0_len): # формування структури вхідних матриць МНК
        S_value = S0[i]
        if type(S_value) == np.ndarray:
            S_value = S_value[0]
        Yin[i, 0] = float(S_value) # формування матриці вхідних даних
        for p in range(1, 3):
            F[i, p] = float(i ** p)

    C, *_ = np.linalg.lstsq(F, Yin, rcond=None)

    a_3 = 3 * (C[3, 0] / C[2, 0])
    a_2 = (2 * C[2, 0]) / (a_3 ** 2)
    a_0 = C[0, 0] - a_2
    a_1 = C[1, 0] - (a_2 * a_3)

    print('Регресійна модель:')
    print(f'y(t) = {a_0} + {a_1} * t + {a_2} * exp({a_3} * t)')

    for i in range(S0_len):
        Yout[i, 0] = a_0
        Yout[i, 0] += a_1 * i
        Yout[i, 0] += a_2 * mt.exp(a_3 * i)

    return Yout

# Expo_scipy
def Expo_Regres (Y, bstart):

    def func_exp(x, a, b, c, d):
        print('Регресійна модель: ')
        print(f'y(t) = {c} + {d} * t + {a} * exp({b} * t)')
        return a * np.exp(b * x) + c + (d * x)

    # емпіричні коефіцієнти старта
    aStart=bstart/10
    bStart=bstart/1000
    cStart=bstart+10
    dStart=bstart/10

```

```

Y_len = len(Y)
x_data = np.ones((Y_len))
y_data = np.ones((Y_len))

for i in range(Y_len):
    x_data[i] = i
    y_data[i] = Y[i]

popt, pcov = curve_fit(func_exp, x_data, y_data, p0=(aStart, bStart, cStart,
dStart))

return func_exp(x_data, *popt)

# алгоритм -a-b фільтра
def ABF(S0):

    S0_len = len(S0)
    Yin, YoutAB = np.zeros((S0_len, 1)), np.zeros((S0_len, 1))
    T0 = 1

    for i in range(S0_len):
        S_value = S0[i]
        if type(S_value) == np.ndarray:
            S_value = S_value[0]
        Yin[i, 0] = float(S_value)

    alpha = 1 #  $2 * (2 * 1 - 1) / (1 * (1 + 1))$ 
    beta = 12 #  $(6 / 1) * (1 + 1)$ 

    y_prev = Yin[0, 0]
    y_speed = (Yin[1, 0] - y_prev) / T0
    y_extra = y_prev + y_speed

    YoutAB[0, 0] = (alpha + 1) * y_prev

    for i in range(1, S0_len):
        y_curr = Yin[i, 0]

        YoutAB[i, 0] = y_extra + alpha * (y_curr - y_extra)

        y_speed += (beta / T0) * (y_curr - y_extra)

        y_extra = YoutAB[i, 0] + y_speed
        alpha =  $(2 * (2 * i - 1)) / (i * (i + 1))$ 
        beta =  $6 / (i * (i + 1))$ 

    return YoutAB

# Виявлення АВ за алгоритмом medium
def Sliding_Window_AV_Detect_medium(S0, n_Wind, Q):

    # параметри циклів
    j_Wind, S_window = mt.ceil(len(S0) - n_Wind)+1, np.zeros((n_Wind))

    # еталон
    j = 0

```

```

for i in range(n_Wind):
    S_window[i] = S0[(j + i)]

# ковзне вікно
for j in range(j_Wind):
    for i in range(n_Wind):
        l=(j+i)
        S_window[i] = S0[l]

    mS = np.median(S_window)
    dS = np.var(S_window)
    scvS = mt.sqrt(dS)

# детекція та заміна АВ
    if scvS > (Q * scvS):
        # детектор виявлення АВ
        S0[j + n_Wind - 1] = mS

return S0

# Виявлення АВ за МНК
def Sliding_Window_AV_Detect_LSM (S0_input, Q, sliding_window_n):

    # параметри циклів
    S0_len = len(S0_input)
    j_Wind, S_window = mt.ceil(S0_len - sliding_window_n) + 1,
    np.zeros((sliding_window_n))

    # еталон
    Speed_standart = LSM_AV_Detect(S0_input)
    Yout_S = LSM(S0_input)

    # ковзне вікно
    for j in range(j_Wind):
        for i in range(sliding_window_n):
            S_window[i] = S0_input[(j + i)]

        scvS = mt.sqrt(np.var(S_window))

        # детекція та заміна аномальних вимірів
        speed_s = abs(Speed_standart * mt.sqrt(S0_len))
        speed = abs(Q * Speed_standart * mt.sqrt(sliding_window_n) * scvS)
        if speed > speed_s:
            # детектор виявлення АВ
            S0_input[j + sliding_window_n - 1] = Yout_S[j + sliding_window_n - 1,
0]

    return S0_input

# ----- Виявлення АВ за алгоритмом sliding window -----
def Sliding_Window_AV_Detect_sliding_wind (S, sliding_window_n):
    # параметри циклів
    S_len = len(S)
    j_Wind, S_window, Midi = mt.ceil(S_len - sliding_window_n) + 1,
    np.zeros((sliding_window_n)), np.zeros((S_len))

```

```
# ковзне вікно
for j in range(j_Wind):
    for i in range(sliding_window_n):
        S_window[i] = S[j + i]

    Midi[j + sliding_window_n - 1] = np.median(S_window)

# очищена вибірка
S0_Mid = Midi.copy()

S0_Mid[:sliding_window_n] = S[:sliding_window_n]

return S0_Mid
```

```

'''
Модуль генерації синтезованих / модельних даних
'''

import numpy as np
import math as mt
import matplotlib.pyplot as plt

# рівномірний закон розподілу номерів АВ в межах вибірки
def randomAM(n, lim, nAV):

    SAV, S = np.zeros((nAV)), np.zeros((n))

    S = np.random.randint(0, lim, size=n)

    mS = np.median(S)
    dS = np.var(S)
    scvS = mt.sqrt(dS)

    # генерація номерів АВ за рівномірним законом
    for i in range(nAV):
        SAV[i] = mt.ceil(np.random.randint(1, lim))
    print('номери АВ: SAV=', SAV)
    print('----- статистичны характеристики РІВНОМІРНОГО закону розподілу ВВ ----')
    print('матиматичне сподівання ВВ=', mS)
    print('дисперсія ВВ =', dS)
    print('СКВ ВВ=', scvS)
    print('-----')

# гістограма закону розподілу ВВ
plt.hist(S, bins=20, facecolor="blue", alpha=0.5)
plt.show()

return SAV

# нормальний закон розподілу ВВ
def randoNORM(dm, dsig, iter):

    S = np.random.normal(dm, dsig, iter) # нормальний закон розподілу ВВ з
    вибіркою ебемом iter та параметрами: dm, dsig
    mS = np.median(S)
    dS = np.var(S)
    scvS = mt.sqrt(dS)

    print('----- статистичны характеристики НОРМАЛЬНОЇ похибки вимірів ----')
    print('матиматичне сподівання ВВ=', mS)
    print('дисперсія ВВ =', dS)
    print('СКВ ВВ=', scvS)
    print('-----')

    # гістограма закону розподілу ВВ
    plt.hist(S, bins=20, facecolor="blue", alpha=0.5)
    plt.show()

    return S

# модель ідеального тренду
def Model(n):

```

```

S0=np.zeros((n))

for i in range(n):
    S0[i] = (0.0000005*i*i)      # квадратична модель реального процесу
return S0

# модель ідеального тренду синтезованих даних
def Model_real(n):

    S0=np.zeros((n))

    for i in range(n):
        # S0[i] = (0.0000005*i*i)      # квадратична модель реального процесу
        S0[i] = 28.03423083418919 + (0.004773157735137576 * i) +
(0.0008240042970011226 * (i ** 2)) - (
        3.204450289838111e-06 * (i ** 3)) + (7.543074486030252e-10 *
(i ** 4)) + (
        6.87528764332385e-12 * (i ** 5))

    return S0

# модель виміру (тренду) з нормальний шумом
def Model_NORM(SN, S0N):
    SV = S0N + SN

    return SV

# модель виміру (тренду) з нормальний шумом + АНОМАЛЬНІ ВИМІРИ
def Model_NORM_AV(S0, SV, nAV, Q_AV, dm, dsig, SAV):

    SV_AV = SV
    SSAV = np.random.normal(dm, (Q_AV * dsig), nAV) # аномальна випадкова
похибка з нормальним законом

    for i in range(nAV):
        k = int(SAV[i])
        SV_AV[k] = S0[k] + SSAV[i]      # аномальні вимірів з рівномірно
розподіленими номерами

    return SV_AV

```

```

'''
Модуль оптимізаційної згорткової нейронної мережі для обробки / прогнозування та
екстраполяції часових рядів:
convolutional artificial neural network (CNN)

Класичний функціонал:
1. Підготовка даних:
    ділянка навчання / прогнозування;
    розбиття ділянки навчання на тренувальну / тестову вибірки в форматі
навчальних пар;
    нормалізація даних
2. Навчання та контроль динаміки метрик;
3. Прогнозування;
4. Екстраполяція поза межами ділянки навчання.

Інноваційний функціонал:
1. Генерація опорної вибірки на екстраполяцію з МНК та шумом.
2. Оптимізація структури моделі;
3. Оптимізаційна модель вхідного прошарку.
'''

import numpy as np
import tensorflow as tf
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error
import warnings
warnings.filterwarnings('ignore')

def file_parsing(File_name="data/Oschadbank (USD).xls", Data_name="Купівля"):
    '''
    Парсинг файлу вхідних даних
    '''

    d = pd.read_excel(File_name)
    S_real = d[[Data_name]].values.reshape(-1, )

    return S_real

class MyModel(tf.keras.Model):
    '''
    Створення моделі нейромережі
    '''

    # def __init__(self, n_stocks, n_neurons, sigma):
    def __init__(self, n_neurons, sigma):
        super(MyModel, self).__init__()
        weight_initializer = tf.keras.initializers.VarianceScaling(seed=7,
scale=sigma, mode="fan_avg",
distribution="uniform")
        bias_initializer = tf.zeros_initializer()
        self.hidden_layers = [tf.keras.layers.Dense(n_neurons[i],
activation=tf.nn.relu,
kernel_initializer=weight_initializer,

```

```

bias_initializer=bias_initializer)
        for i in range(len(n_neurons))]
        self.out = tf.keras.layers.Dense(1,
kernel_initializer=weight_initializer,
        bias_initializer=bias_initializer)

    def call(self, inputs):
        x = inputs
        for layer in self.hidden_layers:
            x = layer(x)
        return self.out(x)

def convolutional_neural_network_model(batch_size, epochs, real_data_in,
data_predicted, n_neurons, k_expo, sigma):
    '''
        Головні процеси зі згортковою неймережею
    '''

    def data_set_in(real_data_in, batch_size, k_expo):
        '''
            Формування потоку вхідних даних із залишком навчальної вибірки для прогнозу
            == екстраполяція
        '''

        # короткий датасет для навчання без прогнозу k_expo
        n_real_data_in = len(real_data_in)
        real_data = np.zeros((n_real_data_in - k_expo))

        for f in range(n_real_data_in - k_expo):
            real_data[f] = real_data_in[f]

        # короткий датасет для навчання з batch_size
        m = len(real_data)
        n = batch_size
        train = np.zeros((m, n))
        test = train.copy()

        for j in range(m):
            for i in range(n):
                train[j, i] = (real_data[j])
                test[j, i] = (real_data[j])

        return m, n, real_data, train, test

    def data_plot(data, text):
        '''
            Візуалізація потоку даних
        '''

        plt.figure(figsize=(15, 5))
        plt.plot(data)
        plt.title(text)
        plt.show()

        return

    def metrics_plot(history, title, text):

```

```

'''
Візуалізація динаміки метрик
'''

plt.figure(figsize=(15, 5))
plt.plot(history.history[title])
plt.title('metrics model')
plt.ylabel(text)
plt.xlabel('Epoch')
plt.show()

return

def predict_plot(y_test, predicted, text):
    '''
    Візуалізація прогнозованих даних
    '''

    plt.figure(figsize=(15, 5))
    plt.plot(np.arange(y_test.shape[0]), y_test, color='blue', label='Real')
    plt.plot(np.arange(y_test.shape[0]), predicted, color='red',
label='Predicted')
    plt.title(text)
    plt.legend()
    plt.show()

    return

def data_norm(train_data, test_data):
    '''
    Нормалізація вхідних даних
    '''

    scaler = MinMaxScaler(feature_range=(-1, 1))
    train_data_norm = scaler.fit_transform(train_data)
    test_data_norm = scaler.transform(test_data)

    return train_data_norm, test_data_norm, scaler

def data_norm_opt(train_data, test_data, m, n):
    '''
    Нормалізація вхідних даних
    '''

    scaler = MinMaxScaler(feature_range=(-1, 1))
    train_data_norm = scaler.fit_transform(train_data)
    test_data_norm = scaler.transform(test_data)

    # модифікація вхідного прохарку до оптимізаційного рівня

    def optimization_network(letter, delta):
        '''
        Оптимізація за Вороніним додана до згортки
        '''

```

```

        x_letter = (1 / (1 / 0.9) * (1 - letter - delta))

        return x_letter

    def optimization_network_norm(letter, delta):
        '''
        Оптимізація за Вороніним додана до згортки - враховано критерії max,
min -> min
        1 -> max = 1/(1+delta) -> min,
        0 -> min
        Користь в наближенні виходу до 1, 0. + швидкість навчання та
ймовірність ідентифікації - покращено
        '''

        if (letter > 0.6):
            x_letter = (1 / (1 / 0.9) * (1 - (1 / (letter - delta))))
        else:
            x_letter = (1 / (1 / 0.9) * (1 - letter + delta))

        return x_letter

    # Модифікація вхідних даних за Вороніним
    for j in range(m):
        for i in range(n):
            train_data[j, i] = optimization_network_norm(train_data[j, i], 0)

    return train_data_norm, test_data_norm, scaler

def dataset_formation(train_data, test_data):
    '''
    Формування навчального / тестового датасет із навчальних пар
    '''

    X_train, y_train = train_data[:, 1:], train_data[:, 0]
    X_test, y_test = test_data[:, 1:], test_data[:, 0]

    return X_train, y_train, X_test, y_test

def network_training(model, X_train, y_train, epochs, batch_size):
    '''
    Тренування штучної нейронної мережі та оцінювання метрик
    Цікавий опис процесів: https://habr.com/ru/articles/482126/
    Функція втрат loss='mse' - мінімізується в процесі навчання найбільш
вдала, оскільки передає природу часових рядів
    Метрики оцінки моделі - metrics:
    R-squared (R2) Score: (r2_score) - ймовірність апроксимації
    Mean Absolute Error (MAE) - середнє абсолютне відхилення
    Mean Squared Error (MSE) - середнє квадратичне відхилення
    loss - втрати - різниця вхід / вихід
    Оптимізатор та функція втрат - ті, що було в оригіналі
    optimizer = tf.keras.optimizers.Adam()
    model.compile(loss='MeanSquaredError', optimizer=optimizer)
    '''

    model.compile(optimizer=tf.keras.optimizers.Adam(0.01),

```

```

        loss='mse', # mean squared error
        metrics=['r2_score', 'mae', 'mse'])

    # тренування моделі
    history = model.fit(X_train, y_train, epochs=epochs,
batch_size=batch_size)

    # побудова динаміки метрик
    metrics_plot(history, 'r2_score', 'metrics: r2_score')
    metrics_plot(history, 'mae', 'metrics: mae')
    metrics_plot(history, 'mse', 'metrics: mse')
    metrics_plot(history, 'loss', 'metrics: loss')

    return model

def network_predict(model_trained):
    '''
    Прогнозування часового ряду на ділянці навчання
    '''

    # прогнозування
    predicted = model_trained.predict(X_test)

    # візуалізація результатів
    predict_plot(y_test, predicted, 'Real vs Predicted normalized')

    return predicted

def metrics_predict(real_data, predicted, text):
    '''
    Оцінювання метрик результатів прогнозування з нейромережею
    '''

    # Обчислення метрик якості моделі на прогнозуванні
    mae = mean_absolute_error(real_data, predicted)
    mse = mean_squared_error(real_data, predicted)
    rmse = np.sqrt(mse)
    r2 = r2_score(real_data, predicted)

    print('\n')
    print(text)
    print("Mean Absolute Error (MAE): " + mae)
    print("Mean Squared Error (MSE): " + mse)
    print("Root Mean Squared Error (RMSE): " + rmse)
    print("R-squared (R2) Score: " + r2)

    return

def inverse_normalization(m, n, predicted, real_data):
    '''
    Ренормалізація - відновлення масштабу даних
    '''

    # повернення в оригінальну область параметрів
    predicted_inverse = np.zeros((m, n))
    for j in range(m):
        for i in range(n):
            predicted_inverse[j, i] = (predicted[j])

```

```

# масштабуємо реальні значення
predicted_n = scaler.inverse_transform(predicted_inverse)

# повертаємо масштабовані до оригінальної шкали дані до необхідної
структури
for j in range(m):
    predicted[j] = predicted_n[j, 0]

# обчислення метрик якості моделі на прогнозуванні
metrics_predict(real_data, predicted, 'Метрики якості моделі відновлених
даних')

# візуалізація результатів
predict_plot(real_data, predicted, 'Real vs Predicted original')

return predicted

def network_extrapolation(real_data_etalon, real_data_in, model, batch_size):
    '''
    Екстраполяція параметрів за межами інтервалу навчання

    По суті традиційно відбувається відтворення та екстраполяція даних в
    оригіналі.
    Для прогнозування потрібна вибірка, схожа на реальні дані.
    Формування опорної вибірки за межами інтервалу спостереження є окрема
    проблема.
    Тоді можливо брати МНК та генерувати опорну / тестову вибірку з моделі та
    інше.....
    Так це і реалізовано зараз
    '''

    m = len(real_data_in)
    n = batch_size

    # ----- блок оригінальних даних -----
    train_data = np.zeros((m, n))
    test_data = train_data.copy()
    for j in range(m):
        for i in range(n):
            train_data[j, i] = (real_data_in[j])
            test_data[j, i] = (real_data_in[j])

    X_test, y_test = test_data[:, 1:], test_data[:, 0]

    # ----- блок нормованих даних -----
    # закоментувати 2 рядка для прогнозування в оригінальних даних
    train_data_norm, test_data_norm, scaler = data_norm(train_data,
test_data)
    X_train, y_train, X_test, y_test = dataset_formation(train_data_norm,
test_data_norm)

    predicted = model.predict(X_test)

    # обчислення метрик якості моделі на прогнозуванні
    metrics_predict(y_test, predicted, 'Метрики якості моделі прогнозованих
даних')

    # візуалізація результатів
    predict_plot(y_test, predicted, 'Real vs Predicted extrapolation')

    # Відновлення масштабу часового ряду на ділянці навчання

```

```

        predicted_original = inverse_normalization(m, n, predicted, real_data_in)

        predict_plot(real_data_etalon, predicted_original, 'Real vs Predicted
original')

    return predicted_original

'''
Головні виклики для convolutional_neural_network_model
'''

# Потік вхідних даних із залишком навчальної вибірки для прогнозу ==
екстраполяція
m, n, real_data, train_data, test_data = data_set_in(real_data_in,
batch_size, k_expo)

# Візуалізація
data_plot(train_data, 'Вхідні параметри для тренування')
data_plot(test_data, 'Вхідні параметри для тестування')

# Нормалізація даних
train_data_norm, test_data_norm, scaler = data_norm(train_data, test_data)
# train_data_norm, test_data_norm, scaler = data_norm_opt(train_data,
test_data, m, n)

# Датасет з x та y - формування навчальних / тестових пар
X_train, y_train, X_test, y_test = dataset_formation(train_data_norm,
test_data_norm)

# Створення екземпляру класу моделі нейромережі
model = MyModel(n_neurons, sigma)

# Тренування штучної нейронної мережі та оцінювання метрик
model_trained = network_training(model, X_train, y_train, epochs, batch_size)

# Прогнозування часового ряду на ділянці навчання
predicted = network_predict(model_trained)

# Відновлення масштабу часового ряду на ділянці навчання
predicted_original = inverse_normalization(m, n, predicted, real_data)

# Екстраполяція параметрів за межами інтервалу навчання
predicted_extrapolation_original = network_extrapolation(real_data_in,
data_predicted, model, batch_size)

metrics_predict(real_data_in, predicted_extrapolation_original, 'Метрики
якості екстраполяції оригінал даних')

return predicted_original, predicted_extrapolation_original

```

```

'''
Модуль визначення інтегрованої оцінки
'''

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import pylab
from data_utility import msx_max, max_and_delta_normal, hurst_index,
hurst_index_my
from model_scor import model_integro_scor
from ts_statistical_learning import LSM, r2_score
import warnings
warnings.filterwarnings('ignore')

def dataframe_parser_and_plot(path_data, filename, ind_x, k):
    '''
    Парсинг вхідних даних з файлу в pandas dataframe
    '''

    dateparse = lambda x: pd.to_datetime(x, dayfirst=True)
    d = pd.read_excel((path_data + filename), index_col=ind_x,
date_parser=dateparse)
    ts = d.iloc[:, k]

    print('d = ', d)
    print('ts= ', ts)

    plt.figure(figsize=(12, 3))
    plt.plot(ts, label='Original Time Series')
    plt.legend()
    plt.show()

    return ts

def datanparray_parsing_and_plot(filename, d_name):
    '''
    Парсинг вхідних даних з файлу в numpy array - побудова гістограми
    '''

    d = pd.read_excel(filename)
    for name, series in d[[d_name]].items():
        print(series)

    plt.clf()
    plt.plot(series)
    plt.show()

    return series

```

```

def restoration_of_measurement_gaps(data_in_real):
    '''
    Наївне відновлення пропусків шляхом заміни попередніми даними
    '''

    data_in_result = data_in_real.copy()

    for i in range(len(data_in_result)):
        if data_in_result[i] == 0:
            data_in_result[i] = data_in_result[i - 1]

    return data_in_result

def data_plot_simslary(data_1, data_2, data_3, text_1, text_2, text_3):
    '''
    Відображення часових рядів
    '''

    plt.figure(figsize=(9, 3))
    plt.plot(data_1, label=text_1, color='r')
    plt.plot(data_2, label=text_2, color='g')
    plt.plot(data_3, label=text_3, color='b')

    plt.legend()
    plt.show()

    return

def data_plot_3d(data_1, data_2, data_3, text_1, text_2, text_3):
    '''
    3d візуалізатор реальних даних
    '''

    xg = np.zeros((len(data_1)))
    for i in range(len(data_1)):
        xg[i] = i

    fig = pylab.figure()
    ax = fig.add_subplot(projection='3d')
    clr = ['g', 'r', 'b']
    ax.bar(xg, data_1, 1, zdir='y', color=clr[0])
    ax.bar(xg, data_2, 2, zdir='y', color=clr[1])
    ax.bar(xg, data_3, 3, zdir='y', color=clr[2])
    ax.set_xlabel('X Time')
    ax.set_ylabel('Y Data')
    ax.set_zlabel('Z quantity')

    pylab.show()

    return

def nparray_1d_to_3d(data_np_1_restoration_norm, data_np_2_restoration_norm,
data_np_3_restoration_norm):
    '''
    Трансформація масивів: 1d to 3d
    '''

```

```

'''
    scor_Nd_ndarray = np.concatenate((np.expand_dims(data_np_1_restoration_norm,
axis=0),
                                     np.expand_dims(data_np_2_restoration_norm,
axis=0),
                                     np.expand_dims(data_np_3_restoration_norm,
axis=0)), axis=0)

    return scor_Nd_ndarray

if __name__ == '__main__':

    # ----- вхідні дані -----#

    # # dataframe_parser: 1 - Купівля, 2 - Продаж, 3 - КурсНБУ
    # ts_1 = dataframe_parser_and_plot('data/', 'Oschadbank (USD).xls', 'Дата',
1)
    # ts_2 = dataframe_parser_and_plot('data/', 'Oschadbank (USD).xls', 'Дата',
2)
    # ts_3 = dataframe_parser_and_plot('data/', 'Oschadbank (USD).xls', 'Дата',
1)

    # datanparray_parsing
    data_np_1 = datanparray_parsing_and_plot('data/Oschadbank (USD).xls',
'Купівля')
    data_np_2 = datanparray_parsing_and_plot('data/Oschadbank (USD).xls',
'Продаж')
    data_np_3 = datanparray_parsing_and_plot('data/Oschadbank (USD).xls',
'КурсНБУ')

    # візуалізація вхідних даних
    data_plot_3d(data_np_1, data_np_2, data_np_3, 'Купівля', 'Продаж', 'КурсНБУ')
    data_plot_simslary(data_np_1, data_np_2, data_np_3, 'Купівля', 'Продаж',
'КурсНБУ')

    # ----- препроцесінг із відновленням -----#

    # відновлення пропусків
    data_np_1_restoration = restoration_of_measurement_gaps(data_np_1)
    data_np_2_restoration = restoration_of_measurement_gaps(data_np_2)
    data_np_3_restoration = restoration_of_measurement_gaps(data_np_3)

    # візуалізатор 3d
    data_plot_3d(data_np_1_restoration, data_np_2_restoration,
                 data_np_3_restoration, 'Купівля', 'Продаж', 'КурсНБУ')
    # візуалізатор 2d
    data_plot_simslary(data_np_1_restoration, data_np_2_restoration,
                      data_np_3_restoration, 'Купівля', 'Продаж', 'КурсНБУ')

    # Нормалізація
    delta = 0.3
    global_max = msx_max(data_np_1, data_np_2, data_np_3)
    data_np_1_restoration_norm = max_and_delta_normal(data_np_1_restoration,
delta, global_max)
    data_np_2_restoration_norm = max_and_delta_normal(data_np_2_restoration,
delta, global_max)
    data_np_3_restoration_norm = max_and_delta_normal(data_np_3_restoration,

```

```

delta, global_max)

# візуалізатор 3d
data_plot_3d(data_np_1_restoration_norm, data_np_2_restoration_norm,
              data_np_3_restoration_norm, 'Купівля', 'Продаж', 'КурсНбу')
# візуалізатор 2d
data_plot_simslary(data_np_1_restoration_norm, data_np_2_restoration_norm,
                   data_np_3_restoration_norm, 'Купівля', 'Продаж',
'КурсНбу')

# ----- аналіз на hurst_index -----#
print('hurst_index ВІДНОВЛЕНИХ вхідних даних')
hurst_index(data_np_1_restoration)
hurst_index(data_np_2_restoration)
hurst_index(data_np_3_restoration)
print('hurst_index НОРМАЛІЗОВАНИХ вхідних даних')
hurst_index(data_np_1_restoration_norm)
hurst_index(data_np_2_restoration_norm)
hurst_index(data_np_3_restoration_norm)
print('Альтернативний херст інтегрованої оцінки після МНК на оригінальних
даних =')
hurst_exponent = hurst_index_my(data_np_1_restoration_norm)
print(f"Hurst Exponent: {hurst_exponent}")
hurst_exponent = hurst_index_my(data_np_2_restoration_norm)
print(f"Hurst Exponent: {hurst_exponent}")
hurst_exponent = hurst_index_my(data_np_3_restoration_norm)
print(f"Hurst Exponent: {hurst_exponent}")

# ----- багатокритеріальний аналіз без препроцесінгу -----#
# трансформація масивів 1d to 3d
print('інтегрована оцінка із даними без МНК')
scor_Nd_narray = ndarray_1d_to_3d(data_np_1_restoration_norm,
                                  data_np_2_restoration_norm,
data_np_3_restoration_norm)
# інтегрована оцінка
integro = model_integro_scor(scor_Nd_narray, data_np_1_restoration_norm,
'data/integro_scor.xlsx')
print('hurst_index інтегрованої оцінки вхідних нормованих даних')
hurst_index(integro)
print('Альтернативний херст інтегрованої оцінки після МНК на оригінальних
даних =')
hurst_exponent = hurst_index_my(integro)
print(f"Hurst Exponent: {hurst_exponent}")

# Побудова МНК тренду для інтегрованої оцінки
integro_LSM_1 = LSM(integro)

plt.title("Multi-criteria integrated Scor")
plt.xlabel("k = t")
plt.ylabel("scor")
plt.grid(which='major')
plt.plot(integro_LSM_1)
plt.show()
print('Альтернативний херст інтегрованої оцінки після МНК на оригінальних
даних =')
hurst_exponent = hurst_index_my(integro_LSM_1)
print(f"Hurst Exponent: {hurst_exponent}")

# ----- багатокритеріальний аналіз з препроцесінгом -----#

```

```

print('інтегрована оцінка із даними після МНК')
data_np_1_restoration_LSM = LSM(data_np_1_restoration)
data_np_2_restoration_LSM = LSM(data_np_2_restoration)
data_np_3_restoration_LSM = LSM(data_np_3_restoration)

print('r2_score після МНК')
r2_score(data_np_1_restoration, data_np_1_restoration_LSM, 'Купівля')
r2_score(data_np_2_restoration, data_np_2_restoration_LSM, 'Продаж')
r2_score(data_np_3_restoration, data_np_3_restoration_LSM, 'КурсНбу')

# Нормалізація
delta = 0.3
global_max = msx_max(data_np_1_restoration_LSM, data_np_2_restoration_LSM,
data_np_3_restoration_LSM)
data_np_1_restoration_norm_LSM =
max_and_delta_normal(data_np_1_restoration_LSM, delta, global_max)
data_np_2_restoration_norm_LSM =
max_and_delta_normal(data_np_2_restoration_LSM, delta, global_max)
data_np_3_restoration_norm_LSM =
max_and_delta_normal(data_np_3_restoration_LSM, delta, global_max)

data_plot_simslary(data_np_1_restoration_norm_LSM,
data_np_2_restoration_norm_LSM,
data_np_3_restoration_norm_LSM, 'Купівля', 'Продаж',
'КурсНбу')
print('hurst_index нормованих та МНК оброблених вхідних даних')
hurst_index(data_np_1_restoration_norm_LSM)
hurst_index(data_np_2_restoration_norm_LSM)
hurst_index(data_np_3_restoration_norm_LSM)

print('Альтернативний херст integro_LSM_LSM =')
hurst_exponent = hurst_index_my(data_np_1_restoration_norm_LSM)
print(f"Hurst Exponent: {hurst_exponent}")
hurst_exponent = hurst_index_my(data_np_2_restoration_norm_LSM)
print(f"Hurst Exponent: {hurst_exponent}")
hurst_exponent = hurst_index_my(data_np_3_restoration_norm_LSM)
print(f"Hurst Exponent: {hurst_exponent}")

# Інтегрований скор МНК оброблених даних
scor_Nd_ndarray_LSM = ndarray_1d_to_3d(data_np_1_restoration_norm_LSM,
data_np_2_restoration_norm_LSM,
data_np_3_restoration_norm_LSM)
integro_LSM = model_integro_scor(scor_Nd_ndarray_LSM,
data_np_1_restoration_norm_LSM, 'data/integro_scor_LSM.xlsx')
print('hurst_index інтегрованої оцінки вхідних нормованих даних та оброблених
МНК')
hurst_index(integro_LSM)
print('Альтернативний херст integro_LSM_LSM =')
hurst_exponent = hurst_index_my(integro_LSM)
print(f"Hurst Exponent: {hurst_exponent}")

print('hurst_index інтегрованої оцінки ПІСЛЯ МНК вхідних нормованих даних та
оброблених МНК')
integro_LSM_LSM = LSM(integro_LSM)
plt.title("Multi-criteria integrated Scor")
plt.xlabel("k = t")
plt.ylabel("scor")
plt.grid(which='major')
plt.plot(integro_LSM_LSM)
plt.show()
print('Альтернативний херст integro_LSM_LSM =')
hurst_exponent = hurst_index_my(integro_LSM_LSM)
print(f"Hurst Exponent: {hurst_exponent}")

```

ДОДАТОК В

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА

Наукові праці в яких опубліковано основні наукові результати дисертації:

1. Algorithms of Statistical Anomalies Clearing for Data Science Applications - Oleksii Pysarchuk, Yurii Mironov, Illya Pysarchuk, Danylo Baran. // System research and information technologies. – No. 1 (2023), P.78-84. <https://doi.org/10.20535/SRIT.2308-8893.2023.1.06>, <http://journal.iasa.kpi.ua/article/view/260175>. Стаття Scopus, категорії A.
2. O. Pysarchuk, T. Andreieva, O. Grinenko, D. Baran. Multi-factor forecasting of statistical trends for data science problems // System Research & Information Technologies, 2024, No 2, P.21-34. <https://doi.org/10.20535/SRIT.2308-8893.2024.2.02> <http://journal.iasa.kpi.ua/article/view/287739>. Стаття Scopus, категорії A.
3. O. O Pysarchuk, O. A. Tuhanskykh, D. R. Baran (2025). Method for determining the structure of nonlinear models for time series processing. Radio Electronics, Computer Science, Control, 1, pp. 53-62 DOI: 10.15588/1607-3274-2025-1-5, <https://ric.zp.edu.ua/article/view/324253>, ISSN: 2313-688X, 1607-3274 | <https://ric.zp.edu.ua/article/view/324253>; Стаття Web of Science, категорії A.
4. Baran D., Pysarchuk O., Pavlova S (2025). The method of adaptation of the parameters of algorithms for the detection and cleaning of a statistical sample from anomalies for data science problems. Radio Electronics, Computer Science, Control, 3, pp. 37-44 DOI: 10.15588/1607-3274-2025-3-4, <https://ric.zp.edu.ua/article/view/339011>, ISSN: 2313-688X, 1607-3274 | <https://ric.zp.edu.ua/article/view/339011> Стаття Web of Science, категорії A.
5. Baran D. Pysarchuk O., Pysarchuk I., Vasylieva M. (2025). Multi-criteria mathematical model of credit scoring in data science problems. System research and information technologies, No. 3 (2025), pp. 99-112. <http://journal.iasa.kpi.ua/article/view/308786>. Стаття Scopus, категорії A.

6. Баран Д.Р., Писарчук О.О., Корочкін О.В., Визначення порядку поліноміальної моделі для побудови лінії тренду в задачах Data Science // Проблеми інформатизації та управління, 3(71), 2022/ – С.35-40.
<https://doi.org/10.18372/2073-4751.71.17001>
<https://jrn1.nau.edu.ua/index.php/PIU/article/view/17001> Стаття фахова, категорії Б.
7. Писарчук, О., Баран, Д., & Туганських, О. (2023). Методика статистичного навчання параметрів експоненційної математичної моделі. Measuring and computing devices in technological processes, (3), 224–227.
<https://doi.org/10.31891/2219-9365-2023-75-26>
<https://vottp.khmnu.edu.ua/index.php/vottp/article/view/177> Стаття фахова, категорії Б.
8. Baran D., Pysarchuk O. (2025). Mathematical Model of Clustering of Informational Messages with Indicators of Activity for the Information Content by Tone and Areas of Society Activity. Information, Computing and Intelligent systems, 7, pp 39-48 DOI: <https://doi.org/10.20535/2786-8729.6.2025.339127>,
<https://itvisnyk.kpi.ua/article/view/339127> Стаття фахова, категорії Б.
9. Pysarchuk O., Baran D. (2025). Optimization neural network for time series processing. Information, Computing and Intelligent systems, 6, pp 230-241 DOI: <https://doi.org/10.20535/2786-8729.7.2025.341480>,
<https://itvisnyk.kpi.ua/article/view/341480> Стаття фахова, категорії Б.

Апробація наукових результатів дисертації:

10. Konrad, T., Pysarchuk, O., Noga, H., Gajewska, A., Volkogon, V., Baran, D. (2024). Method of Multifactor Analysis of the Alternatives Using the Example of Multi-criterial Selection of the Optimal Cargo Transportation Route. In: Yang, XS., Sherratt, S., Dey, N., Joshi, A. (eds) Proceedings of Ninth International Congress on Information and Communication Technology. ICICT 2024 2024. Lecture Notes in Networks and Systems, vol 1002. Springer, Singapore.

https://doi.org/10.1007/978-981-97-3299-9_6

https://link.springer.com/chapter/10.1007/978-981-97-3299-9_6#citeas

11. Tuhanskyk, O., Baran, D., Pysarchuk, O. (2024). Method for Statistical Evaluation of Nonlinear Model Parameters in Statistical Learning Algorithms. In: Yang, X.S., Sherratt, S., Dey, N., Joshi, A. (eds) Proceedings of Ninth International Congress on Information and Communication Technology. ICICT 2024 2024. Lecture Notes in Networks and Systems, vol 1013. Springer, Singapore. https://doi.org/10.1007/978-981-97-3559-4_21
https://link.springer.com/chapter/10.1007/978-981-97-3559-4_21
12. Pysarchuk, O., Tuhanskykh, O., Baran, D. (2026). Structural Adaptation of Nonlinear Models in Time Series Processing Problems. In: Zgurovsky, M., Pankratova, N. (eds) System Analysis and Data Mining. Studies in Systems, Decision and Control, vol 609. Springer, Cham. ISBN 978-3-031-97529-5 https://doi.org/10.1007/978-3-031-97529-5_3/ Розділ монографії Scopus.
13. Метод аналітичного визначення параметрів нелінійних моделей та статистичною вибіркою вимірів – Баран Д., Туганських О., Писарчук О. // Міжнародна науково-практична конференція молодих вчених та студентів «Інженерія програмного забезпечення і передові інформаційні технології» (SoftTech-2023) Кафедра ІІІ ФІОТ, КПІ ім. Ігоря Сікорського, 2023, <https://ela.kpi.ua/handle/123456789/59987> <https://ela.kpi.ua/items/f02db5cb-3489-4c68-941e-179fdca13cbe/full>
14. Algorithms for purification of statistical data samples from anomalies for Data Science tasks – Baran D., Pysarchuk O. // The International Conference on Security, Fault Tolerance, Intelligence” (ICSFTI2022 Online).
15. Oleksii Pysarchuk, Oleksandr Tuhanskykh, and Danylo Baran Structural Adaptation of Nonlinear Models in Time Series Processing Problems // SAIC'2024 (IEEE International Conference on System Analysis & Intelligent Computing), http://saic.ieee.org.ua/?page_id=96
16. Hrishchenko, K., Pysarchuk, O., Baran, D. (2026). Divide and Conquer Method for Constrained Programming Applied to Large Scale Production Scheduling

Problems. In: Choudrie, J., Tuba, E., Perumal, T., Joshi, A. (eds) ICT for Intelligent Systems. ICTIS 2025. Smart Innovation, Systems and Technologies, vol 124. Springer, Singapore. https://link.springer.com/chapter/10.1007/978-981-95-1353-6_9

17. Oleksii Pysarchuk, Oleksandr Tuhanskykh, and Danylo Baran Methodology for statistical learning of nonlinear model parameters with the integral form of ols in the scheme of differential-nontaylor transformations // Security, Fault Tolerance, Intelligence: proceedings of the VIII International Conference ICSFTI2025, Kyiv, Ukraine, may 14, june 30, 2025. – Kyiv : Igor Sikorsky Kyiv Polytechnic Institute, publishing house “Polytechnica”, 2025. <https://icsfti-proc.kpi.ua/article/view/338732>