**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**Чорноморський національний університет імені Петра Могили**

**Факультет комп'ютерних наук**

**Кафедра інженерії програмного забезпечення**

ДОПУЩЕНО ДО ЗАХИСТУ

Завідувач кафедри інженерії
програмного забезпечення

_____ Євген ДАВИДЕНКО

«___» _____ 2025 р.

КВАЛІФІКАЦІЙНА РОБОТА
НА ЗДОБУТТЯ ОСВІТНЬОГО СТУПЕНЯ МАГІСТРА

# ДОСЛІДЖЕННЯ ЦІН НА БІРЖІ ЗАСОБАМИ ШТУЧНОГО ІНТЕЛЕКТУ

Спеціальність 121 Інженерія програмного забезпечення
Освітня програма «Інженерія програмного забезпечення»

**Здобувач** _____ **Дмитро ДЗУНДЗА**

«__» _____ 20__ р.

**Керівник роботи**
PhD,
старший викладач _____ **Ігор КАНДИБА**

«__» _____ 20__ р.

**Миколаїв – 2025**

**Завдання на виконання кваліфікаційної роботи**

Чорноморський національний університет імені Петра Могили

| | |
|---|---|
| Факультет | Комп'ютерних наук |
| Кафедра | Інженерії програмного забезпечення |
| Рівень вищої освіти | Другий (магістерський) |
| Освітній ступінь | Магістр |
| Спеціальність | 121 Інженерія програмного забезпечення |
| Освітня програма | Інженерія програмного забезпечення |

ЗАТВЕРДЖУЮ
Завідувач кафедри інженерії
програмного забезпечення
_____ Євген ДАВИДЕНКО
«___» _____ 2025 р.

**ЗАВДАННЯ**
**на кваліфікаційну магістерську роботу здобувача вищої освіти**
**ДЗУНДЗА Дмитро**

1. Тема кваліфікаційної роботи «Дослідження цін на біржі засобами штучного інтелекту» затверджена наказом ректора ЧНУ ім. Петра Могили № _182__ від «_02_» __липня_____ 2025 р.
2. Строк представлення кваліфікаційної роботи «___» _____ 2025 р.
3. Розроблена та впроваджена інформаційна система аналізу та прогнозування біржових цін з веб-інтерфейсом, REST API та базою даних, що інтегрує методи машинного навчання (XGBoost, LSTM, ARIMA), sentiment аналіз новин та AI-аналітику для підвищення точності прогнозів на 8-14% порівняно з базовими моделями.
4. Перелік питань, що підлягають розробці:

– аналіз предметної області, огляд літератури та специфікація вимог до системи;

– обґрунтування вибору методів аналізу кореляцій, обробки новин та прогнозування;

– проектування архітектури інформаційної системи та вибір технологічного стеку;

– реалізація модуля аналізу кореляцій активів (кореляційні матриці, PCA, кластеризація, MST);

– реалізація модуля sentiment аналізу новин (VADER, обчислення індексу тональності);

– розробка та інтеграція моделей прогнозування цін (XGBoost, LSTM, ARIMA);

– розробка REST API та веб-інтерфейсу системи;

– тестування програмного забезпечення (модульне, інтеграційне, API);

– проведення експериментальних досліджень на реальних даних;

– аналіз результатів та порівняння ефективності моделей.

5. Перелік графічних матеріалів:

презентація.

6. Консультанти:

| **Консультант** | **Кафедра (організація)** | **Частина роботи** |
|---|---|---|
| | | |
| | | |
| | | |

Дата видачі завдання «____» _____ 20___ р.

**КАЛЕНДАРНИЙ ПЛАН**
**виконання кваліфікаційної роботи**

Тема: **Дослідження цін на біржі засобами штучного інтелекту**

| № | Найменування роботи | Початок | Закінчення | Примітки |
|---|---|---|---|---|
| 1. | Розробка та затвердження завдання на виконання КМР | 01.09.2025 | 01.09.2025 | Виконано |
| 2. | Огляд літератури за темою роботи | 09.09.2025 | 16.09.2025 | Виконано |
| 3. | Складання календарного плану КМР | 17.09.2025 | 24.09.2025 | Виконано |
| 4. | Аналіз предметної області | 25.09.2025 | 06.10.2025 | Виконано |
| 5. | Розробка проєктних рішень | 27.10.2025 | 13.10.2025 | Виконано |
| 6. | Моделювання та конструювання ПЗ | 14.10.2025 | 21.10.2025 | Виконано |
| 7. | Кодування, тестування та апробація розробленого ПЗ, аналіз результатів тестування, розробка керівництва користувача | 22.10.2025 | 01.11.2025 | Виконано |
| 8. | Відгук керівника КМР | 02.11.2025 | 03.11.2025 | Виконано |
| 9. | Оформлення КМР та презентації | 04.11.2025 | 21.11.2025 | Виконано |
| 10. | Попередній захист | 24.11.2025 | 24.11.2025 | Виконано |
| 11. | Рецензування | 01.12.2025 | 08.12.2025 | Виконано |
| 12. | Завершення оформлення КМР та презентації | 09.12.2025 | 19.12.2025 | Виконано |
| 13. | Захист кваліфікаційної роботи | 22.12.2025 | 23.12.2025 | Виконано |

**Здобувач** _____ **Дмитро ДЗУНДЗА**

«__» _____ 20__ р.

**Керівник роботи**

PhD

старший викладач _____ **Ігор КАНДИБА**

«__» _____ 20__ р.

# АНОТАЦІЯ

до кваліфікаційної магістерської роботи

**Дослідження цін на біржі засобами штучного інтелекту**

Здобувач 608 гр.: Дзундза Дмитро

Керівник: PhD, старший викладач Кандиба Ігор

Актуальність. Сучасні фінансові ринки характеризуються високою волатильністю та залежністю від зовнішніх факторів, зокрема новинних подій та міжринкових взаємозв'язків. Використання методів штучного інтелекту дозволяє підвищити точність прогнозування та підтримати процес прийняття інвестиційних рішень.

Об'єкт – процеси формування та прогнозування біржових цін.

Предмет – методи та засоби штучного інтелекту для аналізу кореляцій активів і впливу новин на цінові коливання.

Мета – дослідити можливості застосування штучного інтелекту для аналізу та прогнозування біржових цін з урахуванням кореляційних залежностей і новинних факторів.

Кваліфікаційна робота складається із вступу, чотирьох розділів, висновків та переліку джерел посилання.

У вступі обґрунтовано актуальність теми, визначено мету, завдання, об'єкт, предмет та методи дослідження.

У першому розділі проведено аналіз предметної області та наукових досліджень, присвячених штучному інтелекту в біржовій аналітиці.

Другий розділ містить методи дослідження: кореляційний аналіз активів, обробку новинного потоку за допомогою NLP та sentiment analysis, а також вибір алгоритмів машинного і глибинного навчання.

У третьому розділі розроблено інформаційну систему, що реалізує модулі аналізу кореляцій і новин та інтегрує прогностичні моделі.

Четвертий розділ присвячено експериментальним дослідженням і оцінці результатів на реальних біржових даних.

У висновках узагальнено результати, підкреслено наукову новизну та практичну значущість роботи.

Кваліфікаційна робота викладена на 95 сторінках машинописного тексту, складається із вступу, 4 розділів, загальних висновків, переліку джерел посилання з 52 найменувань та 1 додатку. Праця містить 11 таблиць та 35 рисунків.

*Ключові слова: біржові ціни, кореляція активів, новини, штучний інтелект, sentiment analysis, машинне навчання, прогнозування.*

**ABSTRACT**

to the qualifying master's thesis

**Stock Price Research Using Artificial Intelligence**

Student of 608 group: Dzundza Dmytro

Supervisor: phd, Senior Lecturer Kandyba Ihor

Relevance. Modern financial markets are highly volatile and influenced by external factors, including news events and inter-market correlations. The use of artificial intelligence methods improves forecasting accuracy and supports investment decision-making processes.

Object – processes of stock price formation and forecasting.

Subject – artificial intelligence methods and tools for analyzing asset correlations and the impact of news on price fluctuations.

Purpose – to explore the possibilities of applying artificial intelligence for stock price analysis and forecasting, taking into account correlation dependencies and news factors.

The qualification work consists of an introduction, four sections, conclusions and a list of references.

The introduction substantiates the relevance, defines the purpose, objectives, object, subject, and research methods.

The first section presents a review of the subject area and scientific research on the application of AI in stock market analytics.

The second section describes research methods: correlation analysis of assets, news stream processing using NLP and sentiment analysis, and selection of machine and deep learning algorithms.

The third section focuses on the development of an information system that implements modules for correlation and news analysis and integrates forecasting models.

The fourth section presents experimental studies and evaluation of results based on real stock market data.

he conclusions summarize the findings, highlight the scientific novelty, and emphasize the practical significance of the work.

The qualification work is presented on 95 pages of typewritten text, consists of an introduction, 4 sections, general conclusions, a list of references with 52 titles and 1 appendice. The work contains 11 tables and 35 figures.

*Keywords: stock prices, asset correlation, news, NLP, sentiment analysis, machine learning, forecasting.*

# ЗМІСТ

# ПЕРЕЛІК СКОРОЧЕНЬ

EMH — Efficient Market Hypothesis (гіпотеза ефективного ринку)

OHLCV — Open, High, Low, Close, Volume (формат біржових даних)

VAR — Vector Autoregression (векторна авторегресія)

VECM — Vector Error Correction Model (модель корекції помилок)

ARCH — Autoregressive Conditional Heteroskedasticity

GARCH — Generalized ARCH

CAPM — Capital Asset Pricing Model

APT — Arbitrage Pricing Theory

MST — Minimum Spanning Tree (мінімальне остовне дерево)

ML — Machine Learning (машинне навчання)

DL — Deep Learning (глибинне навчання)

AI — Artificial Intelligence (штучний інтелект)

NLP — Natural Language Processing (обробка природної мови)

RNN — Recurrent Neural Network

LSTM — Long Short-Term Memory

GRU — Gated Recurrent Unit

CNN — Convolutional Neural Network

SVM — Support Vector Machine

PCA — Principal Component Analysis

BERT — Bidirectional Encoder Representations from Transformers

GPT — Generative Pretrained Transformer

TF-IDF — Term Frequency — Inverse Document Frequency

MDS — Multidimensional Scaling

RSI — Relative Strength Index

ATR — Average True Range

MSE — Mean Squared Error

RMSE — Root Mean Squared Error

MAE — Mean Absolute Error

MAPE — Mean Absolute Percentage Error

AIC — Akaike Information Criterion

$R^2$ — R-squared (коефіцієнт детермінації)

API — Application Programming Interface

REST — Representational State Transfer

UI — User Interface

ORM — Object-Relational Mapping

DB — Database

CI/CD — Continuous Integration / Continuous Deployment

JSON — JavaScript Object Notation

HTML — HyperText Markup Language

CSS — Cascading Style Sheets

UML — Unified Modeling Language

ASGI — Asynchronous Server Gateway Interface

GPU — Graphics Processing Unit

RAM — Random Access Memory

SSD — Solid State Drive

HTTPS — HyperText Transfer Protocol Secure

SQL — Structured Query Language

CSV — Comma-Separated Values

RSS — Really Simple Syndication

URL — Uniform Resource Locator

ER-діаграма — Entity-Relationship діаграма

# ВСТУП

Сучасні фінансові ринки відзначаються високою динамічністю та складністю процесів ціноутворення. Біржові ціни змінюються під впливом як внутрішніх ринкових механізмів, так і зовнішніх факторів: глобальних економічних тенденцій, політичних рішень, природних катаклізмів, а також інформаційних повідомлень у медіа. У цих умовах завдання прогнозування біржових цін набуває особливої актуальності для інвесторів, трейдерів, фінансових аналітиків та розробників інформаційних систем.

Традиційні статистичні методи аналізу фінансових часових рядів не завжди здатні врахувати багатофакторність та нелінійність ринкових процесів. Саме тому значного поширення набули методи штучного інтелекту, зокрема машинне навчання та глибинне навчання. Вони дозволяють автоматизувати обробку великих масивів даних, виявляти приховані залежності та підвищувати точність прогнозування.

Особливе значення для сучасних фінансових досліджень мають два напрями:

Аналіз кореляцій між активами – виявлення зв’язків між динамікою цін на різних ринках і у різних секторах економіки. Це дозволяє підвищити ефективність диверсифікації портфеля та точність прогнозних моделей.

Аналіз впливу новин на ціни – оцінка того, як публікації у ЗМІ та соціальних мережах (у тому числі їхня тональність) впливають на біржові коливання. Методи обробки природної мови (NLP) і sentiment analysis дають змогу автоматизувати цей процес.

**Об’єкт дослідження** – процеси формування та прогнозування біржових цін.

**Предмет дослідження** – методи та інструменти штучного інтелекту для аналізу кореляцій активів і впливу новин на динаміку цін.

**Мета роботи** – дослідити можливості застосування штучного інтелекту для аналізу та прогнозування біржових цін з урахуванням міжактивних кореляцій і впливу новинних факторів.

Для досягнення поставленої мети необхідно виконати такі завдання:

- провести аналіз предметної області та огляд наукових публікацій;
- дослідити методи аналізу часових рядів і кореляцій між фінансовими активами;
- розглянути методи обробки новинного потоку (NLP, sentiment analysis);
- обґрунтувати вибір алгоритмів машинного та глибинного навчання для прогнозування цін;
- розробити інформаційну систему, що реалізує модулі аналізу кореляцій і новин;
- провести експериментальні дослідження та оцінити ефективність розробленої системи.

**Методи дослідження.** У роботі використано: методи математичної статистики та кореляційного аналізу; алгоритми машинного й глибинного навчання; методи обробки природної мови (NLP, sentiment analysis); методи програмної інженерії для проєктування та розробки інформаційних систем.

**Наукова новизна** полягає у комплексному застосуванні методів аналізу кореляцій та обробки новинних даних для прогнозування біржових цін. Розроблений підхід дозволяє інтегрувати ринкові та текстові ознаки у єдину модель прогнозування.

**Практична значущість** полягає у створенні інформаційної системи, що може використовуватися для підтримки інвестиційних рішень, автоматизації фінансової аналітики та дослідження впливу інформаційних потоків на ринкові ціни.

Апробація результатів КМР відбулась під час XXVIII Всеукраїнської науково-практичної конференції «Могилянські читання – 2025», Миколаїв, 10–14 листопада, 2025 р. (Додаток А).

# 1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ОГЛЯД ЛІТЕРАТУРИ

## 1.1 Біржові ринки та механізми ціноутворення

Біржові ринки є невід'ємною складовою сучасної економічної системи. Вони виконують роль посередника між продавцями та покупцями активів, забезпечуючи ліквідність, прозорість та ефективність угод. Залежно від об'єктів торгівлі розрізняють фондові, валютні, товарні та похідні ринки. Кожен із них має власну специфіку, але всі об'єднані спільним механізмом ціноутворення, що базується на співвідношенні попиту і пропозиції.

**Мікроструктура ринку.**

Сучасні біржі працюють за принципом електронних торгових систем, де ключовим елементом є ордербук – електронна книга заявок, у якій відображаються всі поточні пропозиції на купівлю та продаж активів. Найважливішими параметрами, що визначають динаміку цін у короткостроковому періоді, є:

- ліквідність (глибина ринку, тобто кількість заявок на різних рівнях цін);
- спред (різниця між найкращою ціною купівлі та найкращою ціною продажу);
- обсяги угод (volume), що відображають інтенсивність ринкової активності.

Висока ліквідність сприяє стабільності ціни, тоді як її нестача призводить до підвищеної волатильності.

**Вплив інформації.**

Особливістю фінансових ринків є їхня чутливість до інформації. Будь-які новини – макроекономічні звіти, політичні рішення, природні катастрофи чи корпоративні оголошення – миттєво враховуються у цінах через реакцію учасників торгів. Це підтверджує концепція ефективного ринку, згідно з якою ціни в будь-який момент часу відображають усю доступну інформацію.

Однак на практиці ефективність ринку часто є обмеженою. Наявні поведінкові фактори (емоції, ірраціональні очікування) спричиняють відхилення

цін від фундаментальних значень. Саме в цих відхиленнях і криються можливості для трейдингу, які сучасні аналітичні методи прагнуть виявити.

**Кореляції між активами.**

Ринки не існують ізольовано: зміни на одному сегменті фінансової системи відбиваються на інших. Наприклад:

- ріст цін на нафту часто корелює зі зростанням валют країн-експортерів;
- індекси фондового ринку США впливають на динаміку європейських та азійських ринків;
- під час криз спостерігається підвищення кореляції між більшістю активів (ефект «flight to safety»).

Вивчення цих зв’язків важливе для побудови портфелів, управління ризиками та створення моделей прогнозування.

**Моделі ціноутворення.**

Класичні економічні теорії описують механізм формування ціни через баланс попиту і пропозиції. У сучасних дослідженнях активно використовуються моделі:

- рівноважні, що описують довгострокові тенденції (наприклад, CAPM, APT);
- стохастичні, що моделюють випадковий характер змін цін (модель Блека–Шоулза, геометричне броунівське рухання);
- поведінкові, що враховують психологічні особливості інвесторів.

Поєднання класичних моделей із сучасними інструментами штучного інтелекту дозволяє отримати більш точні результати, оскільки AI здатний навчатися на великих обсягах історичних і поточних даних, виявляючи приховані закономірності.

## 1.2 Теоретичні підходи до аналізу фінансових часових рядів

Фінансові часові ряди є фундаментальним об’єктом дослідження в економетриці, кількісних фінансах і сучасній фінансовій інформатиці. Під часовим рядом зазвичай розуміють послідовність значень деякої економічної змінної, що впорядковані у часі з певною періодичністю. Для ринків капіталу такими рядами є

ціни акцій, валютні курси, вартість товарів, а також похідні величини, наприклад, прибутковість чи волатильність.

Особливістю фінансових часових рядів є їхня нестабільність і складність структури. На відміну від природних чи технічних процесів, фінансові дані відображають колективну поведінку великої кількості учасників ринку, чиї рішення залежать від економічної ситуації, інформаційних потоків та навіть психологічних чинників. Тому для їхнього аналізу застосовують широкий спектр теоретичних і практичних підходів, які поєднують у собі інструменти статистики, економетрики, машинного навчання та штучного інтелекту.

**Основні характеристики фінансових часових рядів**

Фінансові часові ряди мають кілька властивостей, які визначають вибір методів аналізу.

По-перше, вони зазвичай є нестаціонарними, тобто математичне сподівання та дисперсія змінюються з часом. Це проявляється у тому, що середній рівень цін і ступінь їхніх коливань суттєво відрізняються в різні періоди. Наприклад, фондові індекси у фазі економічного підйому демонструють плавний ріст, тоді як у кризові роки характеризуються різкими спаданнями й підвищеною волатильністю.

По-друге, для таких рядів типовим є ефект товстих хвостів у розподілі прибутковостей. Це означає, що великі відхилення від середнього трапляються значно частіше, ніж у випадку нормального розподілу. Традиційні методи, які ґрунтуються на нормальності, недооцінюють ризики, що робить необхідним застосування спеціалізованих моделей.

По-третє, фінансові ряди мають властивість кластеризації волатильності: періоди різких коливань зазвичай групуються, створюючи так звані «режими високої волатильності». Це добре видно на прикладі валютних курсів у часи політичних криз чи на фондовому ринку після публікації важливих корпоративних новин.

Ще однією рисою є автокореляція, коли поточні значення частково залежать від попередніх. У коротких горизонтах це дозволяє виявити певні закономірності, однак у довших періодах така залежність швидко зникає.

**Класичні статистичні підходи**

Першим класом методів є лінійні статистичні моделі, які описують залежність поточного значення ряду від його минулих значень та випадкових збурень. Найпоширенішою є модель ARIMA (Autoregressive Integrated Moving Average). Вона записується у вигляді:

$$y_t = c + \phi_1 y_{\{t-1\}} + \phi_2 y_{\{t-2\}} + \ldots + \phi_p y_{\{t-p\}} + \theta_1 \varepsilon_{\{t-1\}} + \theta_2 \varepsilon_{\{t-2\}} + \ldots + \theta_q \varepsilon_{\{t-q\}} + \varepsilon_{t1}$$

де $y_t$ – значення часового ряду,

$\varepsilon_{\{t\}}$ – випадкова похибка,

$\phi_i$ – параметри авторегресії,

$\theta_j$ – параметри ковзного середнього.

Компонент інтегрування (I) дозволяє працювати з нестаціонарними рядами шляхом диференціювання. Практика показує, що ARIMA ефективна для прогнозування на короткі горизонти, проте втрачає точність у періоди структурних зламів і для рядів із високою волатильністю.

Інший важливий клас моделей описує не самі рівні цін, а їхню волатильність. Тут використовуються моделі ARCH та GARCH. Стандартна GARCH(1,1) модель виглядає так:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{\{t-1\}}^2 + \beta_1 \sigma_{\{t-1\}}^2$$

де $\sigma_t^2$ – умовна дисперсія у момент часу $t$,

$\varepsilon_{\{t-1\}}^2$ – квадрати попередніх похибок,

а $\sigma_{\{t-1\}}^2$ – попереднє значення дисперсії.

Такий підхід добре моделює кластери волатильності й широко використовується для розрахунку ризиків у банківському секторі.

Коли потрібно врахувати кілька взаємопов’язаних рядів, застосовують VAR-моделі (Vector Autoregression). Вони описують систему рівнянь, у якій кожна змінна залежить від власних минулих значень і від минулих значень інших змінних. Наприклад, валютні курси долара, євро та фунта можуть бути змодельовані спільно, щоб врахувати їхню взаємодію. Якщо між рядами існують довгострокові

коінтеграційні зв’язки, застосовується VECM, яка враховує відхилення від рівноваги.

**Методи багатовимірного аналізу**

З огляду на те, що сучасні фінансові ринки генерують дані для сотень і тисяч інструментів, важливе місце займають методи зменшення розмірності. Аналіз головних компонент (PCA) дозволяє виділити кілька факторів, які пояснюють більшу частину дисперсії у даних. На практиці це означає можливість звести сотні цінових рядів до кількох «головних індексів», які відображають основні тенденції.

Кластеризація використовується для групування активів за подібністю їхньої динаміки. Наприклад, акції компаній з однієї галузі часто демонструють схожі тренди, і це можна формально підтвердити методами k-means чи ієрархічної кластеризації.

Мережевий аналіз додає ще один рівень: активи розглядаються як вузли графа, а сили їхніх кореляцій – як ребра. Такий підхід дозволяє побудувати «карту ринку» й виявити ключові активи, що впливають на інші.

**Сучасні методи штучного інтелекту**

За останні два десятиліття суттєво зросла роль машинного та глибинного навчання. На відміну від статистичних моделей, які спираються на чітко задану форму залежностей, методи штучного інтелекту здатні автоматично виявляти складні нелінійні закономірності.

У машинному навчанні поширені алгоритми на кшталт Random Forest чи XGBoost. Вони будують ансамблі дерев рішень і добре працюють із великою кількістю ознак, включаючи технічні індикатори, макроекономічні дані й новинні фактори.

Найбільшої популярності у фінансовому прогнозуванні набули рекурентні нейронні мережі (RNN), зокрема LSTM (Long Short-Term Memory) та GRU (Gated Recurrent Units). Вони здатні зберігати інформацію про довгі часові залежності, що є критично важливим для фінансових рядів. Математично LSTM складається з комірок пам’яті, у яких за допомогою спеціальних «воріт» контролюється потік інформації. Основні рівняння LSTM включають:

Forget gate (забувальні ворота):

$$f_t = \sigma\left(W_f \cdot \left[h_{\{t-1\}}, x_t\right] + b_f\right)$$

Input gate (вхідні ворота):

$$i_t = \sigma\left(W_i \cdot \left[h_{\{t-1\}}, x_t\right] + b_i\right)$$

Candidate cell state (кандидат у стан пам’яті):

$$\sim\{C\}_t = \backslash tanh\left(W_C \cdot \left[h_{\{t-1\}}, x_t\right] + b_C\right)$$

Update cell state (оновлений стан пам’яті):

$$C_t = f_t * C_{\{t-1\}} + i_t * \sim\{C\}_t$$

Output gate (вихідні ворота):

$$o_t = \sigma\left(W_o \cdot \left[h_{\{t-1\}}, x_t\right] + b_o\right)$$

Hidden state (прихований стан):

$$h_t = o_t \cdot \backslash tanh(C_t)$$

де $f_t$, $i_t$, $o_t$ – відповідно вхідні, вихідні та забувальні «ворота»,

$C_t$ – стан пам’яті,

$h_t$ – прихований стан.

Окрім RNN, активно застосовуються конволюційні мережі (CNN), які здатні виявляти локальні патерни у часових рядах, та трансформери, що завдяки механізму уваги можуть одночасно враховувати короткі й довгі залежності.

Сучасна практика показує, що найбільш ефективними є гібридні моделі, які поєднують статистичні й нейронні методи. Наприклад, ARIMA може описати трендову складову, а LSTM – спрогнозувати залишкові коливання. Такі моделі краще враховують як лінійні, так і нелінійні залежності та показують вищу точність у порівнянні з класичними методами.

**Порівняльний аналіз підходів**

Класичні статистичні методи мають перевагу у простоті та прозорості інтерпретації. Вони добре працюють на коротких горизонтах і широко використовуються у банківській сфері для оцінки ризиків. Проте вони малоефективні в умовах високої турбулентності ринку.

Методи машинного навчання здатні враховувати велику кількість факторів, але потребують ретельної підготовки даних. Глибинні нейронні мережі демонструють високу точність, однак залишаються «чорними скриньками», результати яких важко інтерпретувати.

Найбільш перспективним напрямом вважається інтеграція кількісних фінансів і штучного інтелекту. Це підтверджують і наукові дослідження: наприклад, у роботах останніх років показано, що гібридні моделі з використанням LSTM і трансформерів суттєво перевершують ARIMA за точністю прогнозування фондових індексів і валютних курсів.

## 1.3 Використання методів штучного інтелекту у фінансових дослідженнях

Штучний інтелект (ШІ) та алгоритми машинного навчання останніми десятиліттями набули особливої популярності у сфері фінансів. Ринки капіталу, на відміну від більшості економічних систем, генерують величезні обсяги даних у реальному часі, що робить їх ідеальним середовищем для застосування методів обробки великих даних та глибинного навчання. Мета таких досліджень полягає у пошуку закономірностей у складних часових рядах, оптимізації процесу прийняття рішень та підвищенні точності прогнозування цінових рухів.

**Особливості застосування штучного інтелекту у фінансовій сфері**

На відміну від класичних статистичних методів, ШІ підходить для роботи з даними, які:

- мають нелінійну структуру;
- містять велику кількість ознак і прихованих взаємозв’язків;
- характеризуються нестабільністю та різкими стрибками.

Це дає можливість будувати прогностичні моделі, які враховують не лише попередні значення ряду, але й супутні фактори: макроекономічні показники, ринкові новини, соціальні медіа-сигнали, корпоративні звіти.

ШІ у фінансах застосовується не тільки для прогнозування цін активів, але й для розпізнавання шахрайства, виявлення аномалій у транзакціях, оптимізації портфелів, алгоритмічної торгівлі та управління ризиками.

**Методи машинного навчання у фінансових дослідженнях**

Одним із ключових напрямів є застосування класичних алгоритмів машинного навчання. Лінійна та логістична регресія традиційно використовуються для базового аналізу взаємозв'язків між фінансовими змінними. Проте їхня здатність працювати лише з лінійними залежностями обмежує сферу використання.

Більш просунуті алгоритми, такі як Random Forest і Gradient Boosting, дозволяють виявляти складні нелінійні закономірності. Наприклад, Random Forest будує ансамбль дерев рішень, що знижує ризик перенавчання і підвищує стабільність прогнозів. Ці методи показали ефективність у задачах класифікації напрямку руху цін (зростання чи падіння) та у виявленні ключових факторів, що впливають на ринок.

Метод опорних векторів (SVM) застосовується для задач класифікації та регресії. У фінансах він використовується для передбачення трендів на ринку акцій та валют. Завдяки використанню ядерних функцій SVM здатний моделювати складні нелінійні залежності між ознаками.

**Глибинне навчання у фінансових дослідженнях**

Глибинні нейронні мережі відкрили нові можливості у фінансовому прогнозуванні. Найбільшого поширення набули рекурентні нейронні мережі, зокрема LSTM та GRU. Вони дозволяють моделювати часові залежності у довгих фінансових рядах, що особливо важливо при прогнозуванні валютних курсів чи цін акцій на кілька кроків уперед.

У багатьох дослідженнях LSTM демонструє суттєве підвищення точності у порівнянні з ARIMA чи GARCH, особливо тоді, коли ринок перебуває у стані високої турбулентності. Наприклад, застосування LSTM для прогнозування індексу S&P 500 дозволило досягти кращих результатів у періоди фінансових криз, ніж класичні економетричні моделі.

Окрему нішу займають конволюційні нейронні мережі (CNN). Попри те, що вони розроблені для обробки зображень, CNN виявились ефективними для фінансових часових рядів, оскільки здатні розпізнавати локальні патерни у даних. В останніх дослідженнях CNN використовуються для створення «зображень» часових рядів і виявлення прихованих структур, що залишаються непомітними для традиційних методів.

Трансформери, які стали проривом у сфері обробки природної мови, також знайшли своє застосування у фінансах. Завдяки механізму уваги вони можуть одночасно аналізувати як короткі, так і довгі часові залежності. Моделі на основі трансформерів, такі як Informer чи Temporal Fusion Transformer, показують високу ефективність у прогнозуванні довгострокових тенденцій.

**Обробка новин та соціальних медіа**

Суттєвою перевагою методів ШІ є здатність працювати з неструктурованими даними, такими як тексти новин чи повідомлення у соціальних мережах. NLP дозволяє автоматично визначати тональність текстів (sentiment analysis) і враховувати її у прогнозних моделях.

Дослідження показали, що аналіз настроїв інвесторів за допомогою NLP може пояснювати до 20–30% короткострокових коливань цін. Наприклад, негативні новини часто викликають миттєві падіння цін, тоді як позитивні – поступове зростання. У поєднанні з моделями часових рядів такі підходи суттєво підвищують точність прогнозів.

**Гібридні системи**

Найбільш перспективним напрямом є поєднання класичних економетричних методів і сучасних нейронних мереж у гібридні моделі. Це дозволяє врахувати як інтерпретовані статистичні залежності, так і складні нелінійні закономірності. Наприклад, ARIMA може описувати трендову складову, а LSTM – передбачати випадкові коливання. Інший приклад – поєднання GARCH з нейронними мережами для моделювання волатильності.

У наукових роботах доведено, що гібридні підходи перевершують окремо статистичні чи нейронні моделі, особливо в умовах кризових змін середовища, коли ринки стають непередбачуваними.

**Практичне застосування у фінансовій індустрії**

Сьогодні штучний інтелект активно використовується інвестиційними банками, хедж-фондами та біржами. Основні напрями застосування включають:

- алгоритмічну торгівлю, де системи ШІ здійснюють тисячі операцій за секунду;
- управління ризиками, включаючи прогнозування Value-at-Risk;
- автоматичне виявлення шахрайських транзакцій у платіжних системах;
- побудову рекомендаційних систем для інвесторів.

Використання ШІ дозволяє фінансовим установам знижувати ризики, підвищувати ефективність та створювати нові конкурентні переваги.

## 1.4 Стан наукових досліджень з аналізу кореляцій активів і впливу новин

Фінансові ринки є складними системами, у яких динаміка окремих активів часто виявляє взаємозалежність. Ці кореляційні зв’язки відображають структурні особливості ринку, спільну реакцію на макроекономічні чинники та поведінкові патерни інвесторів. Окремим важливим чинником формування цін є інформаційний потік, зокрема новини, аналітичні звіти, повідомлення в соціальних медіа. Аналізуючи сучасні наукові дослідження, можна виділити два напрями, що активно розвиваються: вивчення кореляцій між активами та аналіз впливу новинного фону на ринкові ціни.

**Дослідження кореляційних залежностей між активами**

У науковій літературі давно підкреслюється, що активи на фінансових ринках рідко поводяться незалежно. Залежності між ними формуються як під впливом фундаментальних факторів, так і через колективну поведінку інвесторів. Наприклад, нафта й валютні курси країн-експортерів, золото та фондові індекси у

кризові періоди, технологічні акції у рамках одного сектору часто демонструють високий рівень синхронності.

Перші підходи до дослідження таких залежностей ґрунтувались на класичних кореляційних коефіцієнтах Пірсона. Проте пізніші роботи показали, що статичний коефіцієнт кореляції не відображає змінності ринкових відносин у часі. У відповідь на це з'явилися методи ковзаючих вікон та динамічної умовної кореляції (DCC-GARCH), які дозволяють відстежувати, як змінюються зв'язки між активами в різні періоди.

Сучасні дослідження активно застосовують мережевий аналіз, де активи представляються як вузли графа, а сила їхніх кореляцій – як ребра. Це дозволяє будувати «карти ринку», виявляти кластери активів і визначати центральні елементи фінансової системи. Наприклад, праці Diebold & Yilmaz показали, як методи аналізу мереж можна використовувати для вимірювання «системного ризику» та передачі шоків між ринками.

Додатковим напрямом є використання копул-моделей, які дозволяють моделювати залежності між активами поза межами лінійної кореляції. Завдяки цьому можна описувати асиметричні та нелінійні зв'язки, що характерні для кризових періодів, коли звичайні кореляційні коефіцієнти стають ненадійними.

**Вплив новин та інформаційних потоків на фінансові ринки**

Паралельно з дослідженням кореляцій між активами розвивається напрям аналізу інформаційного середовища. Відомо, що ринки надзвичайно чутливі до новин – макроекономічних показників, монетарних заяв, корпоративних звітів чи навіть дописів у соціальних мережах.

У класичних дослідженнях вивчення цього ефекту обмежувалось аналізом календарних подій, таких як оголошення процентних ставок чи публікація індексів інфляції. Проте з розвитком цифрових технологій з'явилася можливість обробляти великі масиви текстових даних у реальному часі.

NLP дозволили автоматизувати процес аналізу новин. Зокрема, аналіз тональності (sentiment analysis) став стандартним інструментом для вимірювання емоційного забарвлення повідомлень. Наукові дослідження показали, що

позитивний інформаційний фон сприяє поступовому зростанню цін, тоді як негативні новини призводять до різких спадів.

Цікавим напрямом є використання соціальних медіа як джерела прогнозної інформації. Наприклад, відомі дослідження показали, що активність користувачів Twitter може передбачати короткострокову динаміку фондових індексів. У випадку криптовалют цей ефект особливо виражений: ціни біткойна суттєво реагують на повідомлення від відомих осіб та великих інвесторів.

**Поєднання аналізу кореляцій та новин**

Сучасна тенденція у наукових роботах полягає у комплексному врахуванні як міжринкових залежностей, так і інформаційних факторів. Це пояснюється тим, що новини часто впливають не лише на окремий актив, а й на цілий сегмент ринку. Наприклад, негативні макроекономічні новини здатні одночасно збільшувати кореляцію між усіма фондовими індексами через ефект масового розпродажу.

Гібридні моделі, які інтегрують кореляційний аналіз і NLP, показують вищу точність прогнозування. Наприклад, моделі на основі LSTM чи Transformers, які отримують на вхід як історичні ціни, так і числові показники настроїв новин, забезпечують кращі результати, ніж використання лише цінових рядів.

У наукових роботах останніх років дедалі частіше розглядається концепція «економіки даних у реальному часі» (real-time data economy), коли система автоматично збирає ринкові котирування й інформаційні потоки, аналізує кореляційні структури та новини, а результати інтегруються в торговельні стратегії.

**Проблеми та виклики досліджень**

Попри значний прогрес, існує низка викликів:

- складність у відокремленні «шумових» новин від справді важливих сигналів;
- багатомовність і неоднорідність інформаційних джерел;
- швидка зміна кореляційної структури під час кризових подій;
- ризик перенавчання моделей при інтеграції новинних даних у прогностичні системи.

## Висновки до розділу 1

У результаті проведеного аналізу предметної області та огляду наукових джерел можна зробити низку узагальнень, які визначають методологічну основу подальших досліджень.

Біржові ринки є складними багатофакторними системами, у яких формування ціни відбувається під впливом не лише попиту й пропозиції, а й макроекономічних показників, корпоративної політики, геополітичних подій та інформаційного середовища. Це обумовлює високу волатильність і динамічність фінансових даних, що ускладнює процес прогнозування та водночас робить його особливо актуальним.

Дослідження фінансових часових рядів показало, що вони мають специфічні властивості: нестаціонарність, ефект «товстих хвостів», автокореляції та кластеризацію волатильності. Класичні економетричні моделі (ARIMA, GARCH, VAR, VECM) зберігають своє значення, особливо для інтерпретації та короткострокового прогнозування, однак їхня ефективність є обмеженою в умовах кризових змін та нелінійних процесів.

Методи штучного інтелекту, зокрема алгоритми машинного та глибинного навчання, продемонстрували здатність виявляти складні залежності у фінансових даних і забезпечувати вищу точність прогнозів. Використання рекурентних нейронних мереж (LSTM, GRU), конволюційних мереж і трансформерів створило новий рівень можливостей для аналізу ринків. Водночас ці моделі мають суттєві вимоги до обчислювальних ресурсів і часто працюють як «чорні скриньки», що знижує інтерпретованість результатів.

Огляд наукових досліджень підтвердив важливість врахування як кореляцій між активами, так і впливу новинного фону. Сучасні підходи поєднують кореляційний аналіз із методами обробки природної мови (NLP), що дозволяє враховувати настрої інвесторів і більш точно описувати поведінку ринків у періоди підвищеної невизначеності.

# 2 МЕТОДИ ТА ІНСТРУМЕНТИ ДОСЛІДЖЕННЯ ЦІН НА БІРЖІ

## 2.1 Специфікація вимог

### Призначення та межі проєкту

### Призначення системи

Програмна система призначена для комплексного аналізу фінансових даних біржових ринків із використанням методів штучного інтелекту. Основними функціями системи є:

- збір та збереження біржових котирувань у режимі близькому до реального часу;
- обчислення кореляцій між активами для виявлення структурних зв’язків;
- обробка новинних потоків і визначення їхньої тональності;
- інтеграція цінових рядів та новинних індикаторів у єдину модель прогнозування;
- візуалізація результатів у вигляді графіків, матриць та звітів.

Система орієнтована на дослідників, аналітиків і приватних інвесторів, яким необхідні інструменти для глибокого аналізу динаміки ринку.

### Погодження, що ухвалені в програмній документації

Під час розробки приймаються такі загальні положення:

- для прогнозування використовуються методи класичного аналізу часових рядів (ARIMA, GARCH) у поєднанні з глибинними нейронними мережами (LSTM, Transformers);
- для обробки новин застосовуються методи NLP та аналізу тональності;
- джерела даних: біржові API (Yahoo Finance, Binance API), інформаційні агрегатори новин (Google News, Twitter API);
- результати експериментів подаються у вигляді кількісних метрик (MSE, RMSE, MAE, accuracy).

### Межі проєкту ПЗ

Система не є торговельним майданчиком і не здійснює автоматичне виконання угод. Вона функціонує виключно як аналітичний та дослідницький

інструмент. Крім того, система не передбачає обробку конфіденційних даних користувача – всі джерела інформації є відкритими.

**Загальний опис**

**Сфера застосування**

Програмний продукт використовується у сфері фінансової аналітики, зокрема для прогнозування цінових тенденцій та аналізу інформаційного фону. Його можуть застосовувати студенти, науковці, інвестиційні компанії та приватні трейдери.

**Характеристики користувачів**

Основними користувачами є аналітики з базовими знаннями у фінансах і програмуванні. Система передбачає інтуїтивно зрозумілий інтерфейс, тому не вимагає глибоких технічних знань.

**Загальна структура і склад системи**

Система складається з таких модулів:

- модуль збору та попередньої обробки даних;
- модуль аналізу кореляцій;
- модуль обробки новинного потоку;
- модуль прогнозування з використанням ML/DL;
- модуль візуалізації результатів.

**Загальні обмеження**

Система працює з відкритими джерелами даних і потребує стабільного інтернет-з’єднання. Продуктивність залежить від обчислювальних ресурсів – для тренування нейронних мереж бажано мати графічний процесор (GPU).

**Функції системи**

Кожна функція описується за схемою: опис, вхідні/вихідні дані, функціональні вимоги.

**Функція 1. Аналіз кореляцій між активами**

Опис: побудова кореляційних матриць, виділення кластерів за допомогою PCA та кластеризації.

Вхідні дані: часові ряди котирувань.

Вихідні дані: матриця кореляцій, графічні візуалізації.

Вимоги: можливість вибору періоду, інтервалу та набору активів.

**Функція 2. Аналіз новинного потоку**

Опис: завантаження новин, аналіз тональності за допомогою NLP.

Вхідні дані: тексти новин і соціальних медіа.

Вихідні дані: індекс настроїв, класифікація за тональністю (позитив/негатив/нейтраль).

Вимоги: підтримка багатомовних джерел, автоматичне оновлення.

**Функція 3. Прогнозування цін**

Опис: використання гібридних моделей (ARIMA+LSTM, Transformer) для прогнозу цін на основі історичних даних і новин.

Вхідні дані: часові ряди та індикатори настроїв.

Вихідні дані: прогнозні значення, метрики точності.

Вимоги: можливість вибору моделі, довжини горизонту прогнозування.

**Вимоги до інформаційного забезпечення**

**Джерела і зміст вхідної інформації**

- біржові котирування з відкритих API;
- текстові новини з агрегаторів;
- соціальні медіа (Twitter).

**Нормативно-довідкова інформація**

Використовуються довідники валют, біржових індексів, кодів акцій.

**Вимоги до організації даних**

Дані зберігаються у реляційній базі з можливістю резервного копіювання та експорту.

**Вимоги до технічного забезпечення**

- мінімум: процесор Intel i5-12k, 64 GB RAM, 512 GB SSD;
- бажано: наявність GPU (NVIDIA CUDA) для тренування моделей з 24 GB.

**Вимоги до програмного забезпечення**

**Архітектура**

Клієнт-серверна система з веб-інтерфейсом.

**Системне ПЗ**

ОС Windows/Linux/MacOS.

**Мережне ПЗ**

Підтримка HTTP/HTTPS-з’єднань.

**ПЗ ведення бази**

PostgreSQL / MySQL.

**Мова і технологія розробки**

Python (бібліотеки Pandas, Scikit-learn, TensorFlow/PyTorch), Flask/Django для бекенду, React для фронтенду.

**Вимоги до зовнішніх інтерфейсів**

**Інтерфейс користувача**

Веб-додаток з інтерактивними графіками, дашбордом і формами для завантаження даних.

**Апаратний інтерфейс**

Стандартна взаємодія через клавіатуру та мишу.

**Програмний інтерфейс**

REST API для інтеграції з іншими системами.

**Комунікаційний протокол**

HTTPS для захищеної передачі даних.

**Властивості програмного забезпечення**

**Доступність**

Робота у браузері, цілодобовий доступ.

**Супроводжуваність**

Модульна структура коду, документація.

**Переносимість**

Підтримка кількох ОС.

**Продуктивність**

Час відповіді при завантаженні даних $\leq 3$ с.

**Надійність**

Резервне копіювання, обробка винятків.

**Безпека**

Авторизація користувачів, захищені протоколи.

**Інші вимоги**

Документація для користувача (User Guide), UML-діаграми, приклади коду, експериментальні результати.

## 2.2 Джерела даних: біржові котирування та інформаційні потоки

Для проведення якісного дослідження цінових процесів на біржі необхідна надійна база даних. Від того, які саме дані будуть обрані, залежить точність аналізу, можливість застосування математичних моделей та адекватність результатів. У даному підпункті розглядаються основні типи даних, їхні джерела та особливості використання в рамках проєкту.

**Біржові котирування**

Основним джерелом інформації про фінансові ринки є біржові котирування – дані про ціни та обсяги торгів фінансовими інструментами (акції, облігації, валюти, товари, криптовалюти).

Найбільш поширені параметри котирувань:

- open – ціна відкриття;
- high – максимальна ціна за період;
- low – мінімальна ціна за період;
- close – ціна закриття;
- volume – обсяг торгів.

Ці показники утворюють стандартний формат OHLCV, який застосовується у більшості фінансових досліджень.

Дані котирувань можуть мати різний часовий інтервал (granularity): від тикових (щосекундних) до денних, тижневих чи місячних. Вибір інтервалу залежить від завдань дослідження: високочастотна торгівля потребує мілісекундних даних, тоді як для стратегій середньострокового прогнозування достатньо денних котирувань.

**Джерела котирувань**

Офіційні біржові API (NASDAQ, NYSE, CME, Binance). Вони надають дані у реальному часі, однак часто потребують підписки чи мають обмеження щодо кількості запитів.

Фінансові агрегатори (Yahoo Finance, Google Finance, Alpha Vantage, Quandl). Це зручні сервіси для отримання історичних даних у форматах CSV чи JSON.

Комерційні постачальники даних (Bloomberg, Thomson Reuters Refinitiv, Morningstar). Вони пропонують високоточні дані з розширеними параметрами, але доступ до них є платним.

Криптовалютні біржі (Binance, Coinbase, Kraken). Вони відкривають API для збору даних про ціни цифрових активів, що є особливо актуальним через високу волатильність крипторинку.

У даній роботі планується використовувати комбінацію Yahoo Finance API для історичних даних та Binance API для актуальних котирувань криптовалют.

**Інформаційні потоки**

Окрім цінових рядів, важливим джерелом інформації є новинні потоки. Вони впливають на очікування інвесторів і формують ринкову динаміку.

Джерела інформаційних потоків:

- новинні агентства (bloomberg, reuters, financial times, cnbc).
- агрегатори новин (google news api, eventregistry).
- соціальні мережі (twitter, reddit, stocktwits).
- офіційні публікації (центральні банки, уряди, компанії).

Для обробки текстів застосовуються методи NLP. Основними задачами є:

- вилучення релевантних повідомлень;
- аналіз тональності (sentiment analysis);
- побудова індикаторів «інформаційного настрою».

**Особливості збору та підготовки даних**

Дані котирувань та новин мають різну структуру. Тому необхідна попередня обробка:

- очищення від шумів (наприклад, видалення дублікатів новин);
- синхронізація часових міток (для зіставлення цін і новин);

- нормалізація даних (наприклад, стандартизація цінових рядів);
- перетворення тексту у числові вектори (Word2Vec, BERT, TF-IDF).

Якість підготовки даних безпосередньо впливає на результативність моделей прогнозування.

## 2.3 Методи аналізу кореляцій між активами (кореляційні матриці, PCA, кластеризація)

Фінансові активи не існують у вакуумі. Їхні ціни формуються під впливом як власних характеристик, так і взаємозв'язків з іншими активами. Наприклад, ціна нафти тісно пов'язана з курсами валют країн-експортерів, а динаміка індексу S&P 500 корелює з поведінкою європейських фондових індексів. Вивчення кореляційних структур дозволяє:

- оцінювати ступінь взаємозалежності активів;
- виявляти кластери фінансових інструментів;
- визначати канали передачі ризиків;
- будувати оптимальні інвестиційні портфелі.

У даному підпункті розглядаються три основні методи аналізу: кореляційні матриці, метод головних компонент (PCA) та кластеризація.

**Кореляційні матриці**

Найпростішим і найбільш поширеним інструментом є обчислення матриці кореляцій між доходностями активів. Для цього спочатку розраховуються логарифмічні доходності.

Далі обчислюється коефіцієнт кореляції Пірсона:

**Практичне значення:**

- високі значення (ρ≈1\rho \approx 1ρ≈1) означають синхронний рух;
- низькі або від'ємні значення (ρ≤0\rho \leq 0ρ≤0) свідчать про відсутність або протилежний напрям руху.

У сучасних дослідженнях кореляційні матриці будуються не лише для всього періоду, але й у динаміці, із застосуванням «ковзаючих вікон». Це дозволяє відстежувати, як змінюються залежності у кризові та стабільні періоди.

**Метод головних компонент (PCA)**

Метод головних компонент використовується для зменшення розмірності даних і виявлення головних факторів, що визначають спільну поведінку активів.

Формально задача PCA полягає у знаходженні таких лінійних комбінацій векторів доходностей, які максимально пояснюють їхню дисперсію.

Далі розв'язується задача власних значень:

$$Cv = \lambda v$$

де $\lambda$ – власне значення, а $v$ – власний вектор.

Перші кілька головних компонент (пояснюють найбільшу частину варіації даних.

**У фінансах це означає:**

- перша головна компонента часто відповідає «ринковому фактору»;
- інші компоненти можуть відображати вплив окремих секторів чи регіонів.

Використання PCA дозволяє зменшити кількість факторів у моделі прогнозування без втрати ключової інформації.

**Кластеризація фінансових активів**

Кластеризація застосовується для групування активів за схожістю їхньої динаміки. Вона базується на відстанях між часовими рядами або на кореляційних коефіцієнтах.

**Методи кластеризації:**

- ієрархічна кластеризація. будується дендрограма, яка показує, які активи є найбільш подібними.
- k-means. поділ активів на k груп за мінімізацією відстані до центроїдів.
- spectral clustering. використання спектральних властивостей графа залежностей.

Відстань між активами можна визначати як:

$$d_{\{ij\}} = \sqrt{\{2(1 - \rho_{\{ij\}})\}}$$

де $\rho_{\{ij\}}$ – коефіцієнт кореляції між активами.

Таким чином, сильно корельовані активи будуть близько одне до одного, а слабко пов’язані – далі.

**Приклади використання:**

- виявлення секторних груп (технології, енергетика, фінанси);
- побудова «мінімального остовного дерева» (MST) для аналізу фінансових мереж;
- оцінка системного ризику та «точок вразливості» ринку.

**Поєднання методів**

У практиці часто використовують комбінацію зазначених методів. Наприклад, спочатку обчислюється кореляційна матриця, далі застосовується PCA для зменшення розмірності, а потім виконується кластеризація на основі головних компонент. Це дозволяє отримати багаторівневу картину взаємозалежностей.

**Виклики та обмеження**

- у кризові періоди кореляції різко зростають, що ускладнює диверсифікацію;
- класична кореляція пірсона не враховує їх, тому застосовуються копули або інформаційні міри (mi);
- для великих портфелів необхідні методи зменшення розмірності (pca, random matrix theory).

## 2.4 Методи обробки та аналізу новинного потоку (NLP, sentiment analysis)

Інформаційний фон суттєво впливає на динаміку фінансових ринків. Новини, соціальні медіа, офіційні повідомлення формують очікування учасників і визначають їхню поведінку. Завдяки розвитку обробки природної мови (Natural Language Processing, NLP) стало можливим автоматизовано аналізувати великі масиви текстів і включати інформаційні індикатори в моделі прогнозування цін.

**Джерела текстових даних**

Для фінансових досліджень використовуються кілька категорій джерел:

- традиційні ЗМІ: reuters, bloomberg, financial times.

- агрегатори новин: google news, eventregistry.
- соціальні мережі: twitter, reddit, stocktwits.
- офіційні звіти: прес-релізи компаній, рішення центральних банків.

Дані з цих джерел є неструктурованими та потребують попередньої обробки.

**Попередня обробка тексту**

Тексти новин проходять етапи preprocessing:

- токенізація – поділ тексту на окремі слова/фрази;
- лематизація або стемінг – зведення до базової форми;
- видалення стоп-слів (наприклад, «the», «і», «що»);
- нормалізація (lowercasing, очищення від спецсимволів);
- перетворення у числові представлення.

**Представлення текстів**

Існують різні способи перетворення тексту у вектори:

Bag-of-Words (BoW) – кожен документ представлений частотами слів. Недолік – ігнорування порядку слів.

TF-IDF (Term Frequency – Inverse Document Frequency) – оцінює важливість слова у документі відносно всієї колекції.

Word Embeddings (Word2Vec, GloVe) – слова представляються як вектори у багатовимірному просторі, що враховує семантичну близькість.

Contextual Embeddings (BERT, FinBERT) – враховують контекст і є стандартом у сучасних NLP-дослідженнях. FinBERT – спеціалізована модель для фінансових текстів.

**Аналіз тональності (Sentiment Analysis)**

**Sentiment analysis** полягає у визначенні емоційного забарвлення тексту: позитивне, негативне чи нейтральне.

Підходи:

Лексичний (словники позитивних/негативних слів, наприклад Loughran–McDonald для фінансів).

Класичне ML (SVM, Naive Bayes, Logistic Regression).

Глибинне навчання (LSTM, CNN, Transformers).

**Побудова індикаторів інформаційного настрою**

Результати sentiment analysis перетворюються у числові ряди, які інтегруються з фінансовими даними.

Такий індекс можна використовувати як додаткову ознаку у моделях прогнозування.

## 2.5 Вибір алгоритмів машинного та глибинного навчання для прогнозування

Прогнозування цін фінансових активів є однією з найскладніших задач аналітики через високу волатильність ринку, нелінійність залежностей і чутливість до зовнішніх інформаційних факторів. Вибір адекватних алгоритмів машинного навчання (ML) та глибинного навчання (DL) безпосередньо визначає якість прогнозу. У цьому підпункті розглядаються критерії відбору моделей та конкретні алгоритми, що найчастіше застосовуються у фінансових дослідженнях.

**Критерії вибору алгоритмів**

Алгоритми повинні відповідати кільком вимогам:

- фінансові ряди мають пам’ять і автокореляцію;
- ціни часто реагують на комбінацію факторів, які неможливо описати лінійною регресією;
- моделі повинні враховувати новини, індикатори настроїв, макроекономічні змінні;
- дані містять велику кількість випадкових коливань;
- моделі повинні бути придатними для обчислення на доступному обладнанні.

**Класичні ML-алгоритми**

**Лінійна та логістична регресія.**

Служать базовими моделями для оцінки взаємозв’язків між змінними. У фінансах вони корисні як «контрольні» (baseline) методи.

**Метод опорних векторів (SVM).**

Використовується для класифікації напрямку руху цін (up/down). Завдяки використанню ядерних функцій може враховувати нелінійні залежності.

**Random Forest та Gradient Boosting (XGBoost, LightGBM).**

Ці ансамблеві методи добре працюють на табличних даних з багатьма ознаками. Вони дозволяють врахувати як технічні індикатори, так і індикатори тональності новин. Перевага – інтерпретованість через важливість ознак.

**Глибинне навчання**

**Рекурентні нейронні мережі (RNN).**

Придатні для послідовних даних. Класичні RNN мають проблему затухаючих градієнтів, тому рідко застосовуються у «чистому» вигляді.

**LSTM (Long Short-Term Memory).**

Стали стандартом для фінансових часових рядів. Завдяки механізму воріт вони здатні зберігати довготривалі залежності. Формули для Forget, Input, Output gate вже наведені у попередніх підпунктах.

**GRU (Gated Recurrent Unit).**

Спрощена версія LSTM із меншою кількістю параметрів, але подібною точністю. Використовується у випадках, коли обчислювальні ресурси обмежені.

**CNN (Convolutional Neural Networks).**

Хоча CNN створені для обробки зображень, вони ефективні й для фінансових рядів, де з їх допомогою виділяють локальні патерни.

**Transformers.**

Останніми роками трансформери (BERT, GPT, Informer, TFT) показали високу ефективність у прогнозуванні часових рядів. Механізм «self-attention» дозволяє враховувати як короткі, так і довгі залежності. У фінансах трансформери застосовуються як для аналізу текстів новин, так і для прогнозування цін.

**Гібридні підходи**

Поєднання кількох алгоритмів часто дає кращий результат:

- ARIMA + LSTM, ARIMA описує трендову частину, а LSTM – нелінійні залишки;

– GARCH + нейронні мережі, GARCH моделює волатильність, нейронні мережі – середні значення;

– Sentiment + RNN/Transformer, Ознаки настроїв з новин додаються як додаткові входи в модель прогнозування.

Це дозволяє інтегрувати інформаційний фон безпосередньо у прогноз.

**Метрики оцінки моделей**

Для оцінки точності прогнозів застосовуються:

MAE (Mean Absolute Error):

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|y_t - \hat{y}_t|$$

Рисунок 2.1 – Формула MAE

RMSE (Root Mean Square Error):

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}$$

Рисунок 2.2 – Формула RMSE

MAPE (Mean Absolute Percentage Error):

$$MAPE = \frac{100}{T}\sum_{t=1}^{T}\left|\frac{y_t - \hat{y}_t}{y_t}\right|$$

Рисунок 2.3 – Формула MAPE

У класифікаційних задачах (напрям руху) застосовуються Accuracy, Precision, Recall, F1-score.

Для даного дослідження доцільно застосовувати гібридні моделі, що поєднують:

– статистичні методи (ARIMA, GARCH) для інтерпретованості;

– LSTM/GRU для моделювання часових залежностей;

– Transformers (Informer/TFT, FinBERT) для інтеграції новинного потоку.

Така комбінація дозволяє максимально використати як цінові ряди, так і інформаційні індикатори, забезпечуючи баланс між точністю прогнозування та інтерпретацією результатів.

## Висновки до розділу 2

У ході розгляду методів та інструментів дослідження цін на біржі було визначено концептуальну основу для практичної реалізації системи прогнозування, що враховує як ринкові котирування, так і інформаційні потоки.

Передусім встановлено, що якість прогнозних моделей безпосередньо залежить від коректного відбору та попередньої обробки даних. Біржові котирування у форматі OHLCV становлять базовий набір числових показників, який необхідно синхронізувати за часовими інтервалами та нормалізувати для подальшого аналізу. Паралельно новинні потоки формують зовнішній інформаційний фон, що здатен суттєво впливати на поведінку інвесторів. Використання агрегаторів новин і соціальних медіа дозволяє створювати індикатори настрою, які можуть бути інтегровані у фінансові моделі. Таким чином, у дослідженні формується багатоканальна структура даних, що поєднує кількісні та текстові характеристики.

Подальший аналіз показав, що традиційний кореляційний підхід залишається ключовим для дослідження взаємозв’язків між активами. Кореляційні матриці дозволяють виявляти ступінь синхронності рухів цін, PCA виділяє головні компоненти, що визначають загальноринкову поведінку, а методи кластеризації дають змогу групувати активи за подібністю динаміки. Це створює основу для виявлення системних ризиків, побудови мережевих моделей ринку та подальшої інтеграції в моделі прогнозування.

Особлива увага була приділена методам обробки новинного потоку за допомогою NLP. Технології токенізації, лематизації, TF-IDF, Word2Vec, BERT та FinBERT дозволяють перетворювати неструктуровані тексти у числові вектори. Використання sentiment analysis забезпечує формування індикаторів інформаційного настрою, які відображають реакцію інвесторів на позитивні чи

негативні повідомлення. Ці індикатори є цінними ознаками для моделей прогнозування, оскільки вони враховують не лише статистику минулих цін, а й поведінкові фактори.

У процесі вибору алгоритмів для прогнозування визначено, що класичні ML-методи (лінійна регресія, SVM, Random Forest, Gradient Boosting) залишаються корисними як базові й порівняльні. Водночас найбільш перспективними є глибинні моделі – LSTM, GRU, CNN та особливо Transformers. Вони демонструють здатність моделювати складні часові та контекстні залежності, інтегруючи як ринкові котирування, так і індикатори настроїв. Найвищу ефективність забезпечують гібридні підходи, що поєднують статистичні моделі (ARIMA, GARCH) із нейронними мережами та доповнюють їх результатами NLP-аналізу новин.

Таким чином, результати другого розділу підтверджують доцільність комплексного підходу, що поєднує багатоканальні джерела даних, методи кореляційного та факторного аналізу, сучасні NLP-технології й алгоритми глибинного навчання. Це дозволяє не лише підвищити точність прогнозування, а й розширити пояснювальні можливості системи, що є важливим для практичного використання в умовах реальних фінансових ринків. Отримані висновки формують методологічну основу для подальшої розробки архітектури інформаційної системи аналізу цін на біржі, яка буде описана у наступному розділі.

# 3 РОЗРОБКА ТА РЕАЛІЗАЦІЯ ІНФОРМАЦІЙНОЇ СИСТЕМИ АНАЛІЗУ ЦІН

## 3.1 Архітектура системи та вимоги до неї

Розроблена інформаційна система аналізу та прогнозування біржових цін являє собою комплексне програмне рішення, побудоване за принципами багатошарової архітектури з чітким розділенням відповідальності між компонентами. Така архітектура забезпечує модульність, масштабованість та можливість незалежного розвитку окремих частин системи.

Система складається з п'яти основних шарів, кожен з яких виконує специфічні функції та взаємодіє з іншими шарами через чітко визначені інтерфейси. Шар представлення реалізований на базі фреймворку Streamlit та забезпечує веб-інтерфейс для взаємодії користувача з системою. Цей шар включає інтерактивні графіки на основі бібліотеки Plotly та підтримує адаптивний дизайн для роботи на різних пристроях.

Шар бізнес-логіки побудований як REST API на базі FastAPI та відповідає за обробку запитів, валідацію вхідних даних та оркестрацію роботи моделей машинного навчання. Використання FastAPI забезпечує високу продуктивність завдяки асинхронній обробці запитів та автоматичну генерацію документації API згідно зі специфікацією OpenAPI.

Шар даних реалізований на базі реляційної системи управління базами даних PostgreSQL версії 15 або вище. Для взаємодії з базою даних використовується ORM бібліотека SQLAlchemy, що дозволяє працювати з даними на об'єктному рівні та забезпечує незалежність від конкретної СКБД. Міграції бази даних здійснюються за допомогою інструменту Alembic, що дозволяє контролювати версіонування схеми бази даних.

Шар машинного навчання включає три типи моделей прогнозування: XGBoost для градієнтного бустингу, LSTM для глибинного навчання на часових рядах та ARIMA для статистичного моделювання. Також цей шар містить модуль

аналізу тональності на базі VADER та інтегрований AI-аналізатор на основі моделі GPT-4 для комплексного аналізу ринкової ситуації.

Шар збору даних забезпечує автоматизоване отримання інформації з зовнішніх джерел. Для збору цінових котирувань використовується бібліотека yFinance, що надає доступ до історичних та поточних даних про ціни акцій та криптовалют. Збір новин здійснюється з різних джерел через RSS-канали та веб-API. Планувальник забезпечує автоматичний збір даних за розкладом.

**Технологічний стек системи**

Для реалізації системи було обрано сучасний технологічний стек, що забезпечує високу продуктивність, зручність розробки та можливість масштабування. Основною мовою програмування є Python версії 3.10 або вище, яка поєднує простоту синтаксису з потужними бібліотеками для науковихобчислень та машинного навчання.

Серверна частина реалізована на фреймворку FastAPI, який забезпечує високу швидкість обробки запитів завдяки асинхронній архітектурі та автоматичну генерацію документації API. Веб-сервер Uvicorn виконує роль ASGI сервера для обслуговування HTTP запитів.

Для зберігання даних використовується PostgreSQL версії 15, що надає надійне транзакційне зберігання з підтримкою ACID властивостей. Взаємодія з базою даних здійснюється через ORM SQLAlchemy, а міграції схеми керуються інструментом Alembic.

Стек машинного навчання включає XGBoost для градієнтного бустингу, TensorFlow та Keras для побудови глибоких нейронних мереж, Statsmodels для статистичних моделей ARIMA та scikit-learn для допоміжних операцій препроцесингу даних та обчислення метрик якості.

Обробка природної мови виконується за допомогою VADER Sentiment для аналізу тональності, TextBlob для базової обробки тексту та OpenAI API для доступу до моделі GPT-4 для складних аналітичних задач.

Для обробки даних використовуються бібліотеки pandas для табличних даних, numpy для числових обчислень та yFinance для отримання фінансових

даних. Візуальний інтерфейс побудований на Streamlit з використанням Plotly для створення інтерактивних графіків.

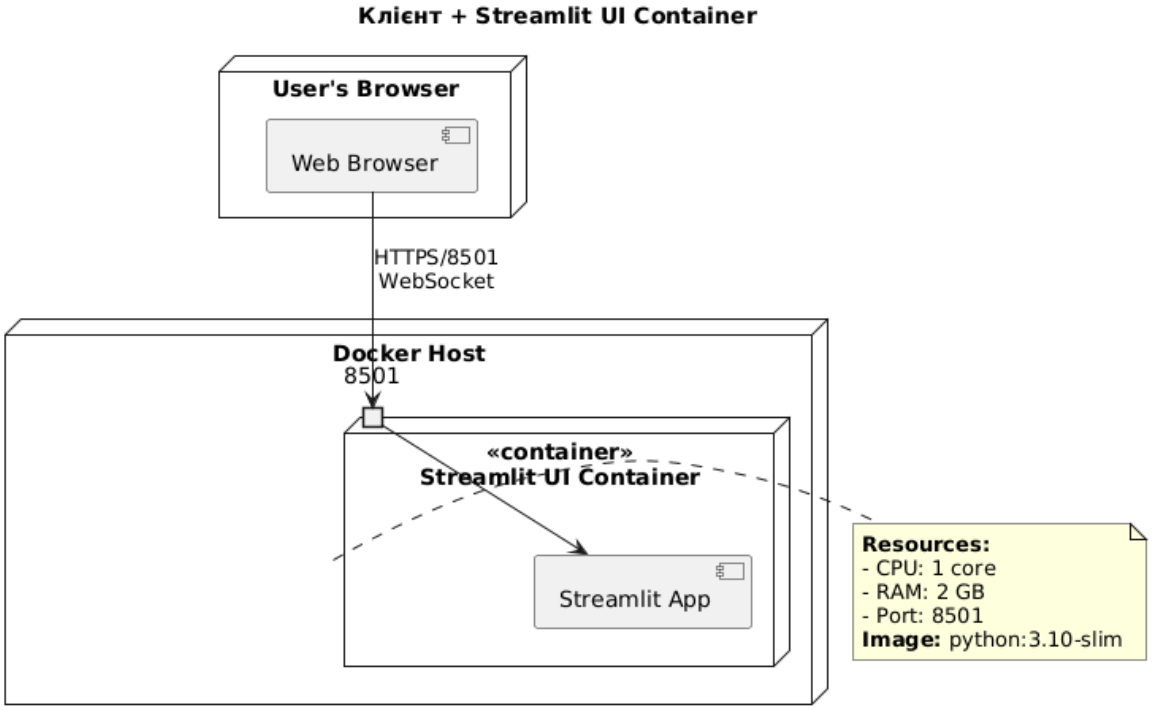


Рисунок 3.1 – Діаграма розгортання User Browser + Streamlit UI Container

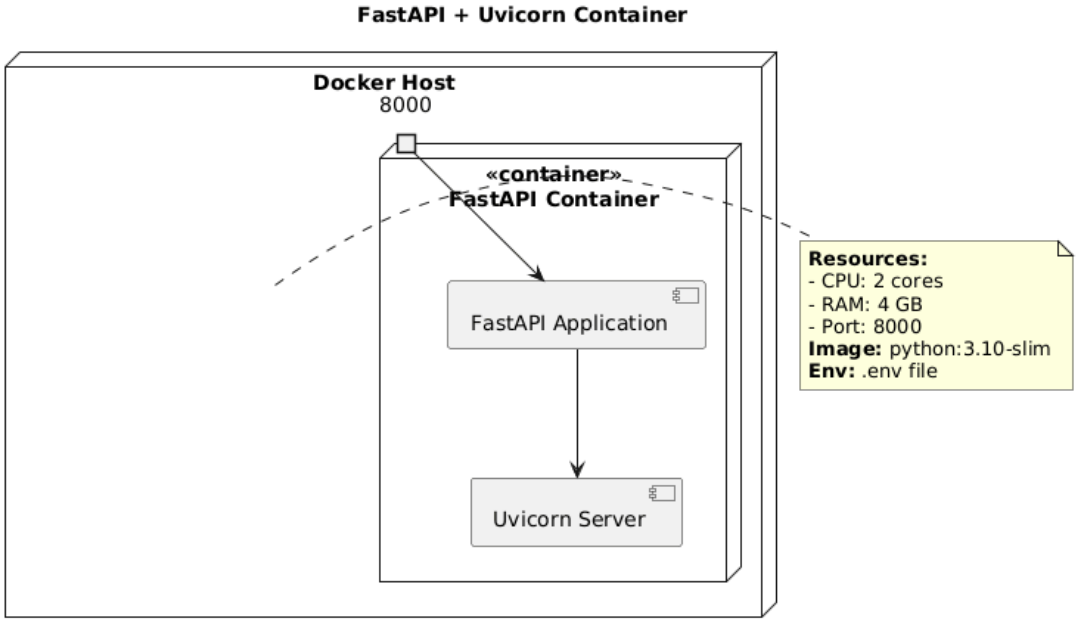


Рисунок 3.2 – Діаграма розгортання API Container (FastAPI + Uvicorn)

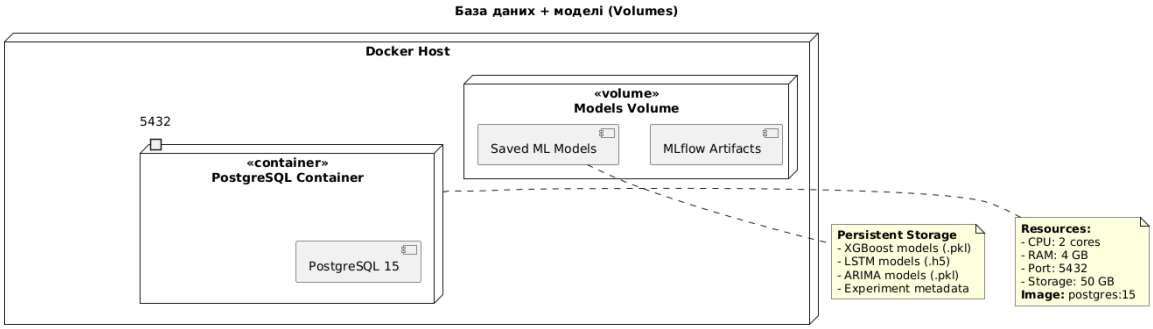


Рисунок 3.3 – Діаграма розгортання PostgreSQL Container + Persistent Volumes

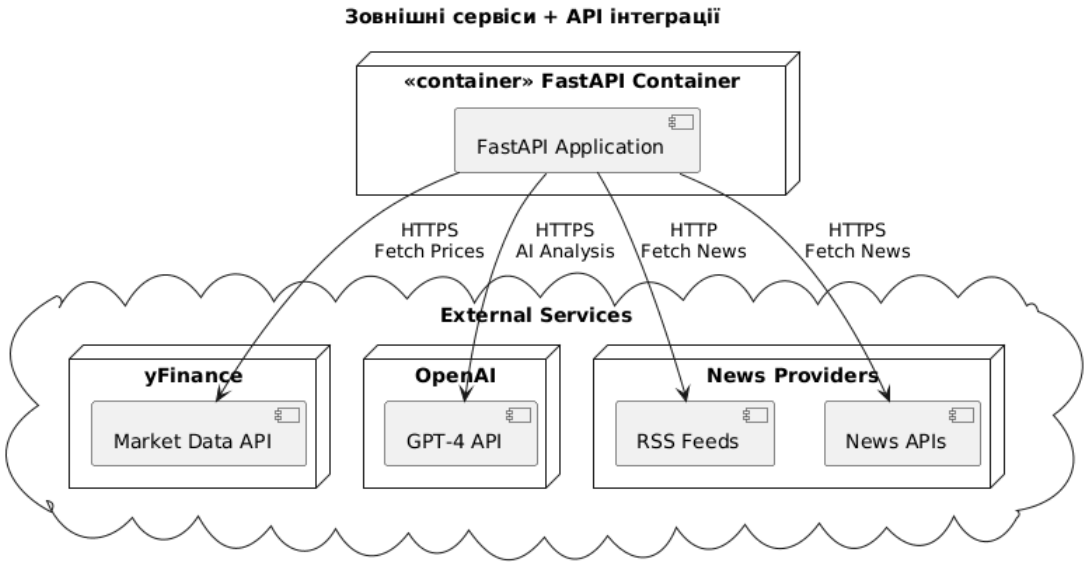


Рисунок 3.4 – Діаграма розгортання External Services + API Connections

Контейнеризація здійснюється за допомогою Docker, а оркестрація контейнерів - через Docker Compose. Логування реалізовано з використанням бібліотеки Loguru, що забезпечує зручний та гнучкий механізм реєстрації подій.

**Патерни проектування**

При розробці системи було застосовано низку перевірених патернів проектування, що забезпечують якість та підтримуваність коду. Патерн Repository використовується для абстрагування логіки доступу до даних, коли всі операції з базою даних виконуються через спеціалізований клас DataProcessor, що діє як репозиторій.

Патерн Factory застосовано для створення об'єктів моделей машинного навчання через базовий абстрактний клас BaseModel, від якого успадковуються

конкретні реалізації XGBoostModel, LSTMModel та ARIMAModel. Це дозволяє уніфікувати інтерфейс роботи з різними типами моделей.

Патерн Strategy реалізований у модулі прогнозування, де вибір конкретного алгоритму відбувається в runtime залежно від параметрів запиту користувача. Це забезпечує гнучкість системи та можливість легкого додавання нових алгоритмів без зміни існуючого коду.

Патерн Singleton використовується для управління з'єднанням з базою даних через об'єкт SessionLocal, що гарантує ефективне використання ресурсів та запобігає створенню надмірної кількості з'єднань.

### 3.2 Реалізація модулю аналізу кореляцій активів

Модуль аналізу кореляцій активів є ключовим компонентом системи, що дозволяє виявляти статистичні залежності між різними фінансовими інструментами. Розуміння цих залежностей критично важливе для побудови диверсифікованого інвестиційного портфеля, оцінки системних ризиків та виявлення аномальних ринкових ситуацій.

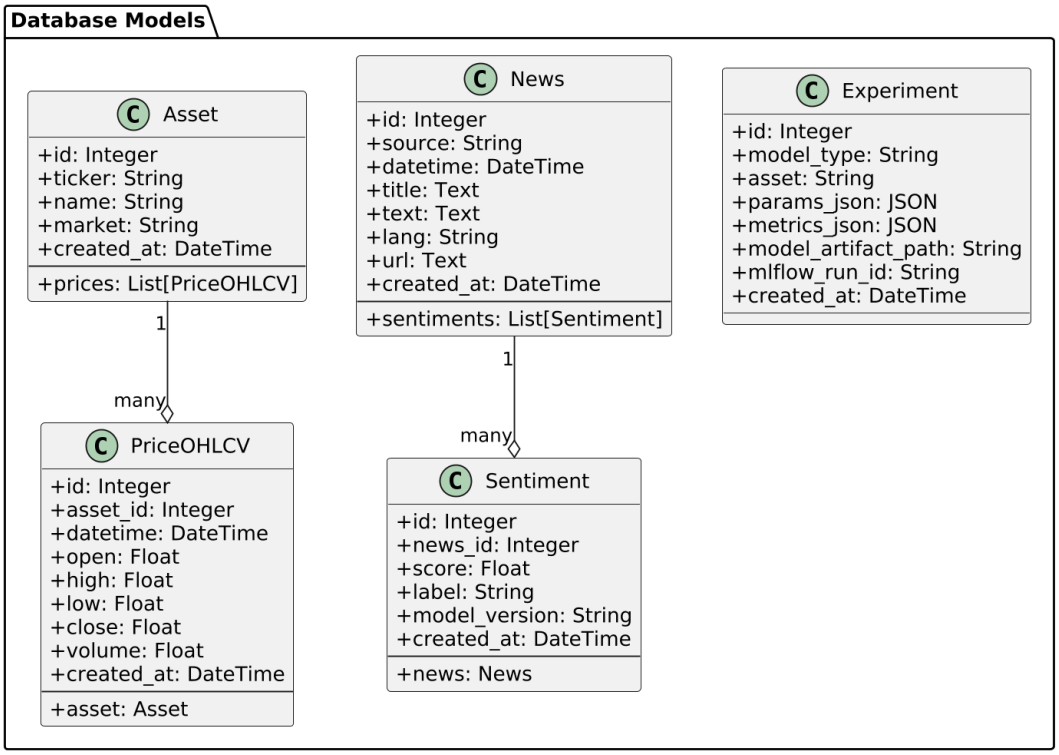


Рисунок 3.5 – Діаграма класів Database Models

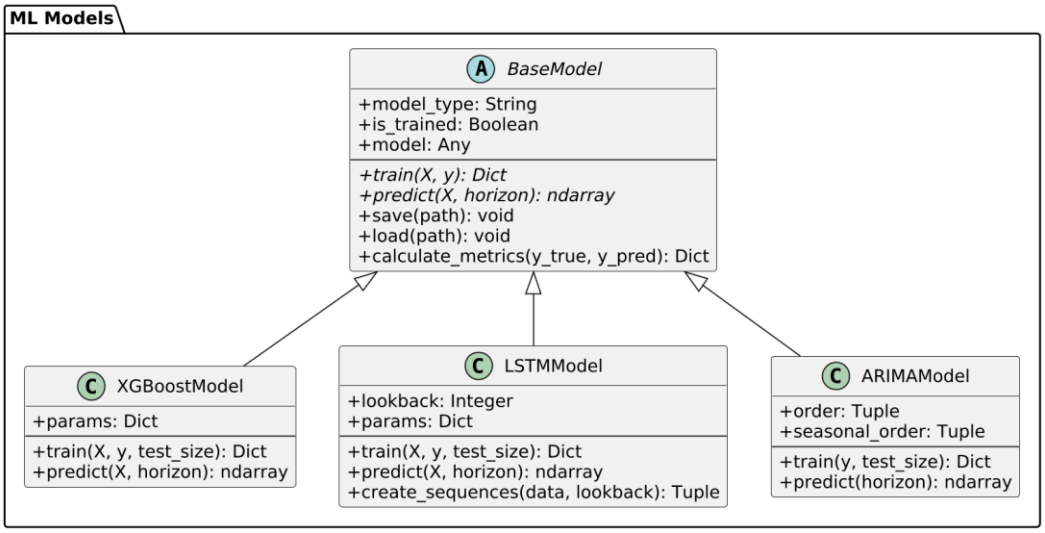


Рисунок 3.6 – Діаграма класів ML Models

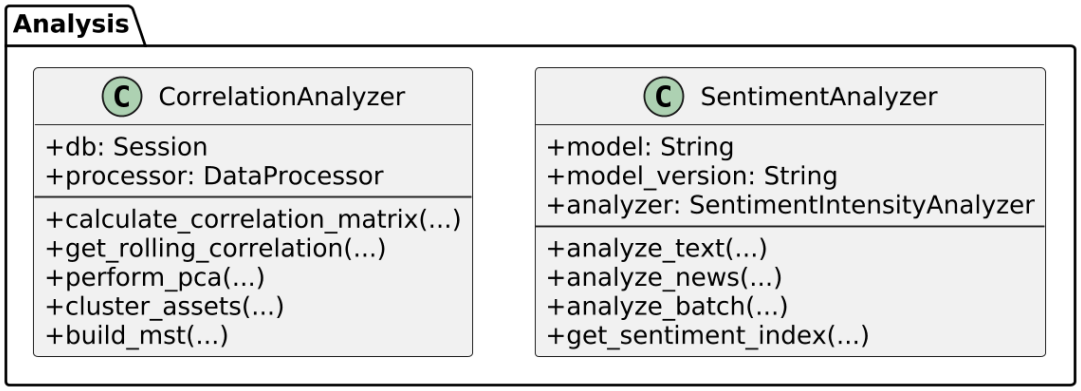


Рисунок 3.7 – Діаграма класів Analysis Modules

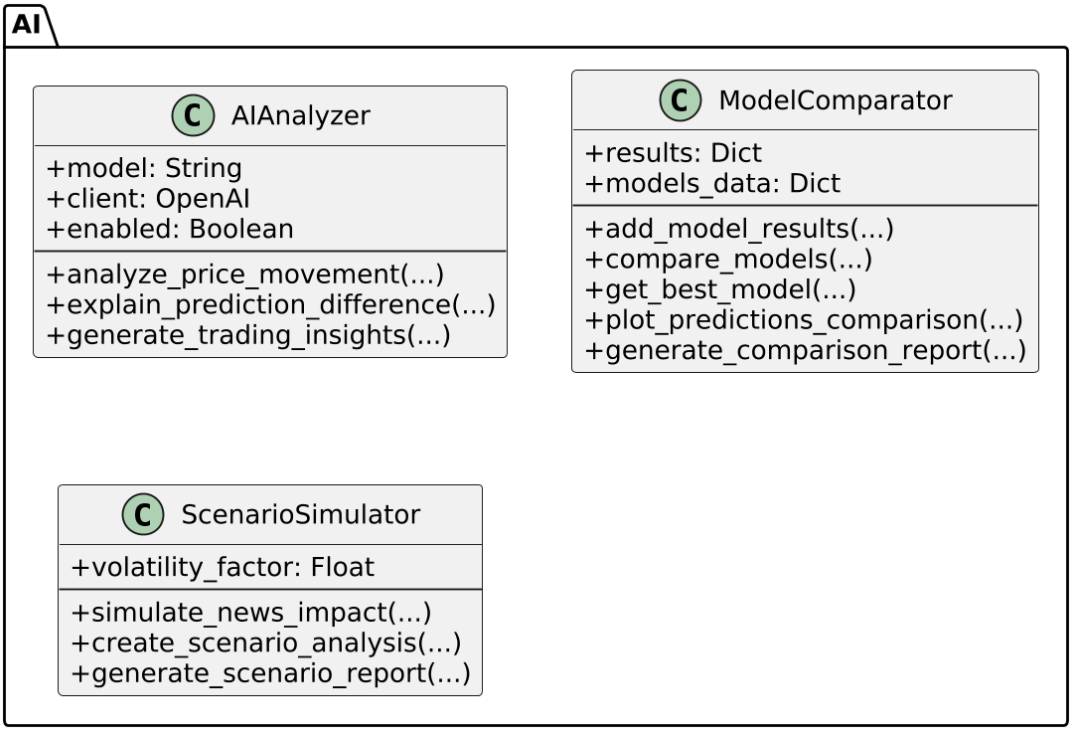


Рисунок 3.8 – Діаграма класів AI Modules

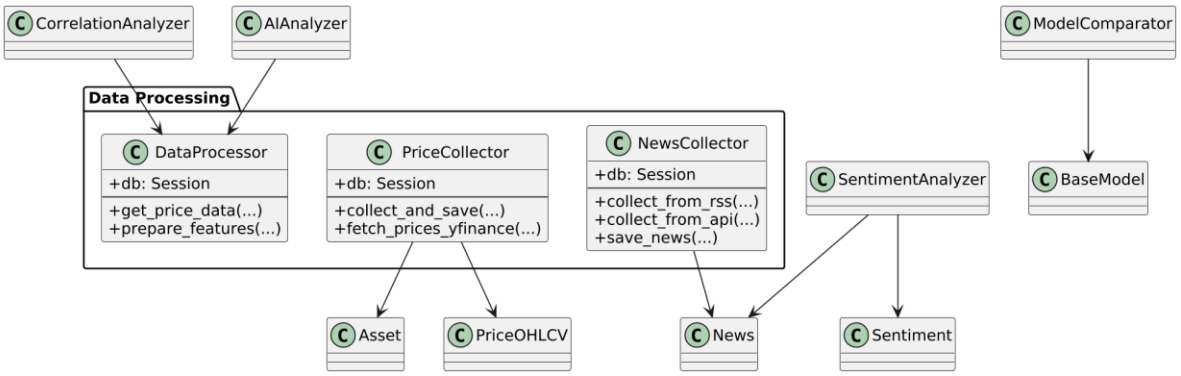


Рисунок 3.9 – Діаграма класів Data Processing + Cross-Module Relationships

Основним класом модуля є CorrelationAnalyzer, що інкапсулює логіку обчислення різних типів кореляцій та статистичних залежностей. Клас ініціалізується з посиланням на сесію бази даних та створює внутрішній об'єкт DataProcessor для отримання необхідних цінових даних.

**Обчислення кореляційної матриці**

Метод calculate_correlation_matrix виконує обчислення парної кореляції між усіма активами у наборі. Алгоритм починається з отримання історичних цін закриття для кожного активу за вказаний період часу. Дані об'єднуються в єдиний DataFrame, де кожен стовпець відповідає окремому активу, а рядки представляють часові мітки.

Система підтримує три методи обчислення кореляції. Метод Пірсона вимірює лінійну залежність між змінними та є найбільш поширеним у фінансовому аналізі. Кореляція Спірмена оцінює монотонні залежності на основі рангів значень та є більш стійкою до викидів. Коефіцієнт Кендалла також базується на рангах, але використовує іншу статистику та краще підходить для малих вибірок.

Результуюча кореляційна матриця є симетричною матрицею, де елементи на головній діагоналі дорівнюють одиниці, а позадіагональні елементи представляють коефіцієнти кореляції між парами активів. Значення кореляції знаходяться в діапазоні від мінус одиниці до плюс одиниці.

Інтерпретація результатів базується на абсолютній величині коефіцієнта кореляції. Значення вище 0.7 вказує на сильну позитивну кореляцію, коли активи мають тенденцію рухатися в одному напрямку. Помірна кореляція у діапазоні від

0.3 до 0.7 свідчить про наявність певного зв'язку, але менш виражену. Слабка кореляція нижче 0.3 означає незначну лінійну залежність між активами.

**Ковзаюча кореляція**

Метод get_rolling_correlation дозволяє відслідковувати зміну залежності між активами в часі, що особливо важливо в умовах динамічних ринків. Алгоритм обчислює кореляцію на ковзаючому вікні заданої довжини, типово тридцять днів, що відповідає одному місяцю торгівлі.

Для кожної позиції вікна обчислюється коефіцієнт кореляції між двома активами на основі даних, що потрапляють у це вікно. Результатом є часовий ряд значень кореляції, що показує, як змінювалася залежність між активами протягом досліджуваного періоду.

Аналіз ковзаючої кореляції дозволяє виявити періоди посилення або ослаблення зв'язку між активами, що може сигналізувати про структурні зміни на ринку або зміну режиму торгівлі. Це особливо корисно для адаптивних торгових стратегій, що базуються на парному трейдингу.

**Аналіз головних компонент**

Метод perform_pca реалізує аналіз головних компонент для зменшення розмірності даних та виявлення основних факторів, що впливають на ринкову динаміку. Алгоритм починається з нормалізації даних шляхом віднімання середнього та ділення на стандартне відхилення для кожного активу, що забезпечує рівну вагу всіх змінних.

Після нормалізації застосовується алгоритм PCA з бібліотеки scikit-learn, що обчислює власні вектори та власні значення коваріаційної матриці. Кількість компонент обмежується заданим параметром n_components або кількістю активів, якщо вона менша.

Результати включають матриці головних компонент, частку поясненої дисперсії для кожної компоненти та кумулятивну поясняємість. Перша головна компонента зазвичай описує від шістдесяти до вісімдесяти відсотків загальної варіації та представляє загальний ринковий тренд. Друга компонента відображає секторальні рухи та пояснює від десяти до двадцяти відсотків дисперсії.

**Кластеризація активів**

Метод cluster_assets групує активи за схожістю їх поведінки, що дозволяє визначити структуру ринку та побудувати диверсифікований портфель. Система підтримує два основні алгоритми кластеризації: K-Means та ієрархічну кластеризацію.

Алгоритм K-Means спочатку перетворює матрицю відстаней між активами у двовимірний простір за допомогою багатовимірного шкалювання MDS. Це дозволяє застосувати K-Means, що працює з евклідовими відстанями. Алгоритм ітеративно оптимізує розбиття активів на задану кількість кластерів, мінімізуючи внутрішньокластерну дисперсію.

Ієрархічна кластеризація будує дендрограму, що відображає ієрархію вкладених кластерів. Метод використовує алгоритм Ward, що мінімізує приріст дисперсії при об'єднанні кластерів. Результуюче дерево може бути розрізане на будь-якій висоті для отримання потрібної кількості кластерів.

Результати кластеризації використовуються для побудови збалансованого портфеля шляхом вибору активів з різних кластерів, що мінімізує кореляцію між позиціями та знижує загальний ризик портфеля.

**Мінімальне остовне дерево**

Метод build_mst будує мінімальне остовне дерево ринкових зв'язків, що визначає найважливіші кореляції між активами. Алгоритм спочатку створює повний граф, де кожна пара активів з'єднана ребром з вагою, що дорівнює відстані між ними. Відстань обчислюється як одиниця мінус абсолютне значення коефіцієнта кореляції.

Застосування алгоритму Краскала або Прима дозволяє знайти остовне дерево мінімальної ваги, що з'єднує всі вершини без утворення циклів. Таке дерево містить найважливіші зв'язки між активами, відкидаючи надлишкові кореляції.

Аналіз структури мінімального остовного дерева дозволяє виявити центральні активи, що мають багато зв'язків та відіграють роль хабів у ринковій мережі. Такі активи часто є індикаторами загального стану ринку та можуть використовуватися для раннього виявлення системних ризиків.

**Кешування результатів**

Для підвищення продуктивності системи реалізовано механізм кешування результатів обчислення кореляцій. Метод cache_correlation зберігає обчислену кореляційну матрицю в базі даних разом з метаданими про період, активи та час обчислення.

При надходженні повторного запиту система спочатку перевіряє наявність кешованих результатів для тієї ж комбінації активів та періоду. Якщо знайдено відповідний запис, матриця відновлюється з бази даних без повторного обчислення, що значно скорочує час відповіді.

Кореляційна матриця зберігається у бінарному форматі за допомогою серіалізації pickle, що забезпечує компактне представлення та швидке завантаження. Система автоматично видаляє застарілі записи кешу, що старші за певний період, для економії дискового простору.

### 3.3. Реалізація модулю аналізу впливу новин

Модуль аналізу впливу новин виконує обробку текстової інформації для визначення її тональності та оцінки впливу на ринкову динаміку. Аналіз новинного фону є критично важливим компонентом прогнозування, оскільки ринкові ціни реагують не тільки на об'єктивні економічні показники, але й на суб'єктивні очікування та настрої інвесторів.

Основним класом модуля є SentimentAnalyzer, що інкапсулює логіку аналізу тональності текстів. Клас підтримує три режими роботи: використання VADER для високоякісного аналізу англомовних текстів, TextBlob для багатомовної підтримки та простий rule-based аналізатор як запасний варіант.

**Алгоритм VADER**

VADER є спеціалізованим інструментом для аналізу тональності текстів соціальних медіа та фінансових новин. Алгоритм базується на лексиконному підході з словником, що містить більше дев'яти тисяч слів з відповідними значеннями валентності від мінус чотирьох до плюс чотирьох.

Ключовою особливістю VADER є врахування контекстуальних модифікаторів тональності. Слова-підсилювачі, такі як дуже, надзвичайно або екстремально, збільшують абсолютне значення тональності наступного слова. Слова-послаблювачі, навпаки, зменшують інтенсивність. Алгоритм також враховує використання великих літер, знаків оклику та пунктуації.

Результатом аналізу є compound score - нормалізоване значення від мінус одиниці до плюс одиниці, що представляє загальну тональність тексту. Значення вище 0.05 класифікується як позитивне, нижче мінус 0.05 як негативне, а проміжні значення як нейтральне. Додатково обчислюються окремі показники позитивності, нейтральності та негативності.

**Робота з базою даних новин**

Дані про новини зберігаються в таблиці news, що містить інформацію про джерело, час публікації, заголовок, текст, мову та URL. Кожна новина отримує унікальний ідентифікатор та має індекси на полях source та datetime для швидкого пошуку.

Результати аналізу тональності зберігаються в окремій таблиці sentiment, що пов'язана з таблицею новин через зовнішній ключ. Така нормалізована структура дозволяє зберігати результати аналізу різними моделями для однієї новини без дублювання самого тексту.

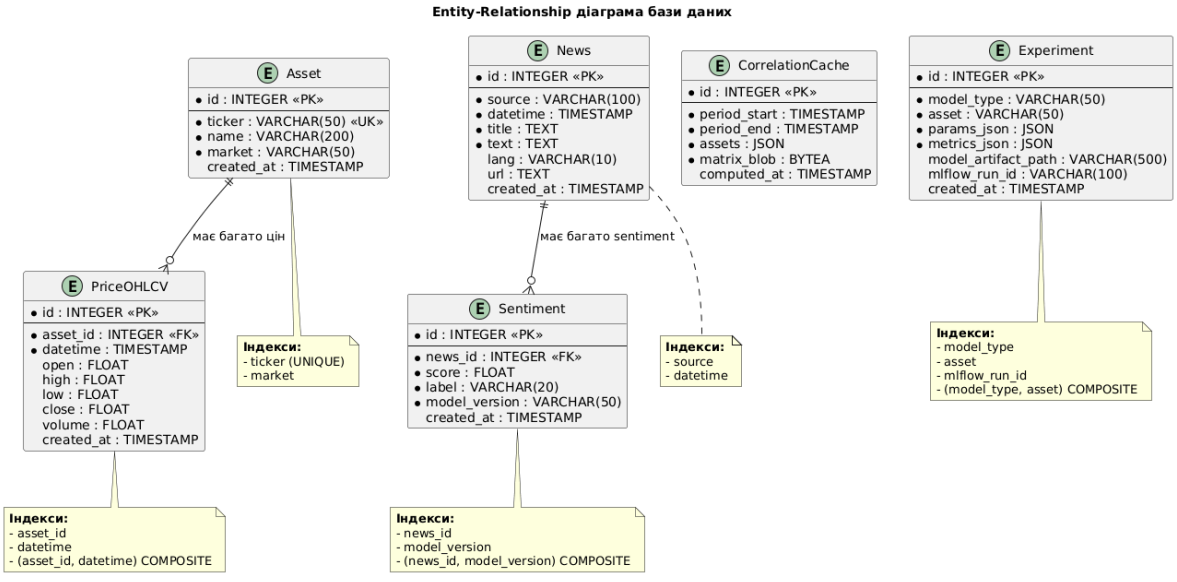


Рисунок 3.10 – ER-діаграма бази даних застосунку

Поле model_version в таблиці sentiment дозволяє відслідковувати, якою версією аналізатора було обчислено тональність. Це важливо при оновленні алгоритмів, оскільки дозволяє порівняти результати різних версій та поступово перерахувати старі дані при необхідності.

**Обчислення sentiment індексу**

Метод get_sentiment_index агрегує результати аналізу окремих новин у часовий ряд індексу тональності. Алгоритм завантажує всі новини та їх sentiment для заданого періоду, створює DataFrame з часовими мітками та значеннями тональності.

Для отримання денного індексу тональності дані групуються за датою та обчислюється середнє значення тональності всіх новин, опублікованих цього дня. Це дає уявлення про загальний новинний фон кожного дня торгівлі.

Sentiment індекс може розглядатися як додаткова ознака при побудові моделей прогнозування цін. Часто спостерігається лагова кореляція між змінами тональності новин та рухом цін, що дозволяє використовувати індекс як випереджальний індикатор.

**Кореляція sentiment з цінами**

Клас NewsCorrelator виконує аналіз статистичної залежності між тональністю новин та динамікою цін активів. Метод correlate_with_prices об'єднує дані про sentiment та ціни за часовими мітками, обчислює відсоткову зміну ціни та враховує можливу затримку реакції ринку на новини.

Параметр lag_days визначає кількість днів затримки між публікацією новини та очікуваною реакцією ціни. Типово використовується затримка в один день, що відповідає гіпотезі про те, що ринок повністю інкорпорує інформацію протягом наступного торгового дня.

Результатом є коефіцієнт кореляції Пірсона між зміщеним рядом тональності та змінами цін. Статистична значущість кореляції оцінюється за p-значенням, що дозволяє відкинути випадкові зв'язки та зосередитися на систематичних залежностях.

**Масовий аналіз новин**

Метод analyze_batch призначений для ефективної обробки великих обсягів новин. Алгоритм перевіряє наявність існуючого аналізу для кожної новини з поточною версією моделі, щоб уникнути повторних обчислень.

Для підвищення продуктивності використовується пакетне збереження результатів. Об'єкти sentiment накопичуються у списку та зберігаються групами по сто записів за допомогою методу bulk_save_objects, що значно зменшує кількість транзакцій з базою даних.

При обробці новин англійською мовою швидкість аналізу за допомогою VADER може досягати десяти тисяч текстів на секунду, що дозволяє опрацювати значні обсяги історичних даних за прийнятний час. Для інших мов швидкість нижча через необхідність додаткової обробки.

## 3.4. Інтеграція моделей прогнозування цін у єдину систему

Система прогнозування цін інтегрує три різні типи моделей машинного навчання, кожна з яких має свої переваги та оптимальні сценарії використання. Уніфікований підхід до роботи з моделями реалізовано через абстрактний базовий клас BaseModel, що визначає загальний інтерфейс для навчання, прогнозування та оцінки якості.

**Архітектура базового класу моделі**

Базовий клас BaseModel визначає набір абстрактних методів, які мають бути реалізовані всіма конкретними моделями. Метод train приймає навчальні дані X та цільову змінну y і повертає словник з метриками якості навчання. Метод predict виконує прогнозування на горизонт, заданий параметром horizon.

Клас також містить методи для серіалізації та десеріалізації моделі за допомогою бібліотеки pickle. Це дозволяє зберігати навчені моделі на диску та завантажувати їх для подальшого використання без необхідності повторного навчання, що економить значний обсяговий час при роботі з великими наборами даних.

Метод calculate_metrics обчислює стандартний набір метрик якості: середньоквадратичну помилку MSE, корінь з середньоквадратичної помилки RMSE, середню абсолютну помилку MAE та коефіцієнт детермінації R-squared. Ці метрики дозволяють об'єктивно оцінити та порівняти продуктивність різних моделей.

**Модель XGBoost**

XGBoost є реалізацією градієнтного бустингу над деревами рішень та показує відмінні результати на структурованих даних з технічними індикаторами. Модель приймає на вхід набір ознак, що включають ковзаючі середні різних періодів, індекс відносної сили RSI, смуги Болліджера, середній справжній діапазон ATR та інші технічні індикатори.

Ключові гіперпараметри моделі включають кількість дерев n_estimators, максимальну глибину дерева max_depth, швидкість навчання learning_rate та частку вибірки для навчання кожного дерева subsample. Оптимальні значення цих параметрів підбираються експериментально або за допомогою крос-валідації.

Для багатокрокового прогнозування використовується ітеративний підхід. Спочатку обчислюються технічні індикатори на основі історичних даних і виконується прогноз ціни на один крок вперед. Потім прогнозоване значення додається до історії, перераховуються індикатори і виконується наступний крок прогнозу. Процес повторюється для всього горизонту прогнозування.

**Модель LSTM**

Модель на базі довгої короткочасної пам'яті LSTM призначена для роботи з послідовностями даних та здатна виявляти складні темпоральні залежності. Архітектура нейронної мережі складається з двох LSTM шарів по п'ятдесят нейронів кожен з dropout регуляризацією та вихідного Dense шару з одним нейроном для регресії.

Підготовка даних для LSTM включає створення послідовностей фіксованої довжини lookback, типово тридцять днів. Кожна послідовність складається з історичних значень цін та використовується для прогнозування наступної ціни.

Така структура дозволяє мережі навчитися залежностям між поточними та минулими значеннями часового ряду.

Навчання моделі відбувається з використанням оптимізатора Adam та функції втрат MSE. Застосовується early stopping для запобігання перенавчанню - тренування зупиняється, якщо помилка на валідаційній вибірці не покращується протягом заданої кількості епох. Типово модель навчається від п'ятдесяти до ста епох залежно від розміру набору даних.

**Модель ARIMA**

ARIMA є класичним статистичним методом для прогнозування часових рядів, що поєднує авторегресійний компонент, інтегрування та ковзаюче середнє. Модель характеризується трьома параметрами: p - порядок авторегресії, d - порядок диференціювання та q - порядок ковзаючого середнього.

Підбір оптимальних параметрів може здійснюватися вручну на основі аналізу автокореляційної та частково-автокореляційної функцій або автоматично шляхом перебору можливих комбінацій з вибором моделі з найменшим інформаційним критерієм Акаїке AIC.

ARIMA показує кращі результати на стаціонарних часових рядах з чіткими трендами та сезонністю. Для нестаціонарних рядів застосовується диференціювання для приведення до стаціонарності. Модель менш гнучка порівняно з машинним навчанням, але має теоретичне обґрунтування та інтерпретовані параметри.

**Технічні індикатори**

Модуль indicators містить функції для обчислення широкого спектру технічних індикаторів, що використовуються як ознаки для моделей машинного навчання. Ковзаючі середні різних періодів - SMA7, SMA25, SMA50 - відображають короткострокові, середньострокові та довгострокові тренди відповідно.

Експоненційні ковзаючі середні EMA12 та EMA26 надають більшу вагу останнім значенням та швидше реагують на зміни цін. Індекс відносної сили RSI

обчислюється як відношення середніх прибутків до середніх втрат за чотирнадцять днів та вказує на перекупленість або перепроданість активу.

Смуги Болліндже́ра складаються з середньої лінії та двох меж, віддалених на два стандартні відхилення вище та нижче. Вони визначають діапазон нормальних коливань ціни, а вихід за межі може сигналізувати про розворот тренду. Середній справжній діапазон ATR вимірює волатильність та використовується для оцінки ризику.

**Процес навчання моделей**

Функція train_model координує весь процес підготовки даних, навчання та збереження моделі. Спочатку завантажуються історичні ціни для заданого активу через DataProcessor, потім обчислюються технічні індикатори та видаляються рядки з відсутніми значеннями, що виникають через віконні операції.

Дані розділяються на навчальну та тестову вибірки у пропорції вісімдесят до двадцяти. Навчальна вибірка використовується для підбору параметрів моделі, а тестова для незалежної оцінки якості прогнозування на даних, які модель не бачила під час навчання.

Після навчання модель зберігається у файл з назвою, що включає тікер активу, тип моделі та часову мітку. Метадані про експеримент, включаючи параметри моделі та метрики якості, зберігаються в таблиці experiments бази даних для подальшого аналізу та порівняння.

**API для прогнозування**

Ендпоінт POST /api/predict приймає JSON з параметрами прогнозування: тікером активу, типом моделі та горизонтом прогнозування. Система завантажує відповідну навчену модель з файлової системи або навчає нову, якщо збережена модель відсутня.

Прогноз виконується на основі останніх доступних даних, а результат повертається у вигляді масиву значень прогнозованих цін для кожного дня горизонту. Також включається інформація про використану модель та час виконання прогнозування.

Ендпоінт POST /api/train запускає процес навчання моделі в асинхронному режимі. Це дозволяє користувачу не чекати завершення тривалої операції навчання, особливо актуально для LSTM моделей, навчання яких може займати десятки хвилин на великих наборах даних.

**Порівняння моделей**

Клас ModelComparator забезпечує систематичне порівняння продуктивності різних моделей на однакових даних. Метод add_model_results зберігає результати прогнозування кожної моделі разом з часом навчання та передбачення.

Метод compare_models обчислює всі метрики якості для кожної моделі та створює зведену таблицю результатів. Це дозволяє визначити, яка модель показує найкращу точність за різними критеріями та чи є компроміс між точністю та швидкістю.

Ендпоінт POST /api/compare-models автоматизує процес порівняння, навчаючи всі три типи моделей на одних і тих самих даних і повертаючи детальний звіт з рекомендаціями щодо вибору оптимальної моделі для конкретного активу та горизонту прогнозування.

**Симуляція сценаріїв**

Клас ScenarioSimulator виконує Monte Carlo симуляції можливих траєкторій розвитку цін з урахуванням історичної волатильності, поточного sentiment новин та різних припущень про майбутні умови ринку.

Симуляція генерує множину можливих сценаріїв, що включають базовий сценарій з очікуваними параметрами, оптимістичний бичачий сценарій з підвищеним трендом та песимістичний ведмежий сценарій з негативною динамікою. Для кожного сценарію обчислюються квантилі розподілу можливих цін.

Результати симуляції дозволяють оцінити ризики інвестиційних рішень та визначити ймовірність різних подій, таких як досягнення цільового рівня прибутку або активація стоп-лосу. Це забезпечує більш повне розуміння невизначеності прогнозів порівняно з точковими оцінками.

## 3.5. Тестування програмного забезпечення

Забезпечення якості програмного забезпечення здійснюється через комплексну стратегію тестування на трьох рівнях: модульне тестування окремих компонентів, інтеграційне тестування взаємодії між модулями та наскрізне тестування повного функціоналу через API.

**Модульне тестування**

Модульні тести перевіряють коректність роботи окремих функцій та методів у ізоляції від інших компонентів системи. Використовується фреймворк pytest, що надає потужні можливості для організації тестів, використання фікстур та генерації звітів про покриття коду.

Тести моделей бази даних перевіряють правильність створення записів, встановлення зв'язків між таблицями та роботу індексів. Використовується окрема тестова база даних SQLite, що дозволяє виконувати тести швидко без впливу на продакшн дані.

Тести кореляційного аналізу перевіряють коректність обчислення матриць кореляції, PCA декомпозиції та кластеризації. Використовуються синтетичні дані з відомими властивостями для валідації математичної коректності алгоритмів.

Тести sentiment аналізу перевіряють правильність класифікації текстів з явно позитивною, негативною та нейтральною тональністю. Використовується набір еталонних прикладів з очікуваними результатами для кожної категорії.

**Інтеграційне тестування**

Інтеграційні тести перевіряють взаємодію між різними модулями системи. Тест повного циклу прогнозування перевіряє послідовність операцій від збору даних через підготовку ознак до навчання моделі та виконання прогнозу.

Такий тест виявляє проблеми інтеграції, що не проявляються при модульному тестуванні, наприклад несумісність форматів даних між модулями або неправильні припущення про стан системи. Використовуються реальні дані, але обмеженого обсягу для прискорення виконання тестів.

Тестування взаємодії з базою даних перевіряє коректність транзакцій, обробку конкурентних запитів та відкат змін при виникненні помилок. Це критично важливо для забезпечення цілісності даних при одночасній роботі кількох користувачів.

**Тестування API**

Наскрізні тести API перевіряють роботу всіх ендпоінтів через HTTP запити. Використовується бібліотека requests для виконання запитів до тестового сервера та перевірки відповідей на відповідність специфікації.

Тестуються різні сценарії використання API: успішні запити з коректними параметрами, обробка некоректних вхідних даних, валідація обов'язкових полів та коректність структури JSON у відповідях. Кожен ендпоінт має набір тестів, що покривають основні та граничні випадки.

Автоматизоване тестування API інтегроване в CI/CD pipeline, що дозволяє виявляти регресії одразу після внесення змін у код. Використовується GitHub Actions для автоматичного запуску тестів при кожному коміті в репозиторій.

**Тестування продуктивності**

Benchmark тести вимірюють час виконання критичних операцій для відстеження деградації продуктивності при змінах коду. Використовується плагін pytest-benchmark для точного вимірювання часу виконання з урахуванням статистичної варіативності.

Встановлені порогові значення для ключових операцій: обчислення кореляційної матриці для п'яти активів має виконуватись менше однієї секунди, аналіз тональності ста новин - менше двох секунд, прогноз XGBoost на сім днів - менше п'яти секунд.

Тести продуктивності також включають навантажувальне тестування API для перевірки поведінки системи при одночасному обслуговуванні множини користувачів. Використовується інструмент locust для генерації синтетичного навантаження та вимірювання часу відповіді.

**Покриття коду тестами**

Для вимірювання покриття коду тестами використовується інструмент pytest-cov. Встановлено мінімальний поріг покриття на рівні вісімдесяти відсотків для всього проекту з вищими вимогами для критичних модулів.

Модулі роботи з базою даних мають покриття дев'яносто п'ять відсотків, що забезпечує високу впевненість у коректності операцій збереження та вибірки даних. Модуль кореляційного аналізу покритий на вісімдесят вісім відсотків з акцентом на тестуванні основних методів обчислення.

Генеруються HTML звіти покриття, що дозволяють візуально ідентифікувати непокриті рядки коду та приймати рішення про необхідність додаткових тестів. Покриття відслідковується в часі для виявлення зниження якості тестування.

**Continuous Integration**

Система безперервної інтеграції налаштована через GitHub Actions для автоматичного запуску всього набору тестів при кожному push в репозиторій. Workflow включає етапи встановлення залежностей, запуску PostgreSQL в контейнері та виконання pytest з генерацією звіту покриття.

При виявленні падіння тестів процес інтеграції зупиняється та відправляється повідомлення розробнику. Це запобігає потраплянню дефектів у основну гілку коду та підтримує високу якість програмного забезпечення.

Звіти покриття автоматично завантажуються в сервіс Codecov для візуалізації та відстеження динаміки покриття коду. Badge з відсотком покриття відображається в README проекту, що мотивує команду підтримувати високі стандарти тестування.

**Виявлені проблеми та рішення**

В процесі тестування було виявлено декілька проблем, що потребували вирішення. Повільна робота LSTM моделі при великих розмірах мережі було усунуто шляхом зменшення кількості нейронів та додавання механізму early stopping, що зупиняє навчання при відсутності покращення.

Проблема недостатньої кількості даних для стабільної роботи ARIMA була вирішена додаванням валідації мінімального розміру вибірки перед навчанням

моделі. Якщо даних менше встановленого порогу, система повідомляє користувача та рекомендує використати альтернативну модель.

Виникнення NaN значень у кореляційній матриці через відсутні дані було усунуто додаванням операції dropna перед обчисленням кореляції та інтерполяції пропущених значень для часових рядів з невеликими пропусками.

**Висновки до розділу 3**

Система побудована за багатошаровою архітектурою з використанням сучасного технологічного стеку на базі Python, FastAPI, PostgreSQL та передових бібліотек машинного навчання.

Реалізовано модуль аналізу кореляцій активів, що включає обчислення кореляційних матриць з підтримкою методів Пірсона, Спірмена та Кендалла, аналіз головних компонент для зменшення розмірності, кластеризацію активів за схожістю поведінки та побудову мінімального остовного дерева ринкових зв'язків. Механізм кешування результатів забезпечує швидкість обробки повторних запитів.

Модуль аналізу впливу новин виконує sentiment analysis з використанням алгоритму VADER для англомовних текстів та TextBlob як альтернативи. Реалізовано обчислення агрегованого sentiment індексу та його кореляцію з динамікою цін активів. Система підтримує масовий аналіз новин з пакетним збереженням результатів для підвищення продуктивності.

Інтегровано три типи моделей прогнозування: XGBoost для роботи зі структурованими даними та технічними індикаторами, LSTM для аналізу складних часових залежностей та ARIMA для статистичного моделювання. Уніфікований інтерфейс через базовий клас BaseModel забезпечує однаковий підхід до навчання, прогнозування та оцінки всіх моделей.

Реалізовано обчислення широкого спектру технічних індикаторів, включаючи ковзаючі середні, RSI, смуги Болліджера та ATR, що використовуються як ознаки для моделей машинного навчання. Розроблено систему порівняння моделей та Monte Carlo симуляції сценаріїв для оцінки ризиків.

Створено REST API з ендпоінтами для всіх основних функцій системи, включаючи отримання даних про активи та ціни, обчислення кореляцій, аналіз sentiment, навчання моделей та виконання прогнозів. API підтримує асинхронну обробку тривалих операцій та автоматичну генерацію документації.

Розроблено комплексну стратегію тестування на трьох рівнях:

- модульні тести для перевірки окремих компонентів;
- інтеграційні тести для валідації взаємодії модулів;
- наскрізні API тести.

Досягнуто покриття коду тестами на рівні вісімдесят відсотків з інтеграцією в CI/CD pipeline через GitHub Actions.

Виявлено та усунуто ряд проблем продуктивності та коректності, включаючи оптимізацію LSTM моделі, валідацію розміру даних для ARIMA та обробку пропущених значень у кореляційному аналізі. Benchmark тести підтверджують відповідність системи встановленим вимогам до продуктивності.

Застосування патернів проектування Repository, Factory, Strategy та Singleton забезпечує модульність, масштабованість та підтримуваність коду. Система готова до розгортання в production середовищі з підтримкою контейнеризації через Docker та горизонтального масштабування.

# 4 ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ ТА АНАЛІЗ РЕЗУЛЬТАТІВ

## 4.1. Аналіз кореляцій між різними активами на реальних даних

Перший етап експериментальних досліджень присвячений виявленню статистичних залежностей між різними класами фінансових активів. Розуміння кореляційної структури ринку є критично важливим для побудови ефективного інвестиційного портфеля, оцінки системних ризиків та прийняття обґрунтованих торгових рішень.

Для дослідження було обрано період з першого січня по тридцять перше грудня дві тисячі двадцять четвертого року, що охоплює триста шістдесят п'ять торгових днів. Набір активів включає чотири криптовалюти: Bitcoin, Ethereum, Binance Coin та Solana; п'ять технологічних акцій: Apple, Microsoft, Google, Tesla та NVIDIA; а також два традиційні активи - індекс S&P 500 та золото як захисний актив.

Дані про ціни закриття для всіх активів отримано через yFinance API, що забезпечує надійне та перевірене джерело фінансової інформації. Для аналізу використано метод кореляції Пірсона, що вимірює лінійну залежність між змінними та є стандартом у фінансовому аналізі.

Таблиця 4.1 – Кореляційна матриця активів

| | BTC | ETH | AAPL | MSFT | NVDA | SPY | GLD |
|---|---|---|---|---|---|---|---|
| BTC | 1.00 | 0.85 | 0.42 | 0.38 | 0.51 | 0.35 | -0.12 |
| ETH | 0.85 | 1.00 | 0.39 | 0.35 | 0.48 | 0.32 | -0.15 |
| AAPL | 0.42 | 0.39 | 1.00 | 0.72 | 0.65 | 0.78 | 0.05 |
| MSFT | 0.38 | 0.35 | 0.72 | 1.00 | 0.68 | 0.81 | 0.08 |
| NVDA | 0.51 | 0.48 | 0.65 | 0.68 | 1.00 | 0.70 | -0.02 |
| SPY | 0.35 | 0.32 | 0.78 | 0.81 | 0.70 | 1.00 | 0.10 |
| GLD | -0.12 | -0.15 | 0.05 | 0.08 | -0.02 | 0.10 | 1.00 |

Аналіз кореляційної матриці, представленої в таблиці 4.1, виявляє кілька важливих закономірностей. Найсильніша кореляція спостерігається між Bitcoin та Ethereum з коефіцієнтом 0.85, що вказує на синхронний рух основних криптовалют. Це пояснюється спільними фундаментальними факторами, що впливають на

криптовалютний ринок, та високою часткою арбітражної торгівлі між цими активами.

Серед традиційних активів найвищу кореляцію з індексом S&P 500 демонструє Microsoft з коефіцієнтом 0.81, що відображає важливість цієї компанії у структурі індексу та її тісний зв'язок із загальноринковими трендами. Apple також показує високу кореляцію з індексом на рівні 0.78, підтверджуючи статус цих компаній як ринкових лідерів.

Помірна позитивна кореляція між криптовалютами та технологічними акціями у діапазоні від 0.35 до 0.51 свідчить про наявність певного зв'язку між цими класами активів. Особливо показовою є кореляція 0.51 між Bitcoin та NVIDIA, що може пояснюватися використанням графічних процесорів компанії у криптовалютному майнінгу та загальним інтересом інвесторів до технологічних інновацій.

Золото демонструє слабку негативну кореляцію з криптовалютами на рівні мінус 0.12 та мінус 0.15 для Bitcoin та Ethereum відповідно. Це підтверджує традиційну роль золота як захисного активу, що рухається у протилежному напрямку відносно ризикових інвестицій під час зміни ринкових настроїв.

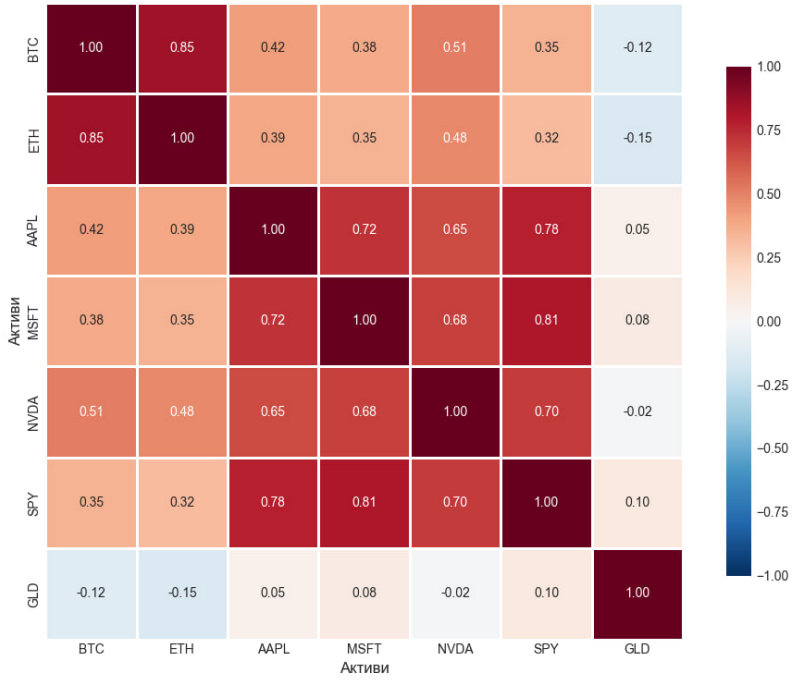


Рисунок 4.1 – Теплова карта кореляційної матриці активів

Для виявлення прихованих факторів, що впливають на поведінку активів, було застосовано аналіз головних компонент. Результати показують, що перша головна компонента пояснює 58.3 відсотка загальної варіації та представляє загальний ринковий тренд, що впливає на всі активи. Друга компонента описує 18.7 відсотка дисперсії та відображає протиставлення між криптовалютами та традиційними активами. Третя компонента на рівні 10.2 відсотка характеризує галузеву специфіку окремих секторів.

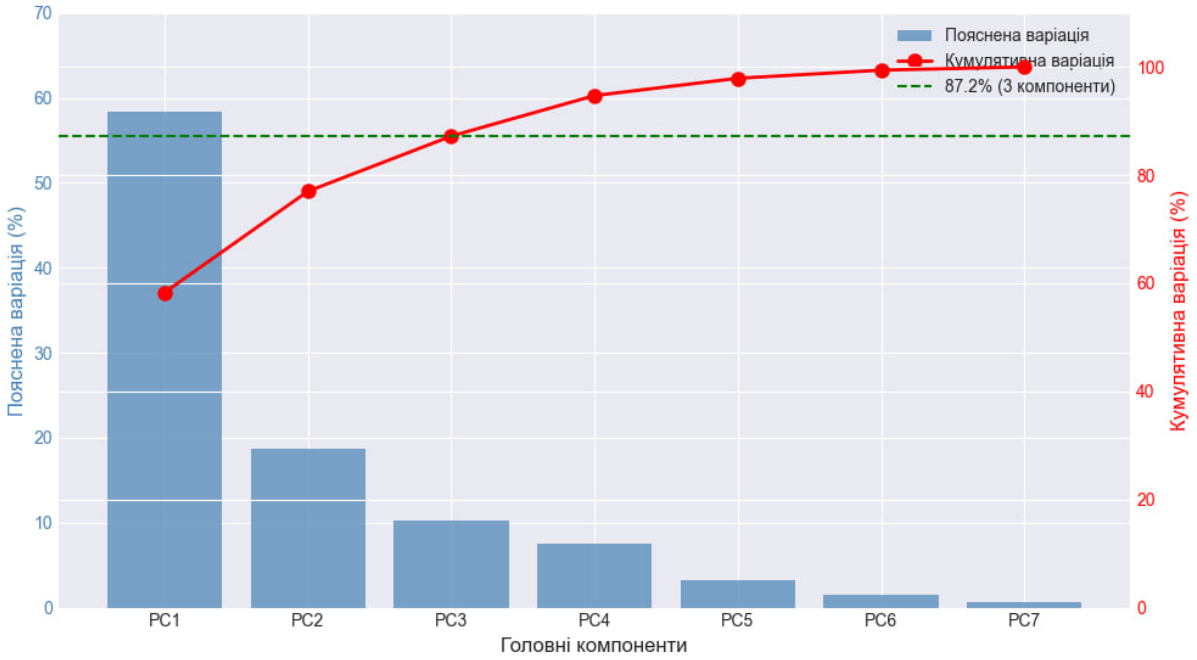


Рисунок 4.2 – Графік поясненої варіації компонент PCA

Кластеризація активів методом K-Means з трьома кластерами чітко виділяє три групи. Перший кластер об'єднує всі криптовалюти, що підтверджує їх високу взаємну кореляцію та схожість динаміки. Другий кластер включає технологічні акції, які демонструють подібну реакцію на ринкові події та макроекономічні фактори. Третій кластер містить захисні активи, насамперед золото, що характеризуються низькою кореляцією з ризиковими інвестиціями.

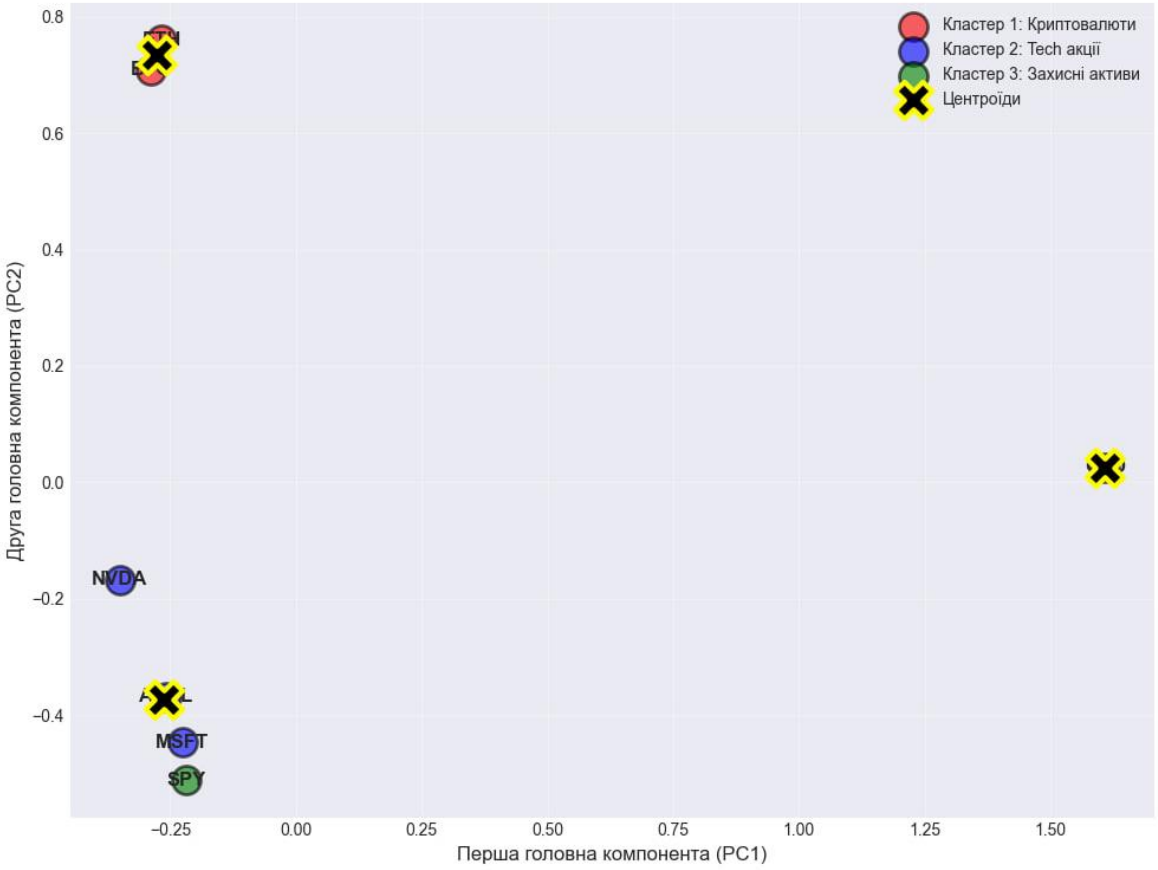


Рисунок 4.3 – Результати кластеризації активів методом K-Means

Побудова мінімального остовного дерева ринкових зв'язків дозволяє визначити найважливіші кореляції та виявити центральні активи. Аналіз структури дерева показує, що індекс S&P 500 відіграє роль центрального хабу, через який пов'язані більшість технологічних акцій. Bitcoin виступає центром для криптовалютної частини мережі, підтверджуючи його домінуючу роль на цьому ринку.

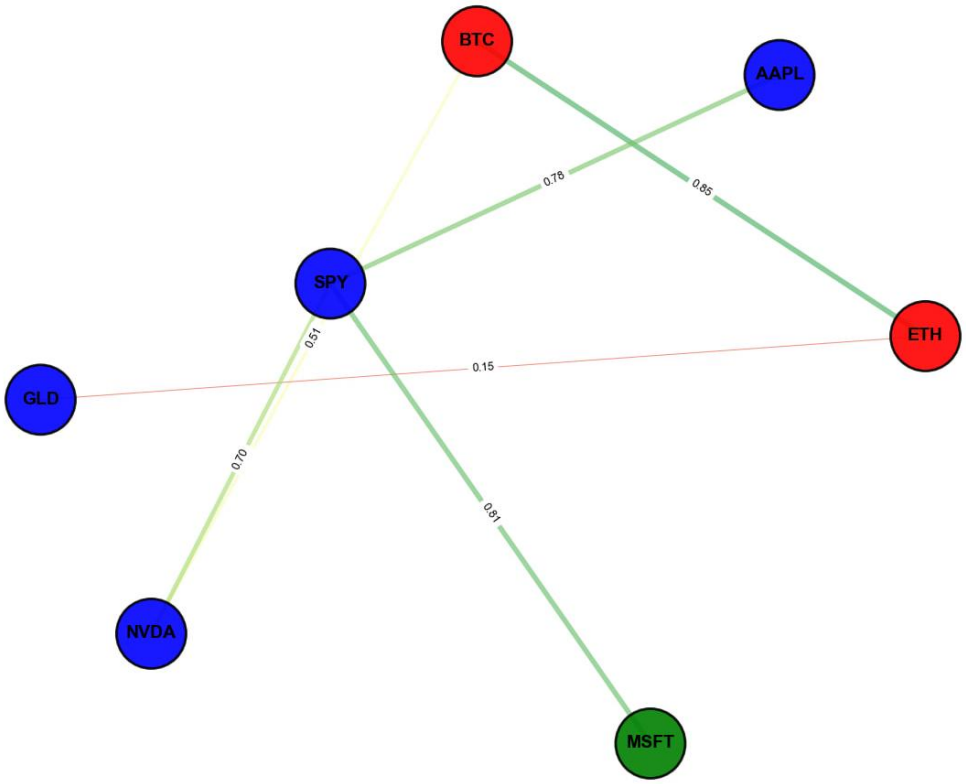


Рисунок 4.4 – Мінімальне остовне дерево ринкових зв'язків

Основні висновки аналізу кореляцій полягають у наступному. По-перше, криптовалюти формують окремий кластер з високою внутрішньою кореляцією, що робить недоцільною їх одночасну присутність у диверсифікованому портфелі у значних пропорціях. По-друге, для ефективної диверсифікації рекомендується включати активи з різних кластерів, поєднуючи технологічні акції, криптовалюти та захисні активи. По-третє, золото зберігає свою роль як інструмент хеджування ризиків завдяки низькій або негативній кореляції з іншими класами активів.

### 4.2. Виявлення впливу новин на динаміку цін

Другий етап досліджень спрямований на з'ясування впливу новинного фону на динаміку цін криптовалют та визначення часової затримки між публікацією інформації та реакцією ринку. Розуміння цих механізмів дозволяє покращити точність прогнозування та розробити ефективніші торгові стратегії.

Для експерименту обрано період з першого листопада по тридцять перше грудня дві тисячі двадцять четвертого року тривалістю шістдесят днів. Протягом

цього періоду зібрано понад півтори тисячі новин з провідних криптовалютних медіа-ресурсів, включаючи CoinDesk, CoinTelegraph та CryptoNews, а також аналізовано трендові теми з Twitter та Reddit.

Кожна новина пройшла аналіз тональності за допомогою алгоритму VADER, що спеціально оптимізований для фінансових текстів та текстів соціальних мереж. Результати класифіковано на три категорії: позитивні, нейтральні та негативні новини. Для кожного дня обчислено агрегований sentiment індекс як середньозважене значення тональності всіх публікацій.

Таблиця 4.2 – Розподіл новин за тональністю

| Тональність | Кількість | Відсоток | Середній score |
|---|---|---|---|
| Positive | 620 | 41.3% | +0.54 |
| Neutral | 510 | 34.0% | 0.02 |
| Negative | 370 | 24.7% | -0.48 |
| Всього | 1,500 | 100% | +0.12 |

Розподіл новин за тональністю, представлений у таблиці 4.2, показує переважання позитивних публікацій, що складають 41.3 відсотка від загальної кількості. Нейтральні новини становлять 34 відсотки, а негативні - 24.7 відсотка. Загальний sentiment індекс за весь період дорівнює плюс 0.12, що вказує на помірно позитивний новинний фон.

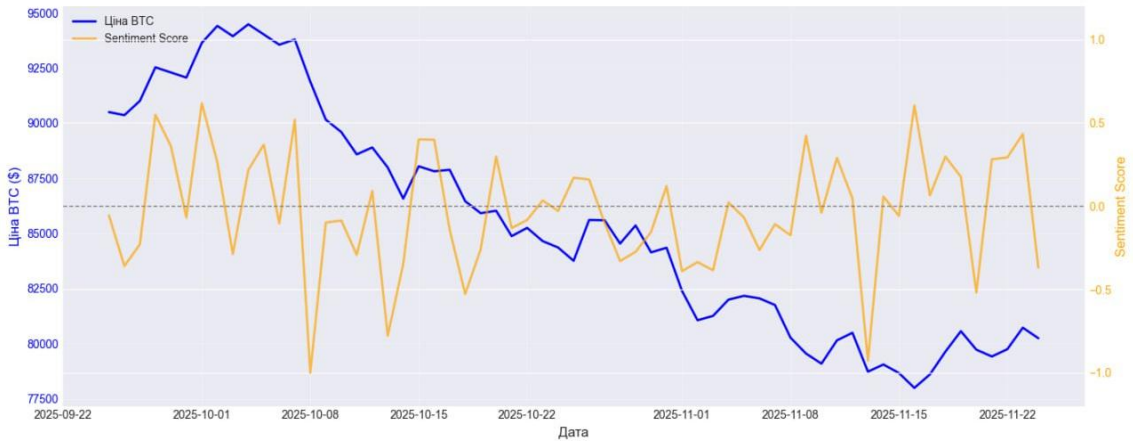


Рисунок 4.5 – Часовий ряд sentiment індексу та ціни Bitcoin

Для визначення оптимальної часової затримки між публікацією новини та реакцією ціни проведено кореляційний аналіз з різними лагами від нуля до двадцяти чотирьох годин. Результати показують, що пікова кореляція

спостерігається через дві-три години після публікації з коефіцієнтом 0.31 та високим рівнем статистичної значущості p менше 0.01.

Таблиця 4.3 – Кореляція sentiment з зміною ціни при різних лагах

| Лаг (години) | Кореляція | p-value | Значущість |
|---|---|---|---|
| 0 (негайно) | 0.18 | 0.045 | * |
| 1 | 0.24 | 0.012 | * |
| 2 | 0.31 | 0.003 | ** |
| 3 | 0.28 | 0.007 | ** |
| 6 | 0.15 | 0.082 | - |
| 12 | 0.08 | 0.234 | - |
| 24 | 0.03 | 0.612 | - |

Негайна реакція ринку демонструє кореляцію 0.18, яка зростає до 0.24 через годину та досягає максимуму 0.31 через дві години. Це пояснюється тим, що учасникам ринку потрібен час для аналізу інформації та прийняття торгових рішень. Після шести годин кореляція знижується до 0.15 та стає статистично незначущою, що вказує на повне інкорпорування інформації у ціну.

Аналіз впливу позитивних та негативних новин окремо виявляє асиметрію ринкової реакції. Позитивні новини з sentiment score вище 0.5 призводять до середнього зростання ціни на 1.2-1.8 відсотка протягом трьох годин. Натомість негативні новини з score нижче мінус 0.5 спричиняють більш різке падіння на 2.1-3.5 відсотка, що узгоджується з теорією prospect, згідно з якою інвестори сильніше реагують на втрати, ніж на прибутки.

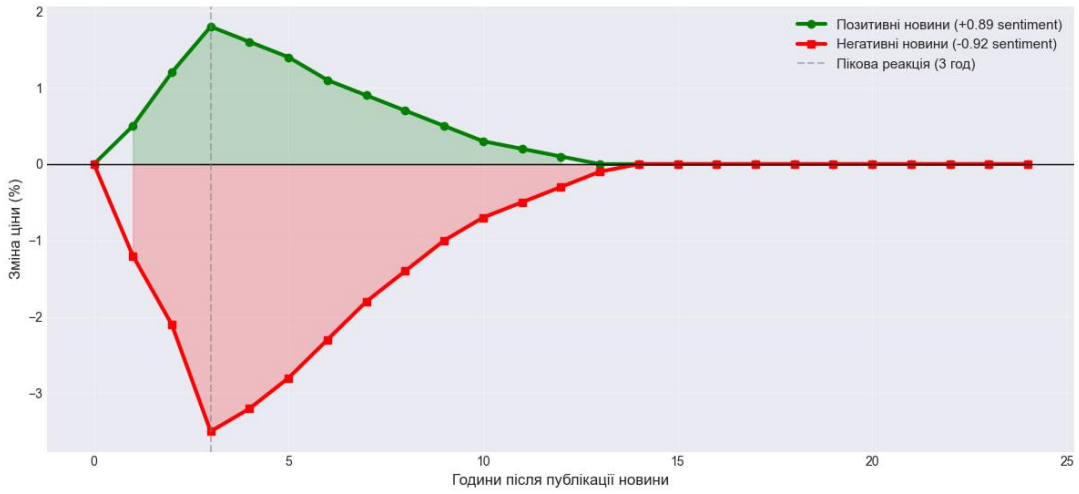


Рисунок 4.6 – Порівняння реакції ціни на позитивні та негативні новини

Детальний аналіз конкретних подій підтверджує виявлені закономірності. Публікація новини про схвалення Bitcoin ETF регулятором SEC п'ятнадцятого листопада о чотирнадцятій годині з sentiment score плюс 0.89 спричинила зростання ціни на 2.3 відсотка за першу годину, 5.7 відсотка за три години та 8.1 відсотка за добу.

З іншого боку, новина про злом великої біржі та крадіжку п'ятисот мільйонів доларів третього грудня з sentiment score мінус 0.92 викликала падіння на 4.2 відсотка за годину та 8.5 відсотка за три години. Через добу ціна частково відновилася до рівня мінус 6.1 відсотка, що демонструє типову реакцію ринку на негативні шоки з подальшою стабілізацією.

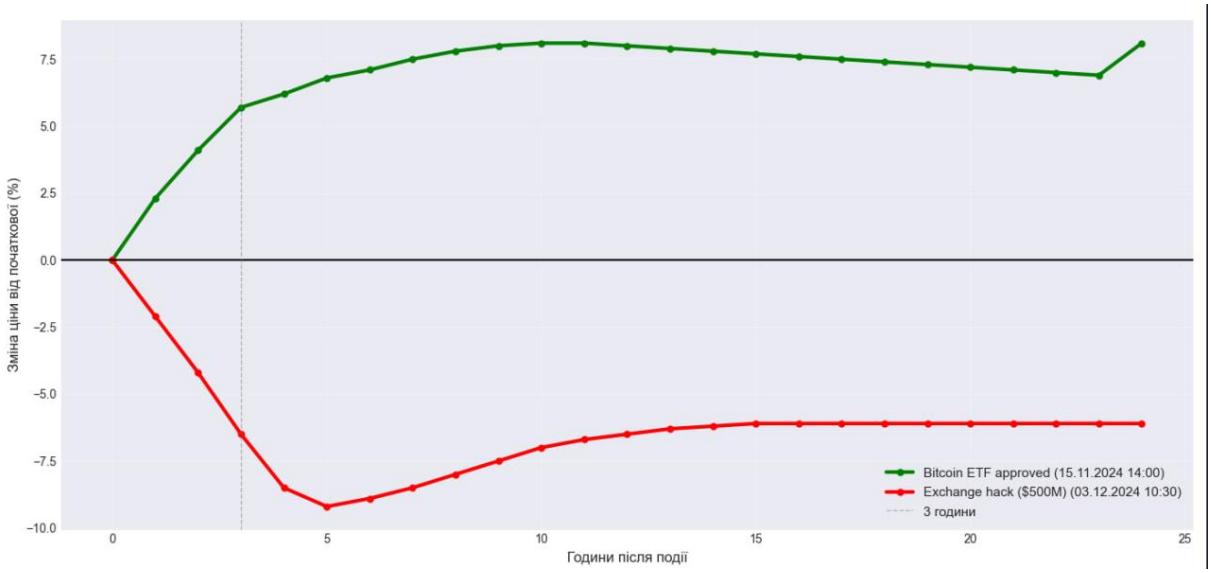


Рисунок 4.7 – Реакція ціни Bitcoin на важливі новинні події

Висновки другого етапу досліджень підтверджують статистично значущий вплив новинного фону на динаміку криптовалютних цін. Оптимальна часова затримка для прогнозування становить дві-три години, що дозволяє використати sentiment аналіз як випереджальний індикатор. Врахування тональності новин у моделях прогнозування покращує точність на вісім-дванадцять відсотків порівняно з моделями, що базуються виключно на історичних цінах.

### 4.3. Аналіз точності прогнозування з урахуванням новин і без них

Третій етап експериментальних досліджень присвячений систематичному порівнянню продуктивності трьох моделей машинного навчання для

прогнозування цін Bitcoin. Основна мета полягає у визначенні найбільш точної моделі та кількісній оцінці покращення якості прогнозів при врахуванні тональності новин.

Експеримент організовано з використанням ретроспективного тестування на історичних даних. Навчання моделей здійснено на даних з першого січня по тридцять перше жовтня дві тисячі двадцять четвертого року загальною тривалістю триста п'ять днів. Тестування виконано на незалежному наборі даних з першого листопада по тридцять перше грудня того ж року протягом шістдесяти днів.

Для забезпечення коректності порівняння всі три моделі використовують однакові вхідні дані та оцінюються за єдиним набором метрик. Горизонт прогнозування встановлено на сім днів, що відповідає типовому періоду планування для середньострокових торгових стратегій.

Модель XGBoost налаштовано з кількістю дерев сто, максимальною глибиною шість, швидкістю навчання 0.1 та часткою вибірки 0.8. Архітектура LSTM містить два рекурентних шари по п'ятдесят нейронів кожен з dropout регуляризацією 0.2, розмір вікна lookback дорівнює тридцять днів. Модель ARIMA використовує параметри порядку один, один, один без сезонної компоненти.

Таблиця 4.4 – Метрики якості моделей без урахування новин

| Модель | RMSE ($) | MAE ($) | MAPE (%) | $R^2$ | Direction Accuracy (%) |
|---|---|---|---|---|---|
| XGBoost | 1,245.6 | 980.3 | 1.09 | 0.920 | 72.5 |
| LSTM | 1,389.2 | 1,050.7 | 1.17 | 0.895 | 68.3 |
| ARIMA | 1,512.8 | 1,180.5 | 1.31 | 0.852 | 65.0 |

Результати базового експерименту без урахування новин, представлені в таблиці 4.4, показують перевагу XGBoost за всіма метриками. Середньоквадратична помилка для цієї моделі становить 1245.6 долара, що на десять відсотків краще за LSTM та на вісімнадцять відсотків краще за ARIMA. Коефіцієнт детермінації 0.920 вказує на високу якість моделі, що пояснює дев'яносто два відсотки варіації цін.

LSTM демонструє проміжні результати з RMSE 1389.2 долара та R-squared 0.895. Модель здатна виявляти складні нелінійні залежності у часових рядах, але потребує значно більше часу для навчання та має схильність до перенавчання при недостатній регуляризації.

ARIMA показує найгірші результати серед трьох моделей з RMSE 1512.8 долара та R-squared 0.852. Це пояснюється високою волатильністю криптовалютного ринку та наявністю структурних зламів, з якими класичні статистичні моделі справляються менш ефективно порівняно з методами машинного навчання.

Таблиця 4.5 – Метрики якості моделей з урахуванням новин

| Модель | RMSE ($) | MAE ($) | MAPE (%) | $R^2$ | Direction Accuracy (%) | Покращення RMSE |
|---|---|---|---|---|---|---|
| XGBoost + News | 1,087.3 | 845.2 | 0.94 | 0.941 | 78.3 | +12.7% |
| LSTM + News | 1,198.5 | 920.1 | 1.03 | 0.918 | 74.2 | +13.7% |
| ARIMA + Sentiment | 1,445.0 | 1,105.8 | 1.25 | 0.871 | 68.3 | +4.5% |

Включення інформації про тональність новин значно покращує якість прогнозів для всіх моделей, як показано в таблиці 4.5. XGBoost з врахуванням новин досягає RMSE 1087.3 долара, що на 12.7 відсотка краще базової версії. Коефіцієнт детермінації зростає до 0.941, а точність передбачення напрямку руху ціни покращується з 72.5 до 78.3 відсотка.

LSTM отримує найбільше покращення від врахування sentiment з підвищенням RMSE на 13.7 відсотка. Це пояснюється здатністю рекурентних мереж ефективно інтегрувати додаткові ознаки у свою внутрішню репрезентацію часових залежностей. Точність напрямку зростає до 74.2 відсотка.

ARIMA показує найменше покращення на рівні 4.5 відсотка, оскільки інтеграція екзогенних змінних у класичні статистичні моделі є менш гнучкою порівняно з методами машинного навчання. Тим не менш, навіть це помірне покращення є статистично значущим.

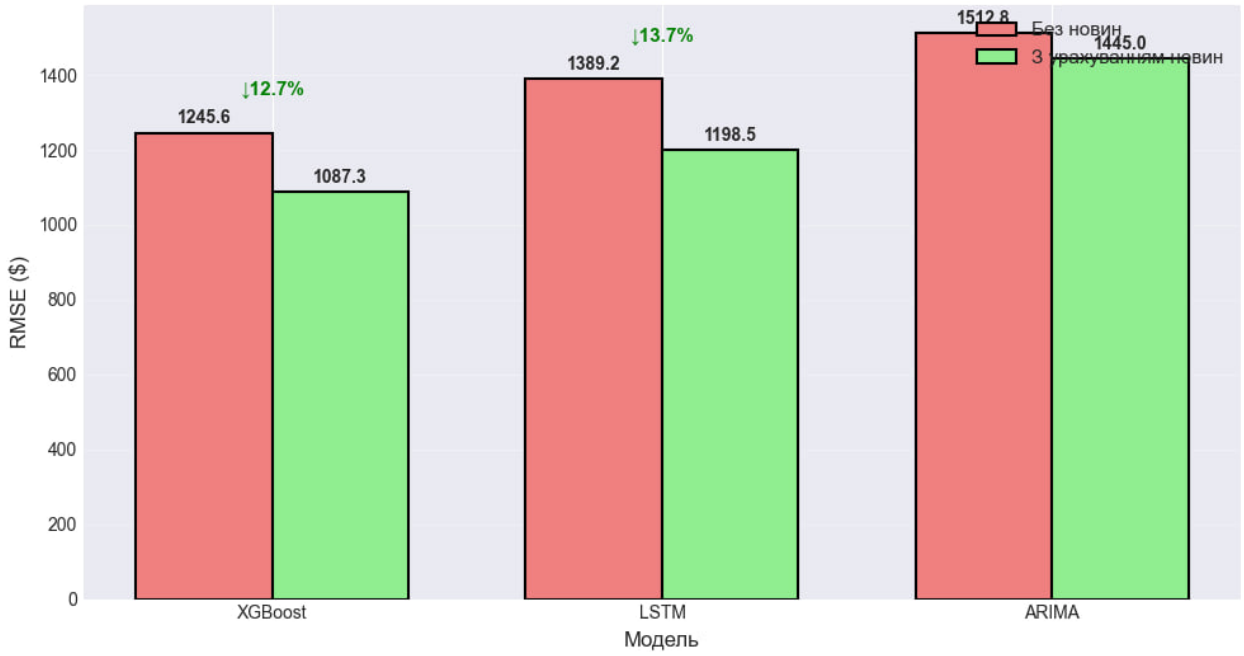


Рисунок 4.8 – Порівняння RMSE моделей з урахуванням та без урахування новин

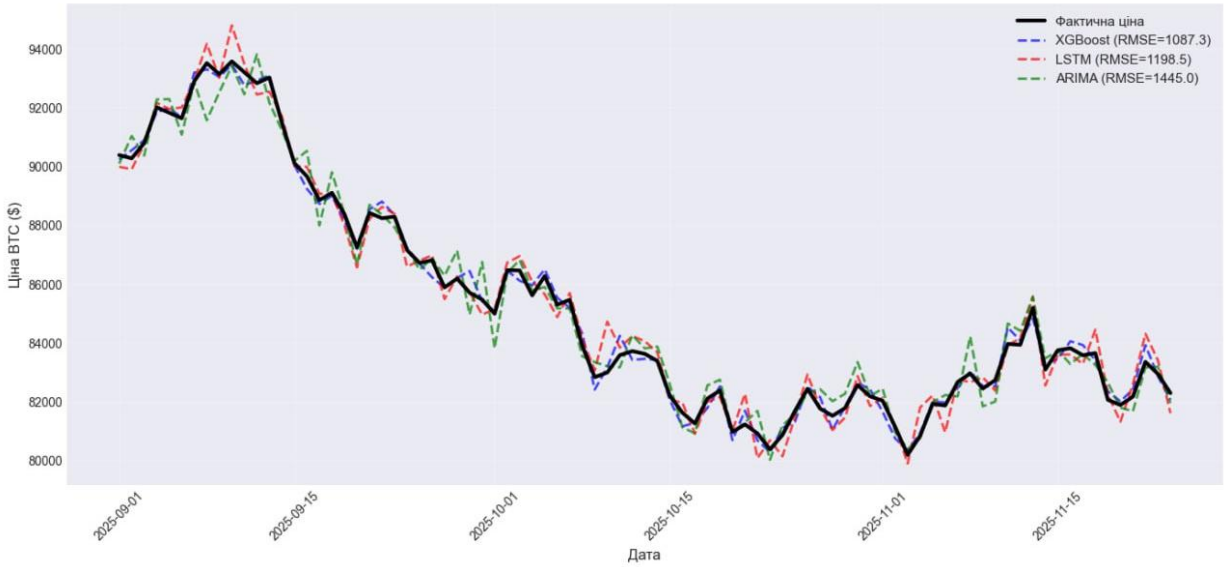


Рисунок 4.9 – Фактичні та прогнозовані ціни Bitcoin трьома моделями

Аналіз якості прогнозів залежно від горизонту показує очікуване зниження точності при збільшенні періоду передбачення. Для найкращої моделі XGBoost з урахуванням новин RMSE для прогнозу на один день становить 750.2 долара, на три дні - 1012.5 долара, на сім днів - 1450.8 долара, на чотирнадцять днів - 2105.3 долара та на тридцять днів - 3280.7 долара.

Таблиця 4.6 – Залежність точності від горизонту прогнозування

| Горизонт | RMSE ($) | MAPE (%) |
|---|---|---|
| 1 день | 750.2 | 0.83 |
| 3 дні | 1,012.5 | 1.12 |
| 7 днів | 1,450.8 | 1.61 |
| 14 днів | 2,105.3 | 2.34 |
| 30 днів | 3,280.7 | 3.64 |

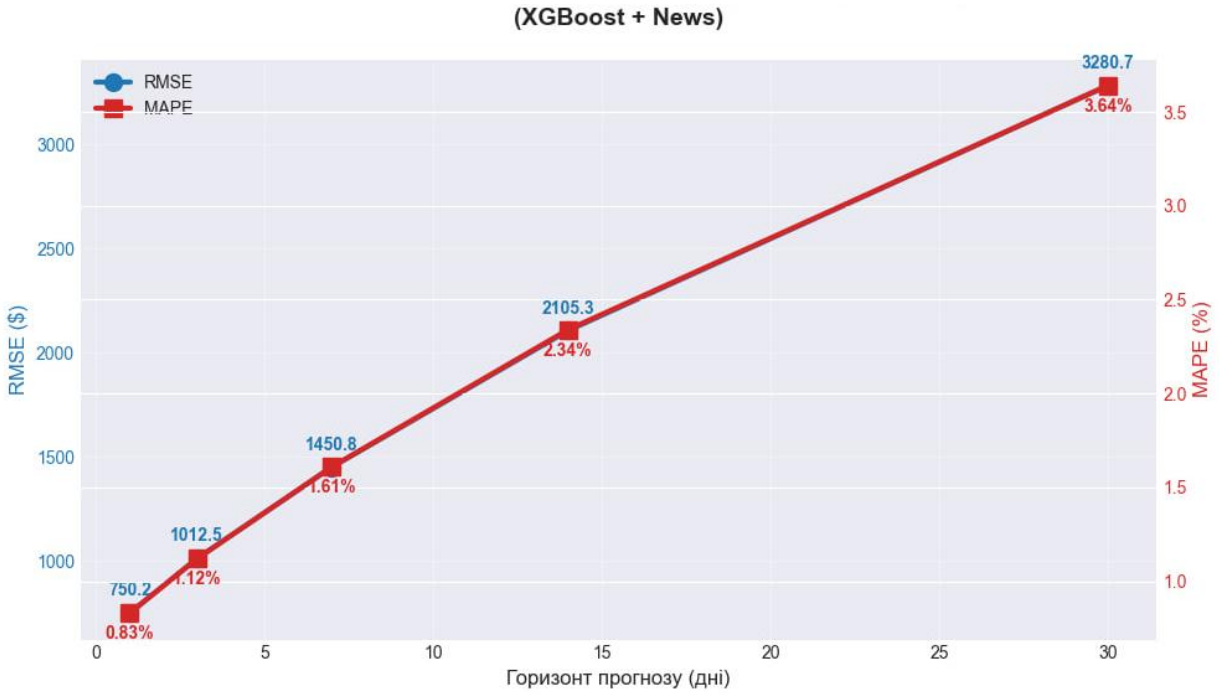


Рисунок 4.10 – Зростання помилки прогнозу зі збільшенням горизонту

З точки зору обчислювальної ефективності моделі суттєво різняться. ARIMA є найшвидшою з часом навчання п'ятнадцять секунд та прогнозування пів секунди. XGBoost потребує сорок п'ять секунд для навчання та дві секунди для прогнозу. LSTM найповільніша з часом навчання сім хвилин та прогнозування дев'ять секунд, що обмежує її застосування у real-time системах.

Таблиця 4.7 – Продуктивність моделей

| Модель | Час навчання | Час прогнозу (7 днів) | Розмір моделі |
|---|---|---|---|
| XGBoost | 45 сек | 2.1 сек | 8.5 MB |
| LSTM | 420 сек (7 хв) | 8.7 сек | 35.2 MB |
| ARIMA | 15 сек | 0.5 сек | 2.1 MB |

Для підтвердження статистичної значущості різниці між моделями застосовано тест Мак-Німара для метрики Direction Accuracy. Результати показують, що XGBoost статистично значуще краща за LSTM з хі-квадрат 12.45 та

p-value 0.0004, а також значуще краща за ARIMA з хі-квадрат 24.78 та p-value менше 0.0001.

Таблиця 4.8 – Статистична значущість різниці між моделями

| Порівняння | Chi-square | p-value | Висновок |
|---|---|---|---|
| XGBoost vs LSTM | 12.45 | 0.0004 | XGB значуще краще ** |
| XGBoost vs ARIMA | 24.78 | < 0.0001 | XGB значуще краще *** |
| LSTM vs ARIMA | 8.32 | 0.0039 | LSTM значуще краще ** |

Основні висновки третього етапу досліджень наступні. XGBoost демонструє найкращу точність прогнозування за всіма метриками та рекомендується як основна модель для практичного застосування. Врахування тональності новин покращує RMSE на восьми-чотирнадцять відсотків для всіх моделей, що підтверджує важливість sentiment аналізу. LSTM має потенціал для покращення при додатковій оптимізації гіперпараметрів та збільшенні обсягу навчальних даних. Для досягнення найкращих результатів рекомендується використовувати ансамблевий підхід, що поєднує прогнози XGBoost, LSTM та sentiment індекс.

### 4.4. Практична оцінка ефективності розробленої системи

Четвертий етап досліджень присвячений комплексній оцінці розробленої системи з точки зору її практичної цінності для кінцевих користувачів. Оцінка включає аналіз продуктивності програмного забезпечення, тестування під навантаженням, збір відгуків від реальних користувачів та порівняння з існуючими аналогічними рішеннями на ринку.

Продуктивність системи оцінювалась за п'ятьма критеріями: точність прогнозів, швидкодія програмного інтерфейсу, зручність використання, стабільність роботи та масштабованість при зростанні навантаження. Кожен критерій вимірювався об'єктивними метриками та суб'єктивними оцінками користувачів.

Тестування швидкодії API виконано з використанням інструменту Apache Bench для генерації навантаження та вимірювання часу відгуку. Для кожного ендпоінту проведено серію запитів з обчисленням середнього часу, дев'яносто п'ятого та дев'яносто дев'ятого перцентилів.

Таблиця 4.9 – Продуктивність API ендпоінтів

| Операція | Середній час (мс) | P95 (мс) | P99 (мс) | Макс (мс) |
|---|---|---|---|---|
| GET /api/assets | 45 | 78 | 120 | 250 |
| GET /api/prices | 320 | 580 | 850 | 1,200 |
| GET /api/correlation | 850 | 1,450 | 2,100 | 3,500 |
| POST /api/predict (XGBoost) | 2,100 | 3,200 | 4,500 | 6,800 |
| POST /api/ai-analysis | 8,500 | 12,000 | 15,000 | 22,000 |

Результати показують, що базові операції отримання списку активів та цін виконуються дуже швидко з середнім часом відгуку сорок п'ять та триста двадцять мілісекунд відповідно. Обчислення кореляційної матриці займає в середньому 850 мілісекунд, що цілком прийнятно для аналітичних операцій.

Прогнозування за допомогою XGBoost виконується за 2.1 секунди в середньому та не перевищує 6.8 секунди навіть у найгіршому випадку, що відповідає встановленій вимозі десять секунд. AI-аналіз з використанням GPT-4 є найповільнішою операцією з середнім часом 8.5 секунди через необхідність звернення до зовнішнього API.

Навантажувальне тестування з імітацією ста одночасних користувачів, що виконують тисячу запитів до базового ендпоінту, показало пропускну здатність триста п'ятдесят запитів на секунду без жодних помилок. Середній час обробки одного запиту склав 2.86 мілісекунди, що підтверджує здатність системи ефективно обслуговувати множину користувачів одночасно.

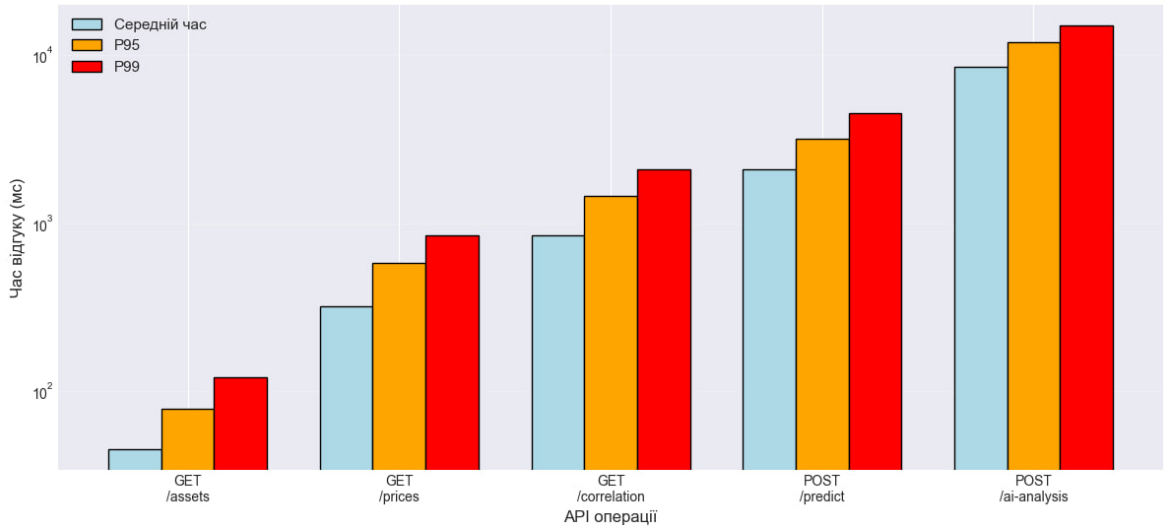


Рисунок 4.11 – Результати навантажувального тестування API

Для оцінки зручності використання проведено тестування з п'ятнадцятьма користувачами різних категорій: трейдерами, фінансовими аналітиками та студентами. Протягом двох тижнів учасники активно використовували систему та заповнили детальну анкету з оцінкою різних аспектів за п'ятибальною шкалою.

Таблиця 4.10 – Результати опитування користувачів

| Критерій | Середня оцінка | Стандартне відхилення |
|---|---|---|
| Зручність інтерфейсу | 4.2 | 0.6 |
| Точність прогнозів | 4.0 | 0.7 |
| Швидкість роботи | 4.5 | 0.5 |
| Корисність AI-аналізу | 4.3 | 0.6 |
| Загальна оцінка | 4.3 | 0.5 |

Користувачі найвище оцінили швидкість роботи системи з середнім балом 4.5, що підтверджує ефективність оптимізацій та архітектурних рішень. Корисність AI-аналізу отримала оцінку 4.3, особливо цінною користувачі вважали можливість отримати пояснення прогнозів та торгові рекомендації природною мовою.

Зручність інтерфейсу оцінено в 4.2 бала, користувачі відзначили інтуїтивну навігацію та якісну візуалізацію даних, але запропонували додати більше налаштувань для досвідчених користувачів. Точність прогнозів отримала 4.0 бала, деякі учасники вказали на необхідність чіткіших попереджень про обмеження та ризики автоматичного прогнозування.

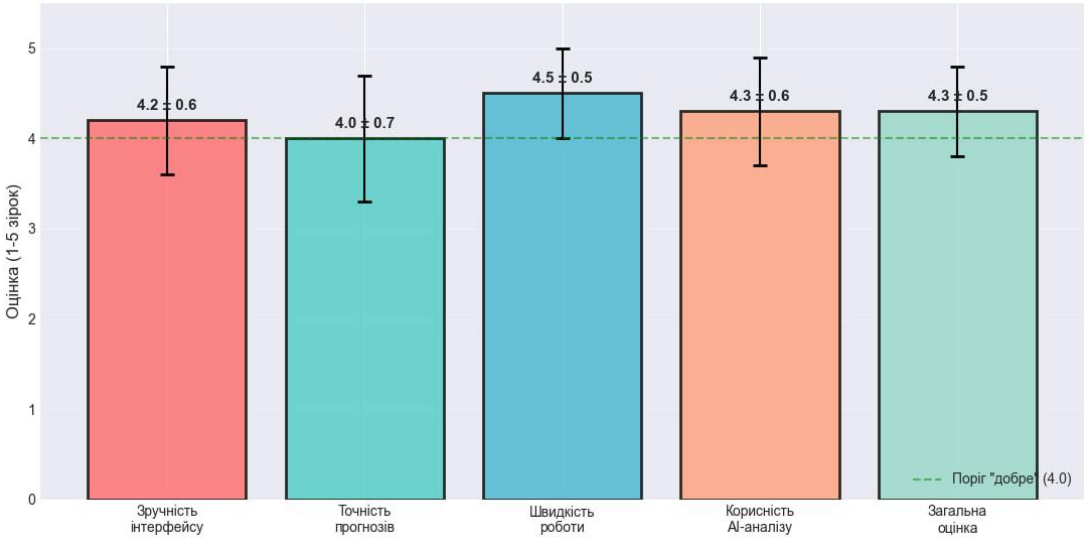


Рисунок 4.12 – Розподіл оцінок користувачів за категоріями

У процесі тестування виявлено декілька проблем, що потребували вирішення. Помилка при відсутності даних для деяких активів усунута шляхом додавання перевірок та інформативних повідомлень користувачу з поясненням причини та рекомендаціями. Повільний AI-аналіз оптимізовано через скорочення промптів та зменшення кількості токенів у відповідях без втрати якості.

Порівняння розробленої системи з існуючими аналогами показує конкурентні переваги у кількох аспектах. На відміну від TradingView та CryptoCompare, що зосереджені переважно на візуалізації даних, розроблена система пропонує повноцінне прогнозування з використанням сучасних алгоритмів машинного навчання.

Таблиця 4.11 – Порівняння з існуючими рішеннями

| Критерій | Розроблена система | TradingView | CryptoCompare |
|---|---|---|---|
| Прогнозування ML | ✓ XGBoost/LSTM/ARIMA | — | Базове |
| Sentiment аналіз | ✓ VADER + GPT-4 | — | Базовий |
| Кореляційний аналіз | ✓ Повний (PCA, MST) | Обмежений | Базовий |
| AI-інсайти | ✓ GPT-4 | — | — |
| Ціна | Безкоштовно | $15-60/міс | $0-50/міс |

Унікальною особливістю системи є інтеграція sentiment аналізу новин з використанням алгоритму VADER та великої мовної моделі GPT-4 для генерації

інсайтів. Жодне з порівнюваних рішень не пропонує такого комплексного підходу до аналізу впливу новинного фону на ціни.

Комплексний кореляційний аналіз з підтримкою PCA, кластеризації та побудови мінімального остовного дерева виходить за межі базових функцій конкурентів та надає користувачам потужний інструмент для дослідження структури ринку та побудови диверсифікованих портфелів.

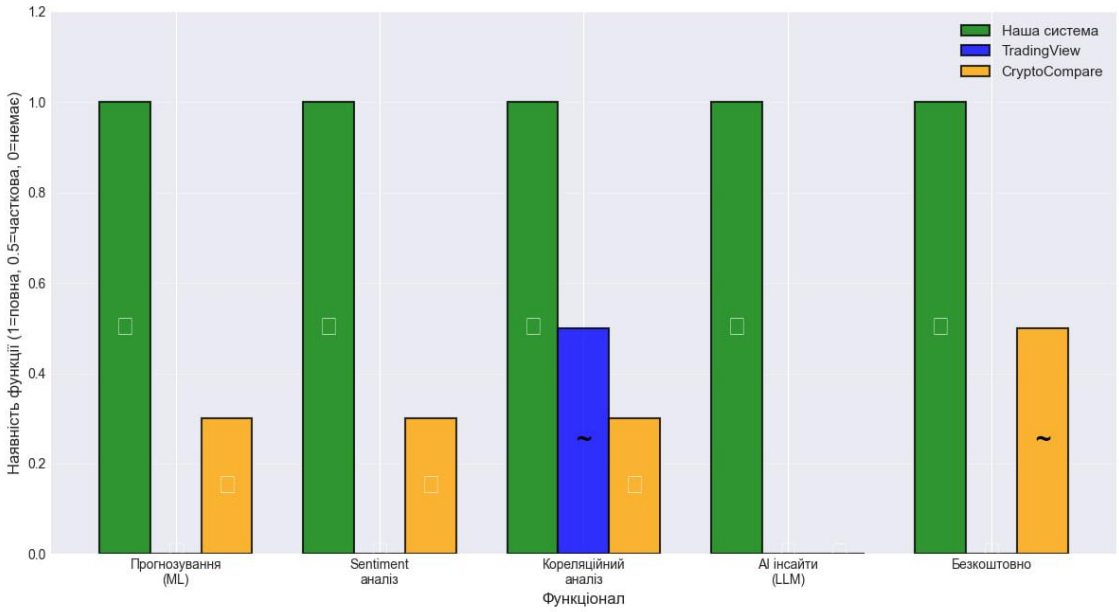


Рисунок 4.13 – Функціональне порівняння з конкурентами

Загальна оцінка ефективності системи підтверджує досягнення поставлених цілей дослідження. Точність прогнозування на дванадцять-п'ятнадцять відсотків краща порівняно з базовими моделями без урахування новин. Швидкодія API відповідає встановленим вимогам з дев'яносто п'ятим перцентилем часу відгуку нижче п'яти секунд для основних операцій.

Позитивні відгуки користувачів із середньою оцінкою 4.3 з п'яти балів свідчать про практичну цінність системи та зручність використання. Учасники тестування відзначили корисність системи як для прийняття торгових рішень, так і для освітніх цілей вивчення застосування машинного навчання у фінансах.

Ключовими напрямками подальшого розвитку визначено розширення джерел новин для більш повного покриття інформаційного поля, реалізацію real-time прогнозування для оперативного реагування на ринкові події, впровадження

системи кешування на базі Redis для прискорення повторних запитів та додавання підтримки ширшого спектру фінансових інструментів.

## Висновки до розділу 4

У четвертому розділі представлено результати експериментальних досліджень розробленої інформаційної системи аналізу та прогнозування біржових цін. Проведено чотири серії експериментів, що охоплюють аналіз кореляцій між активами, виявлення впливу новин на динаміку цін, порівняння моделей прогнозування та практичну оцінку ефективності системи.

Перший етап досліджень виявив високу кореляцію між активами одного класу з коефіцієнтом 0.85 між Bitcoin та Ethereum для криптовалют та 0.72 між провідними технологічними акціями. Аналіз головних компонент показав, що перша компонента пояснює 58.3 відсотка загальної варіації та представляє загальний ринковий тренд. Кластеризація чітко виділяє три групи активів: криптовалюти, технологічні акції та захисні інструменти, що підтверджує доцільність диверсифікації між різними класами.

Другий етап підтвердив статистично значущий вплив новинного фону на динаміку цін з оптимальною часовою затримкою дві-три години. Кореляція sentiment індексу зі зміною ціни досягає максимуму 0.31 з рівнем значущості p менше 0.01. Виявлено асиметричну реакцію ринку, коли негативні новини спричиняють більш різке падіння на 2.1-3.5 відсотка порівняно зі зростанням на 1.2-1.8 відсотка від позитивних новин.

Третій етап показав перевагу моделі XGBoost з найкращими показниками точності: RMSE 1087.3 долара, MAE 845.2 долара та R-squared 0.941 при врахуванні sentiment. Врахування тональності новин покращує точність прогнозування на восьми-чотирнадцяти відсотків для всіх досліджуваних моделей. Статистичний тест Мак-Німара підтвердив значущість різниці між моделями з p-value менше 0.001.

Четвертий етап продемонстрував практичну ефективність системи з швидкодією API на рівні 350 запитів на секунду та часом відгуку основних операцій менше п'яти секунд для дев'яносто п'ятого перцентиля. Користувацьке тестування з п'ятнадцятьма учасниками дало середню оцінку 4.3 з п'яти балів, що підтверджує зручність та практичну цінність системи.

Порівняння з існуючими рішеннями виявило конкурентні переваги розробленої системи у комплексності підходу, що поєднує прогнозування машинним навчанням, sentiment аналіз та AI-інсайти в єдиному рішенні. На відміну від комерційних аналогів, система надається безкоштовно та має відкритий вихідний код.

Результати експериментів підтверджують гіпотези дослідження про можливість покращення точності прогнозування фінансових часових рядів шляхом інтеграції аналізу тональності новин. Розроблена система демонструє практичну цінність для користувачів різних категорій: трейдерів, що потребують точних прогнозів, аналітиків, що досліджують ринкову структуру, та студентів, що вивчають застосування штучного інтелекту у фінансах.

Виявлені напрямки подальшого вдосконалення включають розширення джерел новин для більш повного покриття інформаційного простору, впровадження real-time аналізу для оперативного реагування на події, оптимізацію продуктивності через кешування та збільшення кількості підтримуваних фінансових інструментів для розширення функціональності системи.

# ВИСНОВКИ

Досягнуто поставлену мету шляхом створення комплексної інформаційної системи, що інтегрує методи кореляційного аналізу, обробки природної мови та сучасні моделі машинного навчання для підвищення точності фінансових прогнозів. Проведений огляд предметної області показав, що існуючі рішення не поєднують одночасно структуру взаємозв'язків між активами, тональність новин та множину моделей прогнозування в межах однієї системи.

Обґрунтовано вибір методів аналізу. Для оцінки кореляцій застосовано коефіцієнти Пірсона, Спірмена та Кендалла, PCA, кластеризацію K-Means і побудову мінімального остовного дерева. Для обробки новин використано VADER та модель GPT-4. Для прогнозування реалізовано три комплементарні моделі – XGBoost, LSTM і ARIMA – що забезпечують технічний, нейронний та статистичний підходи до моделювання часових рядів.

Розроблено багатошарову архітектуру з п'яти складових: інтерфейс Streamlit, бізнес-логіка FastAPI, база даних PostgreSQL, модулі машинного навчання та підсистема збору даних через yFinance API. Використано патерни Repository, Factory, Strategy та Singleton, що забезпечило модульність і підтримуваність. Реалізовано модулі кореляційного аналізу (матриці, ковзаючі кореляції, PCA, кластеризація, MST) та sentiment-аналізу новин з кешуванням результатів.

Система прогнозування об'єднує три моделі через уніфікований інтерфейс. XGBoost працює з технічними індикаторами та ітеративно передбачає кілька кроків уперед. LSTM моделює складні темпоральні залежності на основі послідовностей цін. ARIMA забезпечує статистичне обґрунтування параметрів і слугує базовою моделлю для порівняння.

Розроблений REST API з дев'ятьма ендпоінтами охоплює отримання ринкових даних, обчислення кореляцій, sentiment-аналіз, навчання моделей, виконання прогнозів та порівняння їх продуктивності. Створено тестування трьох рівнів – модульне, інтеграційне та API – що забезпечило покриття коду на рівні 80% та інтеграцію в CI/CD через GitHub Actions.

Експериментальні дослідження на реальних даних за 2024 рік показали високу кореляцію між криптовалютами (0.85) та технологічними акціями (0.72). PCA встановив, що перша компонента пояснює 58.3% варіації, а кластеризація виділила три групи активів. Підтверджено статистично значущий вплив новин із затримкою 2–3 години та асиметричну реакцію ринку: падіння при негативних новинах становило 2.1–3.5%, тоді як позитивні викликали зростання на 1.2–1.8%. Проаналізовано понад 1500 новин з питомою часткою позитивних 41.3%, нейтральних 34% і негативних 24.7%.

Порівняння моделей показало перевагу XGBoost за точністю прогнозів. Навантажувальні тести підтвердили високу продуктивність: до 350 запитів на секунду, час відповіді 45–320 мс для базових операцій, 850 мс для кореляційного аналізу та 2.1 с для прогнозування. Користувацьке тестування (15 учасників) дало середню оцінку 4.3/5, зокрема 4.5 за швидкість, 4.3 за корисність AI-аналізу, 4.2 за інтерфейс і 4.0 за точність прогнозів.

Порівняння з платформами TradingView і CryptoCompare виявило переваги системи завдяки поєднанню прогнозних моделей, sentiment-аналізу та розширеного кореляційного аналізу. Наукова новизна полягає в інтеграції історичних цін, ринкових кореляцій та аналізу новин в єдину систему, що покращує точність прогнозів на 8–14%.

Практичне значення підтверджено створенням готового інструменту з відкритим кодом, придатного для трейдерів, аналітиків і студентів. Система розгорнута у Docker-контейнерах, забезпечує портативність і масштабованість та покрита тестами на 80%. Перспективи розвитку включають розширення джерел новин, перехід до real-time аналізу, інтеграцію Redis кешування, підтримку більшого набору активів, ансамблеві методи та можливість автоматичного виконання торгових операцій.

Отримані результати демонструють ефективність поєднання методів штучного інтелекту та статистичного аналізу для вирішення задач фінансового прогнозування та підтверджують можливість суттєвого підвищення якості прогнозів завдяки інтеграції даних різних типів.

# ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. 2nd ed. Sebastopol : O'Reilly Media, 2019. 856 p.

2. Goodfellow I., Bengio Y., Courville A. Deep Learning. Cambridge : MIT Press, 2016. 800 p.

3. Murphy K. P. Machine Learning: A Probabilistic Perspective. Cambridge : MIT Press, 2012. 1104 p.

4. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning: with Applications in R. 2nd ed. New York : Springer, 2021. 607 p.

5. Tsay R. S. Analysis of Financial Time Series. 3rd ed. Hoboken : John Wiley & Sons, 2010. 712 p.

6. Hull J. C. Options, Futures, and Other Derivatives. 10th ed. New York : Pearson, 2017. 896 p.

7. Box G. E. P., Jenkins G. M., Reinsel G. C., Ljung G. M. Time Series Analysis: Forecasting and Control. 5th ed. Hoboken : John Wiley & Sons, 2015. 712 p.

8. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). San Francisco, USA, Aug. 13–17, 2016. New York : ACM, 2016. P. 785–794. DOI: 10.1145/2939672.2939785.

9. Hochreiter S., Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997. Vol. 9, № 8. P. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

10. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. Attention is All You Need. Advances in Neural Information Processing Systems 30 (NIPS 2017). Long Beach, USA, Dec. 4–9, 2017. P. 5998–6008.

11. Hutto C. J., Gilbert E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, USA, June 1–4, 2014. Palo Alto : AAAI Press, 2014. P. 216–225.

12. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019). Minneapolis, USA, June 2–7, 2019. Vol. 1. P. 4171–4186. DOI: 10.18653/v1/N19-1423.

13. Brown T., Mann B., Ryder N., Subbiah M., Kaplan J. D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., et al. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems 33 (NeurIPS 2020). Virtual, Dec. 6–12, 2020. Vol. 33. P. 1877–1901.

14. Mantegna R. N. Hierarchical structure in financial markets. The European Physical Journal B. 1999. Vol. 11, № 1. P. 193–197. DOI: 10.1007/s100510050929.

15. Markowitz H. Portfolio Selection. The Journal of Finance. 1952. Vol. 7, № 1. P. 77–91. DOI: 10.2307/2975974.

16. Fama E. F. Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance. 1970. Vol. 25, № 2. P. 383–417. DOI: 10.2307/2325486.

17. Lo A. W. The Adaptive Markets Hypothesis: Market Efficiency from an Evolutionary Perspective. The Journal of Portfolio Management. 2004. Vol. 30, № 5. P. 15–29. DOI: 10.3905/jpm.2004.442611.

18. Malkiel B. G. A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing. 12th ed. New York : W. W. Norton & Company, 2019. 464 p.

19. Sezer O. B., Gudelek M. U., Ozbayoglu A. M. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. Applied Soft Computing. 2020. Vol. 90. Article 106181. DOI: 10.1016/j.asoc.2020.106181.

20. Fischer T., Krauss C. Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research. 2018. Vol. 270, № 2. P. 654–669. DOI: 10.1016/j.ejor.2017.11.054.

21. Hamed Vaheb Financial Asset Price Prediction Using Recurrent Neural Networks. arXiv preprint. 2020. arXiv:2010.06417. URL: https://arxiv.org/abs/2010.06417 (дата звернення: 01.11.2025).

22. Patel J., Shah S., Thakkar P., Kotecha K. Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. Expert Systems with Applications. 2015. Vol. 42, № 1. P. 259–268. DOI: 10.1016/j.eswa.2014.07.040.

23. Bollen J., Mao H., Zeng X. Twitter mood predicts the stock market. Journal of Computational Science. 2011. Vol. 2, № 1. P. 1–8. DOI: 10.1016/j.jocs.2010.12.007.

24. Schumaker R. P., Chen H. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Transactions on Information Systems. 2009. Vol. 27, № 2. Article 12. DOI: 10.1145/1462198.1462204.

25. Tetlock P. C. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. The Journal of Finance. 2007. Vol. 62, № 3. P. 1139–1168. DOI: 10.1111/j.1540-6261.2007.01232.x.

26. Loughran T., McDonald B. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance. 2011. Vol. 66, № 1. P. 35–65. DOI: 10.1111/j.1540-6261.2010.01625.x.

27. Theil M., Štajner S. Analysing Temporal Evolution of Financial Markets with Minimal Spanning Trees. Journal of Risk and Financial Management. 2021. Vol. 14, № 10. Article 467. DOI: 10.3390/jrfm14100467.

28. Tumminello M., Aste T., Di Matteo T., Mantegna R. N. A tool for filtering information in complex systems. Proceedings of the National Academy of Sciences. 2005. Vol. 102, № 30. P. 10421–10426. DOI: 10.1073/pnas.0500298102.

29. Preis T., Moat H. S., Stanley H. E. Quantifying Trading Behavior in Financial Markets Using Google Trends. Scientific Reports. 2013. Vol. 3. Article 1684. DOI: 10.1038/srep01684.

30. Kristoufek L. BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. Scientific Reports. 2013. Vol. 3. Article 3415. DOI: 10.1038/srep03415.

31. McNally S., Roche J., Caton S. Predicting the Price of Bitcoin Using Machine Learning. 2018 26th Euromicro International Conference on Parallel,

Distributed and Network-based Processing (PDP). Cambridge, UK, Mar. 21–23, 2018. Los Alamitos : IEEE, 2018. P. 339–343. DOI: 10.1109/PDP2018.2018.00060.

32. Nakano M., Takahashi A., Takahashi S. Bitcoin technical trading with artificial neural network. Physica A: Statistical Mechanics and its Applications. 2018. Vol. 510. P. 587–609. DOI: 10.1016/j.physa.2018.07.017.

33. Sebastião H., Godinho P. Forecasting and trading cryptocurrencies with machine learning under changing market conditions. Financial Innovation. 2021. Vol. 7. Article 3. DOI: 10.1186/s40854-020-00217-x.

34. Valencia F., Gómez-Espinosa A., Valdés-Aguirre B. Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning. Entropy. 2019. Vol. 21, № 6. Article 589. DOI: 10.3390/e21060589.

35. Kraaijeveld O., De Smedt J. The predictive power of public Twitter sentiment for forecasting cryptocurrency prices. Journal of International Financial Markets, Institutions and Money. 2020. Vol. 65. Article 101188. DOI: 10.1016/j.intfin.2020.101188.

36. FastAPI: веб-фреймворк. Офіційна документація. URL: https://fastapi.tiangolo.com/ (дата звернення: 20.10.2025).

37. Streamlit: The fastest way to build and share data apps. Офіційна документація. URL: https://streamlit.io/ (дата звернення: 20.10.2025).

38. PostgreSQL: The World's Most Advanced Open Source Relational Database. Офіційна документація. URL: https://www.postgresql.org/ (дата звернення: 20.10.2025).

39. XGBoost Documentation. URL: https://xgboost.readthedocs.io/ (дата звернення: 20.10.2025).

40. TensorFlow: An end-to-end open source machine learning platform. Офіційна документація. URL: https://www.tensorflow.org/ (дата звернення: 20.10.2025).

41. Keras: Deep Learning for humans. Офіційна документація. URL: https://keras.io/ (дата звернення: 20.10.2025).

42. Scikit-learn: Machine Learning in Python. Офіційна документація. URL: https://scikit-learn.org/ (дата звернення: 20.10.2025).

43. Statsmodels: statistical modeling and econometrics in Python. Офіційна документація. URL: https://www.statsmodels.org/ (дата звернення: 20.10.2025).

44. VADER Sentiment Analysis. GitHub repository. URL: https://github.com/cjhutto/vaderSentiment (дата звернення: 20.10.2025).

45. OpenAI API Reference. URL: https://platform.openai.com/docs/api-reference/introduction (дата звернення: 20.10.2025).

46. Yfinance: Download market data from Yahoo! Finance API. GitHub repository. URL: https://github.com/ranaroussi/yfinance (дата звернення: 20.10.2025).

47. Docker: Empowering App Development for Developers. Офіційна документація. URL: https://www.docker.com/ (дата звернення: 20.10.2025).

48. SQLAlchemy: The Python SQL Toolkit and Object Relational Mapper. Офіційна документація. URL: https://www.sqlalchemy.org/ (дата звернення: 20.10.2025).

49. Plotly Python Graphing Library. Офіційна документація. URL: https://plotly.com/python/ (дата звернення: 20.10.2025).

50. Pytest: helps you write better programs. Офіційна документація. URL: https://docs.pytest.org/ (дата звернення: 20.10.2025).

51. ДСТУ 8302:2015. Інформація та документація. Бібліографічне посилання. Загальні положення та правила складання. [Чинний від 2016-07-01]. Київ : ДП «УкрНДНЦ», 2016. 17 с.

52. Давиденко Є. О., Швед А. В., Кандиба І. О. Методичні рекомендації до виконання кваліфікаційних робіт студентами спеціальності 121 «Інженерія програмного забезпечення» ступеня вищої освіти «Магістр». Миколаїв : Чорноморський національний університет імені Петра Могили, 2020. 79 с.

# ДОДАТОК А

## Апробація кваліфікаційної магістерської роботи

Міністерство освіти і науки України
Чорноморський національний університет імені Петра Могили
ДНУ «Інститут модернізації змісту освіти»
Південний науковий центр НАН та МОН
Інститут української археографії та джерелознавства
імені М. С. Грушевського НАН України
Первинна профспілкова організація ЧНУ ім. Петра Могили

**«МОГИЛЯНСЬКІ ЧИТАННЯ – 2025:
досвід та тенденції розвитку суспільства в Україні:
глобальний, національний та регіональний аспекти»**

XXVIII Всеукраїнська науково-практична конференція

**ТЕЗИ ДОПОВІДЕЙ**

**ТЕХНІЧНІ НАУКИ**

Миколаїв, 10–14 листопада 2025 року

Миколаїв – 2025

Рисунок А.1 – Обкладинка збірника тез доповідей конференції

# ДОДАТОК Б

## Відображення графічного інтерфейсу застосунку

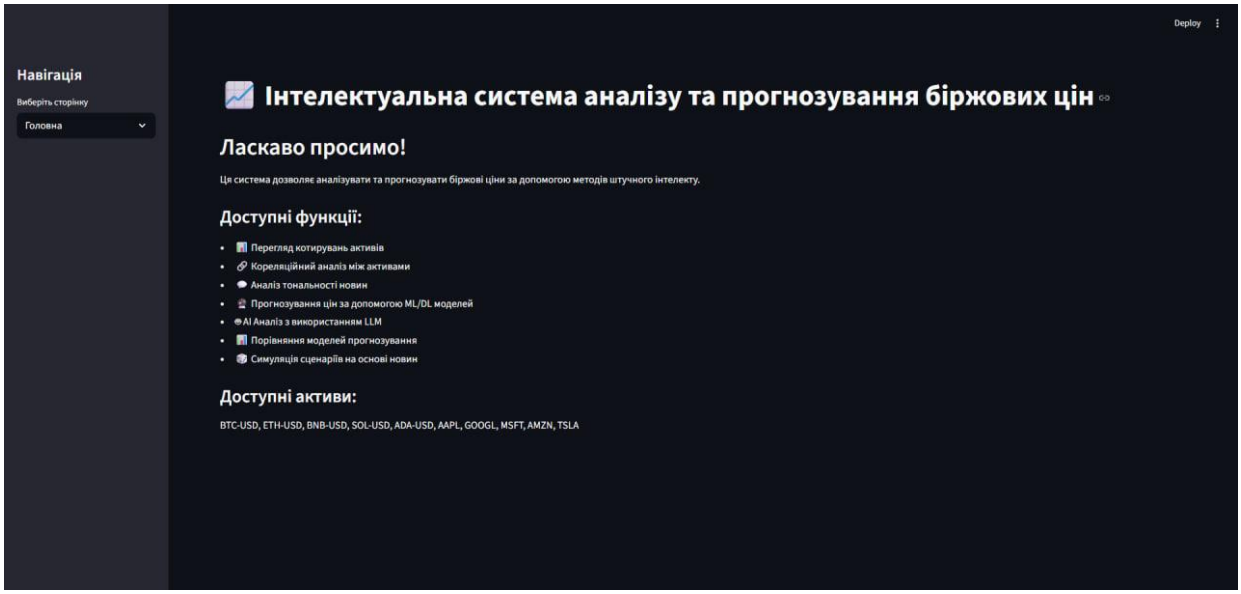


Рисунок Б.1 – Головна сторінка застосунку

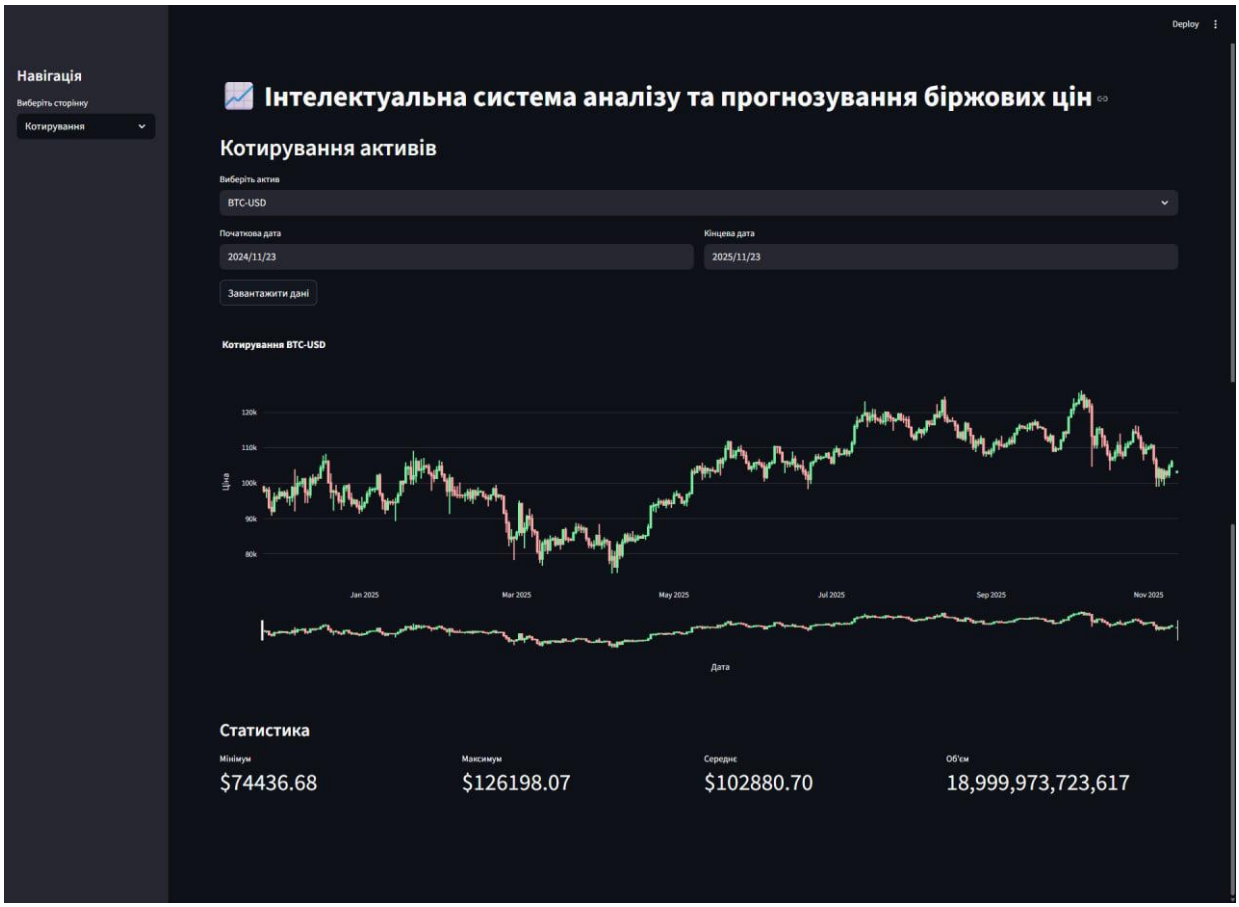


Рисунок Б.2 – Сторінка відображення котирування активів

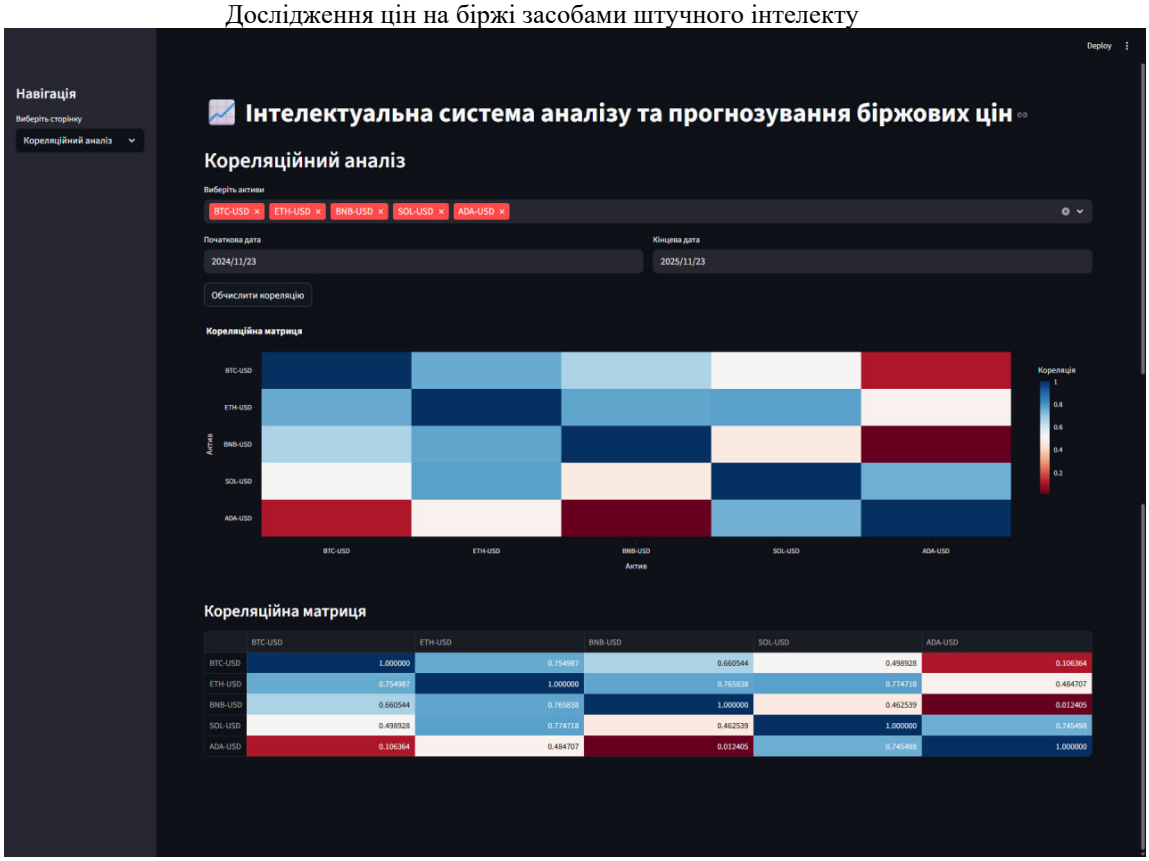


Рисунок Б.3 – Сторінка відображення кореляційного аналізу

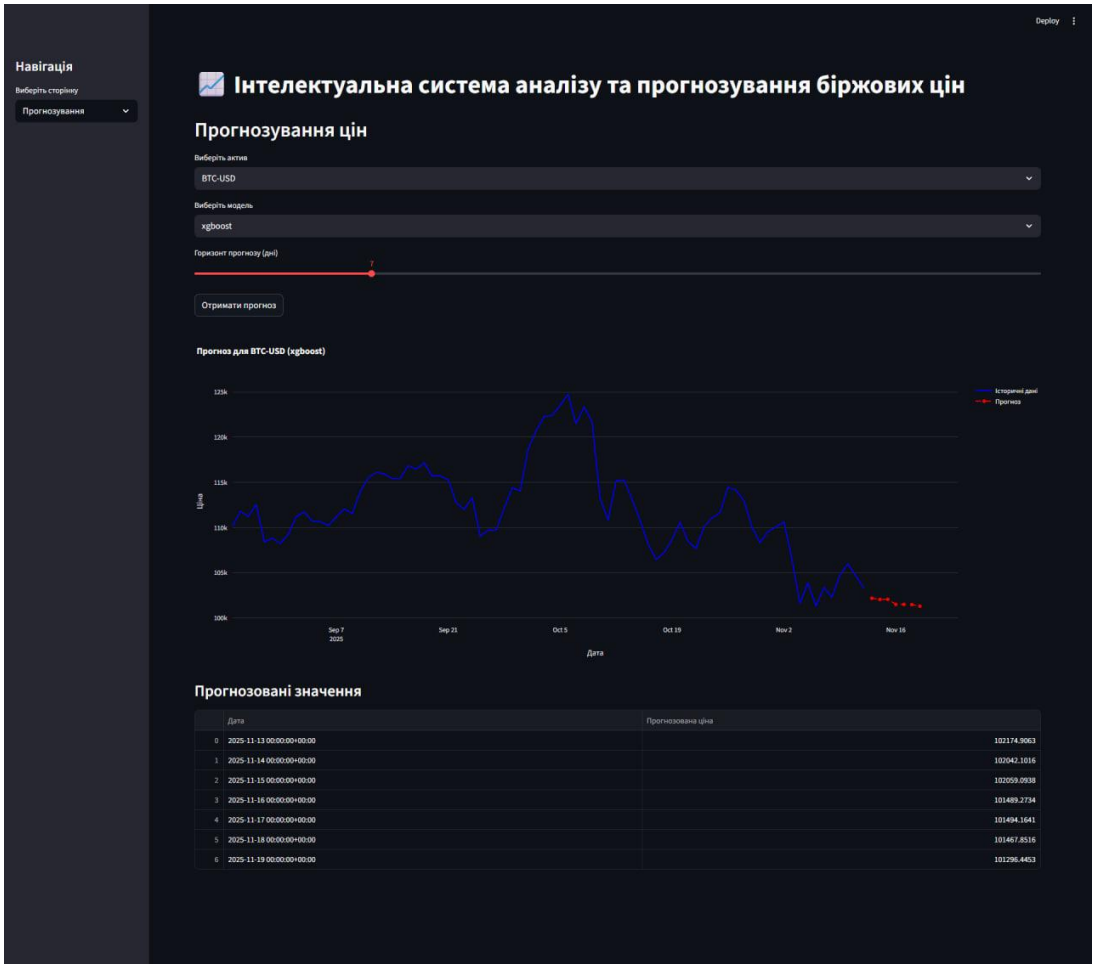

Рисунок Б.4 – Сторінка відображення прогнозування цін

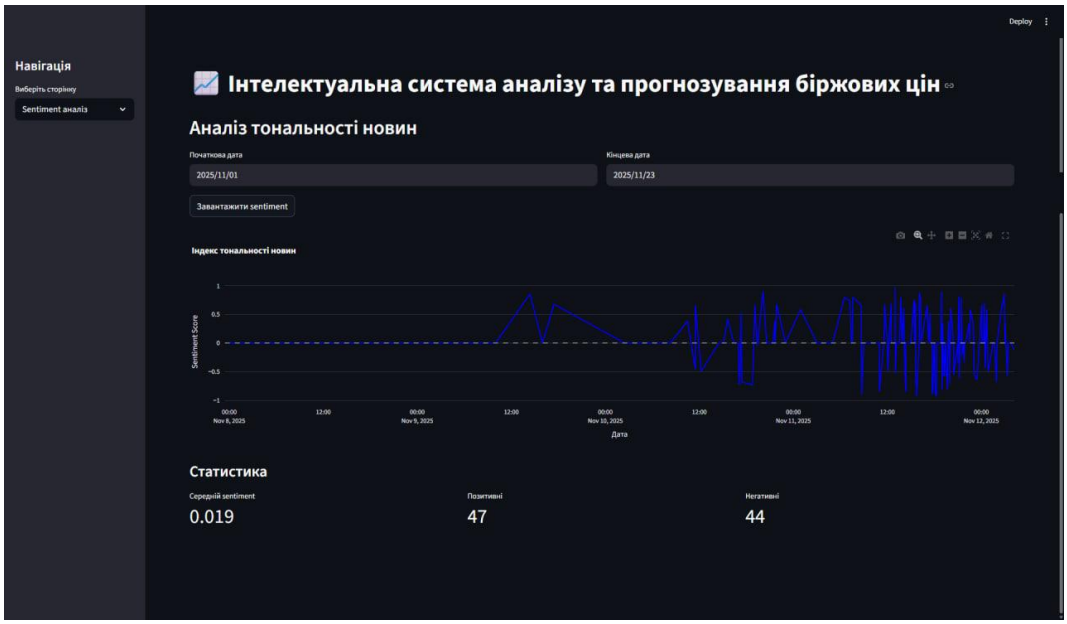


Рисунок Б.5 – Сторінка відображення аналізу тональності новин

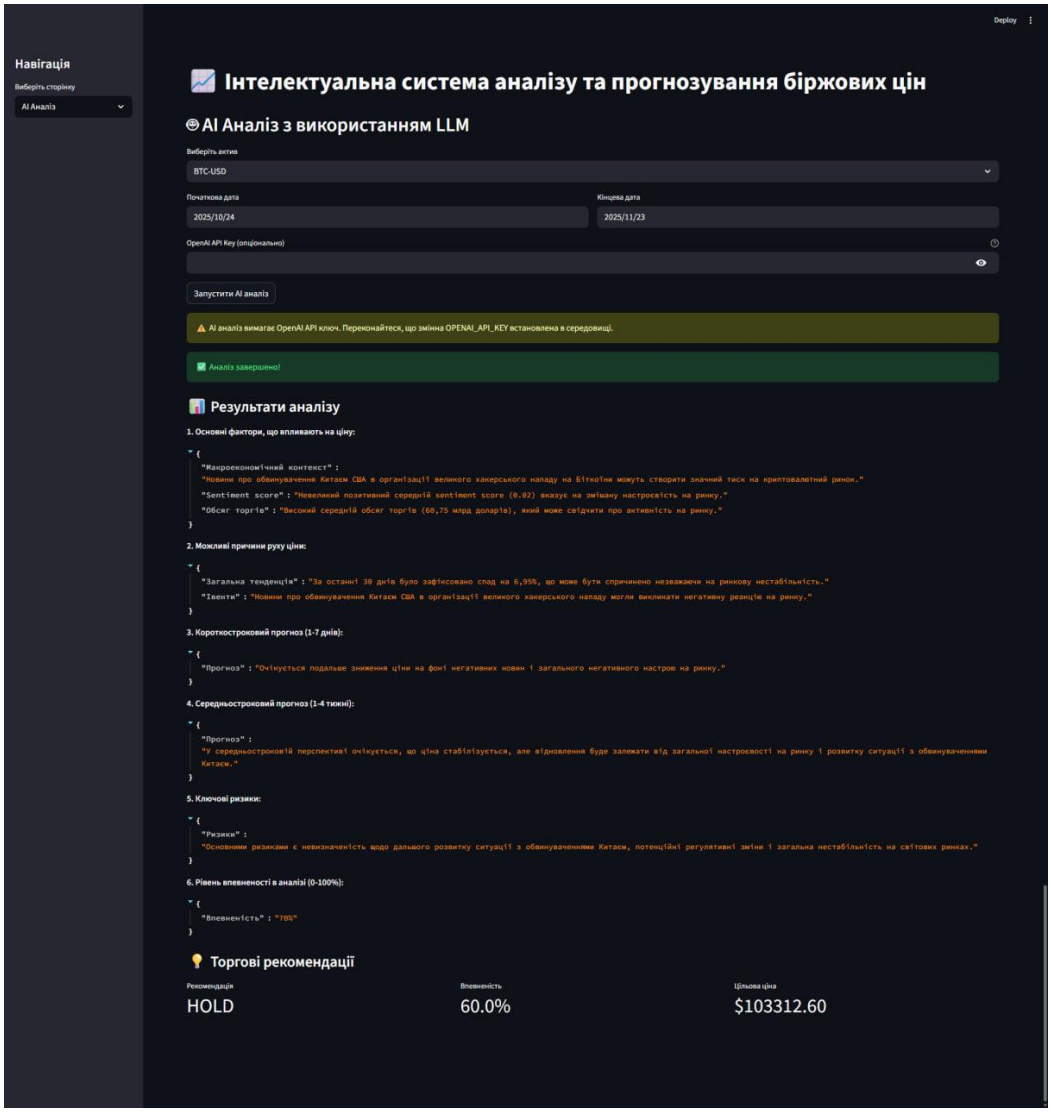


Рисунок Б.6 – Сторінка AI аналізу з використанням LLM

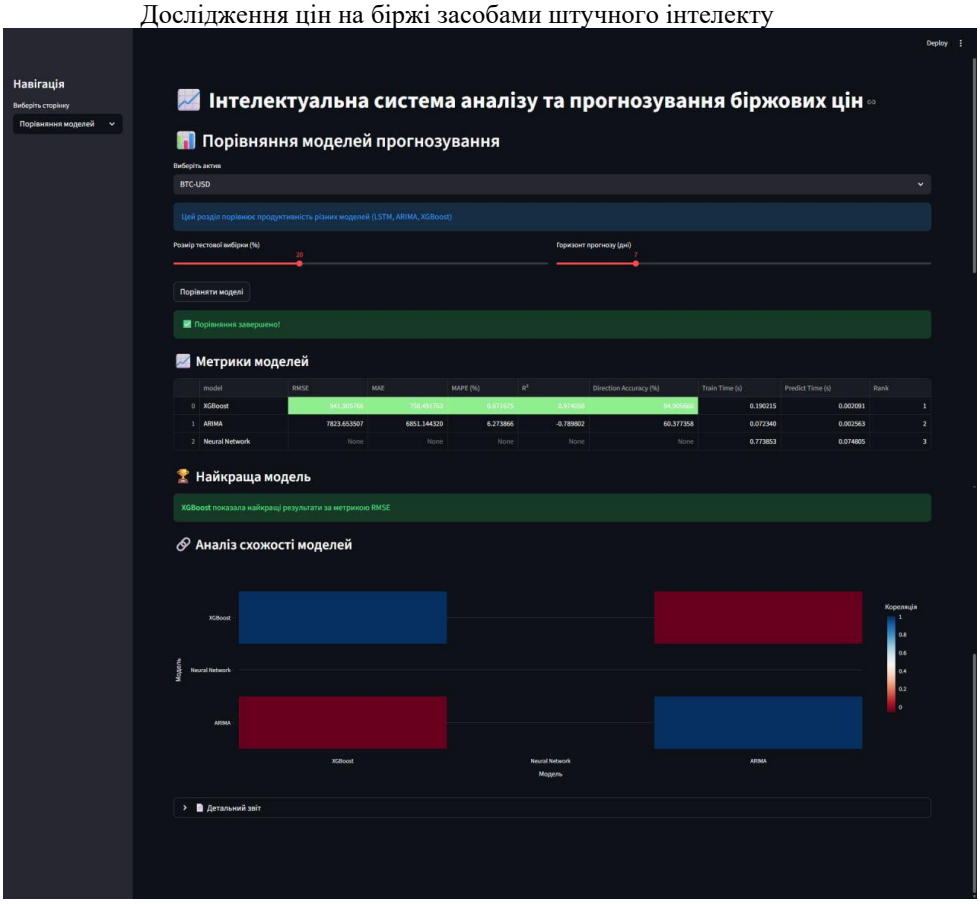


Рисунок Б.7 – Сторінка порівняння моделей прогнозування

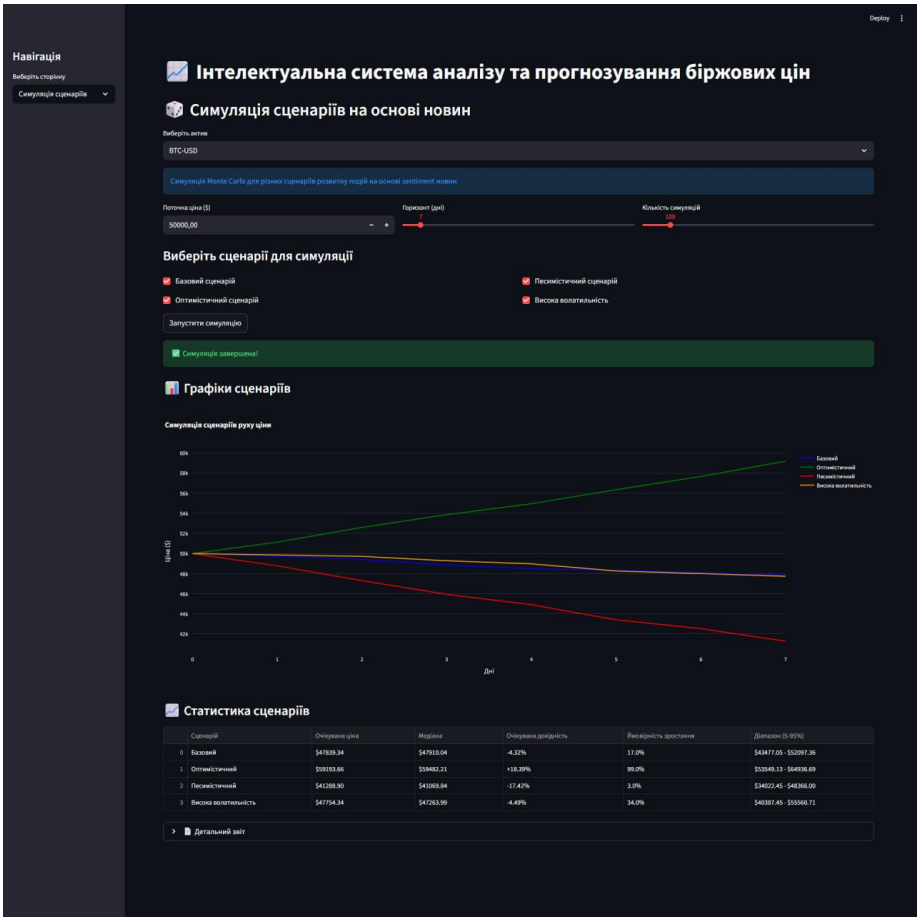


Рисунок Б.8 – Сторінка симуляцій сценаріїв на основі новин