ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ
ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
КАФЕДРА ІНФОРМАЦІЙНИХ СИСТЕМ ТА ТЕХНОЛОГІЙ

КВАЛІФІКАЦІЙНА РОБОТА

на тему: «Мультиканальний сентимент-аналіз та його застосування у прогнозуванні ринкових тенденцій криптовалют»

на здобуття освітнього ступеня магістра
зі спеціальності 124 Системний аналіз
освітньо-професійної програми «Інтелектуальні системи управління»

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

_____ Дмитро ПРОКОПЕНКО
(підпис)

Виконав: здобувач вищої освіти групи САДМ-61

Дмитро ПРОКОПЕНКО

Керівник: Михайло КУЗЬМІЧ
Доктор філософії (PhD)

Рецензент: _____
науковий ступінь, вчене звання *Ім'я, ПРІЗВИЩЕ*

Київ 2024

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

Навчально-науковий інститут інформаційних технологій

Кафедра інформаційних систем та мереж

Ступінь вищої освіти магістр

Спеціальність 124 Системний аналіз

Освітньо-професійна програма Інтелектуальні системи управління

ЗАТВЕРДЖУЮ
Завідувач кафедрою ІСТ
_____Каміла СТОРЧАК
«_____» _____20__р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Прокопенко Дмитро Олександрович
(прізвище, ім'я, по батькові здобувача)

1. Тема кваліфікаційної роботи: Мультиканальний сентимент-аналіз та його застосування у прогнозуванні ринкових тенденцій криптовалют

керівник кваліфікаційної роботи *Михайло КУЗЬМІЧ, доктор філософії (PhD)*,
(Ім'я, ПРІЗВИЩЕ, науковий ступінь, вчене звання)

затверджені наказом Державного університету інформаційно-комунікаційних технологій

від «____» _____20___р. № ____

2. Строк подання кваліфікаційної роботи «____» _____20___р.

3. Вихідні дані до кваліфікаційної роботи: наукова література з питань мультиканального сентимент-аналізу та методів машинного навчання для прогнозування ринкових тенденцій криптовалют; науково-технічна література з питань інтеграції даних із соціальних мереж, форумів і новинних платформ; аналітичні матеріали та статистичні дані щодо динаміки криптовалютних ринків; методичні матеріали та результати власних експериментальних досліджень для валідації запропонованих методів.

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)
   1. теоретико-методологічні основи мультиканального сентимент-аналізу в контексті системного аналізу
   2. загальна характеристика сучасних підходів до прогнозування ринкових тенденцій криптовалют
   3. розробка інтегрованої системи мультиканального сентимент-аналізу з використанням методів машинного навчання та оцінка її ефективності на реальних ринкових даних

5. Перелік ілюстративного матеріалу: *презентація*
6. Дата видачі завдання «_____» _____20____р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів роботи	Примітка
1	Аналіз науково-технічної літератури з питань мультиканального сентимент-аналізу та прогнозування ринкових тенденцій криптовалют		
2	Аналіз та відбір релевантних джерел даних з урахуванням специфіки високоволатильних ринків		
3	Дослідження методів машинного навчання та підходів до обробки великих даних для настроєвого аналізу		
4	Розробка концепції інтегрованої системи мультиканального збору й попередньої обробки інформації		
5	Проєктування й написання коду для аналізу настроїв		
6	Тестування та налаштування запропонованої системи на реальних ринкових даних із оцінкою ефективності прогнозів		
7	Оформлення кваліфікаційної роботи: підготовка вступу, висновків, реферату та розробка презентаційних матеріалів		
8	Завершальна перевірка, узгодження та підготовка до публічного захисту		

Здобувач вищої освіти _____ Дмитро ПРОКОПЕНКО
(підпис) *(Ім'я, ПРІЗВИЩЕ)*

Керівник
кваліфікаційної роботи _____ Михайло КУЗЬМІЧ
(підпис) *(Ім'я, ПРІЗВИЩЕ)*

РЕФЕРАТ

Текстова частина кваліфікаційної роботи на здобуття освітнього ступеня магістра: 73 стор., 7 рис., 4 табл., 20 джерел.

Мета роботи – розробка ефективного підходу до мультиканального сентимент-аналізу для прогнозування ринкових тенденцій криптовалют.

Об’єкт дослідження – процес передбачення ринкових тенденцій у криптовалютній індустрії.

Предмет дослідження – мультиканальний сентимент-аналіз, який дає змогу об’єднати різнорідні дані з соціальних мереж, форумів та новинних платформ для оцінювання впливу настроїв на динаміку цін криптовалют.

Короткий зміст роботи:

У роботі обґрунтовано актуальність застосування мультиканального сентимент-аналізу для високоволатильних ринків, зокрема криптовалют. На основі огляду наукової літератури визначено ключові підходи до аналізу текстових даних і великих обсягів інформації з різних джерел. Розглянуто методи машинного навчання для автоматичної оцінки настроїв та їх інтеграції у процес прогнозування ринкових тенденцій. Запропоновано алгоритм обробки даних, що включає збір, фільтрацію, нормалізацію й моделювання настроїв, а також побудову системи для короткострокового прогнозування цін криптовалют. Проведено серію експериментів на реальних ринкових даних, результати яких засвідчили позитивний вплив мультиканального підходу на точність прогнозування та оперативність ухвалення рішень. Робота містить теоретичне обґрунтування методів, схеми інтеграції підсистем аналізу, а також дає практичні рекомендації щодо впровадження системи у діяльність трейдерів, фінансових аналітиків та регуляторних органів.

КЛЮЧОВІ СЛОВА: МУЛЬТИКАНАЛЬНИЙ СЕНТИМЕНТ-АНАЛІЗ, КРИПТОВАЛЮТИ, МАШИННЕ НАВЧАННЯ, СОЦІАЛЬНІ МЕРЕЖІ, ВИСОКОВОЛАТИЛЬНИЙ РИНОК, BIG DATA, ПРОГНОЗУВАННЯ.

ABSTRACT

The text part of the qualification work for the master's degree: 73 pages, 7 figures, 4 tables, 20 sources.

Purpose of the study – to develop an effective approach to multi-channel sentiment analysis for forecasting cryptocurrency market trends.

Object of the study – the process of predicting market trends in the cryptocurrency industry.

Subject of the study – multi-channel sentiment analysis, which enables the integration of heterogeneous data from social networks, forums, and news platforms to assess the impact of sentiment on cryptocurrency price dynamics.

Brief summary of the work:

This study substantiates the relevance of applying multi-channel sentiment analysis to highly volatile markets, particularly cryptocurrencies. Based on a review of the scientific literature, key approaches to analyzing textual data and large volumes of information from various sources were identified. Machine learning methods were examined for the automated assessment of sentiment and its integration into the market trend forecasting process. An algorithm for data processing was proposed, encompassing collection, filtering, normalization, and sentiment modeling, as well as the development of a system short-term forecasting of cryptocurrency prices. A series of experiments on real market data demonstrated the positive impact of the multi-channel approach on forecasting accuracy and decision-making efficiency. The work includes theoretical justification of the methods, outlines the integration schemes for analysis subsystems, and offers practical recommendations for implementing the system in the activities of traders, financial analysts, and regulatory bodies.

KEYWORDS: MULTI-CHANNEL SENTIMENT ANALYSIS, CRYPTOCURRENCIES, MACHINE LEARNING, SOCIAL NETWORKS, HIGH-VOLATILITY MARKET, BIG DATA, FORECASTING.

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

Навчально-науковий інститут інформаційних технологій

ПОДАННЯ
ГОЛОВІ ЕКЗАМЕНАЦІЙНОЇ КОМІСІЇ
ЩОДО ЗАХИСТУ КВАЛІФІКАЦІЙНОЇ РОБОТИ
на здобуття освітнього ступеня магістра

Направляється здобувач *Прокопенко Д. О.* до захисту кваліфікаційної роботи
(прізвище та ініціали)
за спеціальністю *124 Системний аналіз*
(код, найменування спеціальності)
освітньо-професійної програми Інтелектуальні системи управління
(назва)
на тему: «Мультиканальний сентимент-аналіз та його застосування у прогнозуванні ринкових тенденцій криптовалют».

Кваліфікаційна робота і рецензія додаються.

Директор ННІІТ _____ Катерина НЕСТЕРЕНКО
(підпис) *(Ім'я, ПРІЗВИЩЕ)*

Висновок керівника кваліфікаційної роботи

Здобувач Дмитро ПРОКОПЕНКО представив кваліфікаційну роботу, присвячену мультиканальному сентимент-аналізу для прогнозування ринкових тенденцій криптовалют. Робота викладена чітко, науковою мовою і відповідає встановленим стандартам оформлення. До переваг слід віднести системність, послідовність викладення матеріалу та застосування прогресивних методів машинного навчання для аналітичних завдань. Результати дослідження можуть бути успішно впроваджені у практиці прогнозування високоволатильних фінансових ринків.

Все це дозволяє оцінити виконану кваліфікаційну роботу здобувача Дмитра ПРОКОПЕНКА на оцінку «_____» та присвоїти йому кваліфікацію магістр з системного аналізу.

Керівник кваліфікаційної роботи _____ Михайло КУЗЬМІЧ
(підпис) *(Ім'я, ПРІЗВИЩЕ)*

«____» _____20_____року

Висновок кафедри про кваліфікаційну роботу

Кваліфікаційна робота розглянута. Здобувач *Прокопенко Д.О.* допускається до захисту даної роботи в Екзаменаційній комісії.

Завідувач кафедрою ІСТ _____ Каміла СТОРЧАК
(назва) *(підпис)* *(Ім'я, ПРІЗВИЩЕ)*

ВІДГУК РЕЦЕНЗЕНТА
на кваліфікаційну роботу
на здобуття освітнього ступеня магістра

здобувача вищої освіти Прокопенка Дмитра Олександровича
(прізвище, ім'я, по батькові)

на тему «Мультиканальний сентимент-аналіз та його застосування в прогнозуванні ринкових тенденцій криптовалют»

Актуальність.

Кваліфікаційна робота є актуальною та відповідає вимогам сучасного аналізу фінансових ринків, зокрема ринку криптовалют. Використання мультиканального сентимент-аналізу для прогнозування ринкових тенденцій є новаторським підходом з практичним значенням для інвесторів та трейдерів. Здобувач продемонстрував високий рівень теоретичних та практичних знань у галузі аналізу даних і машинного навчання

Позитивні сторони.

1. Практична значущість: Запропонований підхід має потенціал для впровадження в аналітичні системи криптовалютних ринків.
2. Інтеграція різних джерел даних: В роботі реалізовано підхід, який передбачає використання даних з різних каналів, що дозволяє забезпечити більш точний і комплексний аналіз ринкових тенденцій.
3. Комплексний підхід: Здобувач продемонстрував системний підхід до вирішення завдання, детально описавши методи збору, обробки та аналізу даних, що є важливою складовою роботи.

Недоліки.

1. Обмежений аналіз альтернатив: У роботі не було проведено достатнього порівняння з іншими підходами або інструментами, що використовуються для аналізу ринкових тенденцій у криптовалютах.
2. Стислість теоретичної частини: Не достатньо детально розглянути питання щодо теоретичних засад сентимент-аналізу, розвитку криптовалютних ринків.

Відзначені зауваження не впливають на загальну позитивну оцінку кваліфікаційної роботи магістерської

Висновок: *кваліфікаційна робота на здобуття ступеня магістра заслуговує оцінку "_____", а здобувач Дмитро Прокопенко заслуговує присвоєння кваліфікації магістр з системного аналізу*

Рецензент:

науковий ступінь, вчене звання	*підпис*	*Ім'я, ПРІЗВИЩЕ*

ЗМІСТ

ВСТУП

З розвитком цифрових технологій та глобалізацією фінансових ринків криптовалюти стали однією з найбільш обговорюваних і перспективних сфер інвестицій. Їхній ринок характеризується високою волатильністю та значною залежністю від настроїв учасників. У цьому контексті виникає необхідність у розробці інструментів, які дозволяють аналізувати багатоканальну інформацію для передбачення ринкових тенденцій.

Актуальність теми обумовлена високою волатильністю ринків криптовалют, що вимагає нових підходів до аналізу даних, зокрема з використанням мультиканального сентимент-аналізу. Така методологія дозволяє враховувати широкий спектр інформації з соціальних мереж, форумів, новинних платформ та інших джерел. Розробка таких підходів є актуальною не лише для глобального ринку, а й для України, враховуючи зростаючий інтерес до криптовалют та блокчейн-технологій у країні.

У сфері криптовалют значну увагу приділяють дослідженням на основі аналізу великих даних (Big Data) та застосуванню методів машинного навчання для передбачення ринкових тенденцій. Наприклад, робота Джейна, Джохарі та Делгібабу (2023) [1] аналізує вплив активності в Twitter, зокрема настроїв твітів, на ціни криптовалют. Інше дослідження Єлесковича та Макая (2023) [2] розглядає вплив інформативних твітів на поведінку трейдерів та рух цін криптовалют у короткостроковій перспективі. Автори виявили статистично значущі зміни в рівнях прибутковості протягом перших трьох хвилин після публікації твітів, що підкреслює важливість соціальних мереж у формуванні ринкових тенденцій. Водночас питання інтеграції даних з різних джерел з метою формування цілісної картини ринку залишаються недостатньо опрацьованими. Це, а також необхідність адаптації методів до специфіки високоволатильного ринку криптовалют, зумовлює потребу у подальших дослідженнях.

Метою роботи є розробка ефективного підходу до мультиканального сентимент-аналізу для прогнозування ринкових тенденцій криптовалют. Для досягнення цієї мети необхідно вирішити такі *завдання*:

— проаналізувати існуючі методи сентимент-аналізу та їх застосування у криптоіндустрії;

— розробити методику збору та обробки даних з різних інформаційних каналів;

— інтегрувати та адаптувати методи машинного навчання для аналізу настроїв ринку;

— перевірити ефективність розробленого підходу на реальних даних.

Об'єктом дослідження виступає процес передбачення ринкових тенденцій у криптовалютній індустрії.

Предметом дослідження виступає мультиканальний сентимент-аналіз, який дозволяє об'єднати дані з різних джерел для аналізу впливу настроїв на цінові коливання криптовалют.

Методи дослідження. У роботі використано такі методи, як:

— сентимент-аналіз текстів із застосуванням технологій глибокого навчання (LSTM, трансформери);

— методи обробки великих даних, включаючи збирання даних із соціальних мереж та API;

— кількісний аналіз, зокрема кореляційний аналіз між настроями аудиторії та ринковими змінами;

— візуалізація даних для оцінки результатів.

Наукова новизна роботи полягає у розробці інтегрованого підходу до аналізу настроїв на основі мультиканальних даних, що враховує специфіку високоволатильного криптовалютного ринку. Практична значущість полягає у можливості застосування розробленої методології для створення автоматизованих інструментів прогнозування, які можуть використовуватись як у приватному секторі, так і на державному рівні.

Практична значущість полягає у можливості застосування розробленої методології для створення автоматизованих інструментів прогнозування, які можуть використовуватись як у приватному секторі, так і на державному рівні.

Апробація результатів та публікації були представлені на науковій конференції "II всеукраїнська науково-технічна конференція "Технологічні горизонти: дослідження та застосування інформаційних технологій для технологічного прогресу України і світу"", що відбувалася 18.11.2024 в Державному університеті інформаційно-телекомунікаційних технологій та викладені у публікації. Теми робіт "Мультиканальний підхід до сентимент-аналізу як інструмент прогнозування ринкових коливань криптовалют" та "Алгоритми мультиканального сентимент-аналізу в контексті системного прогнозування фінансових ринків".

Теоретична, методична та практична значущість представлена тим, що дані результати можуть бути адаптовані для інших ринків, таких як акції чи валютні ринки, а також використані для створення освітніх матеріалів у сфері системного аналізу.

Структура роботи: робота складається зі вступу, трьох розділів, висновків, списку використаних джерел та додатків.

РОЗДІЛ 1. ТЕОРЕТИЧНІ ОСНОВИ СЕНТИМЕНТ-АНАЛІЗУ ТА КРИПТОВАЛЮТНИХ РИНКІВ

1.1 Поняття та методи сентимент-аналізу

Сентимент-аналіз (sentiment analysis, SA), або аналіз тональності, є міждисциплінарним науковим напрямом, що поєднує методи обробки природної мови (Natural language processing, NLP), машинного навчання (Machine learning, ML) та статистичного аналізу з метою автоматичного виявлення та класифікації емоційного забарвлення текстових повідомлень.

Під "*сентиментом*" зазвичай розуміють суб'єктивне ставлення автора до певного об'єкта, події чи явища. Це ставлення може бути позитивним, негативним або нейтральним і відіграє ключову роль у формуванні тональності тексту. Початково сформований у галузі лінгвістики та обробки тексту, сентимент-аналіз з часом став важливим інструментом у маркетингових, соціологічних, політичних та фінансових дослідженнях.

Зростання обсягів текстових даних, доступних в мережі Інтернет (новинні статті, публікації у соціальних мережах, відгуки про товари, коментарі на форумах тощо) зумовлює необхідність розробки та вдосконалення інструментів, здатних швидко, масштабовано й точно інтерпретувати настрої користувачів чи суспільства. У контексті фінансових ринків, а особливо ринків криптовалют, здатність автоматизованих алгоритмів оперативно визначати тональність публікацій має значний вплив на прийняття рішень інвесторами. Своєчасна інтерпретація настроїв дозволяє покращити прогнозування цінових трендів, ідентифікувати потенційні маніпулятивні інформаційні кампанії та загалом розуміти динаміку ринкових очікувань.

Вихідною точкою є поняття "*сентименту*", що охоплює спектр суб’єктивних емоційних та оцінних реакцій. Сентимент можна розподілити за трьома базовими категоріями:

— *позитивний*: свідчить про схвалення, ентузіазм чи задоволення;

— *негативний*: виражає критику, розчарування або занепокоєння;

— *нейтральний*: не містить яскраво вираженого емоційного забарвлення.

У дослідженнях, спрямованих на глибинний аналіз емоцій, категорії можуть деталізуватися до ширшого спектра станів (гнів, страх, радість, сум тощо) та їхніх комбінацій. Це має особливе значення для складних сфер, зокрема фінансової, де емоційний тон повідомлення може бути завуальованим або контекстуально обумовленим. Наприклад, повідомлення про зростання прибутків компанії інтерпретується як позитивний сигнал, що вказує на потенційне підвищення вартості активу, тоді як новини про фінансові втрати чи організаційні проблеми найчастіше відображають негативні очікування.

Сентимент може вимірюватися кількісно за допомогою дискретних або неперервних шкал. Дискретні шкали, наприклад {-1, 0, +1}, дозволяють чітко класифікувати текст за категоріями тональності. Неперервні шкали, такі як інтервал від -100 до +100 або від 0% до 100%, забезпечують більш точну оцінку інтенсивності емоційного забарвлення. Вибір методу квантифікації залежить від конкретних завдань: дискретна оцінка краще підходить для розв’язання класифікаційних задач, тоді як неперервна — для глибинного аналізу тонких відтінків настрою.

Розширеним поняттям у сентимент-аналізі є "*емоційний тон*", що відображає не лише полярність (позитивну, негативну чи нейтральну), а й спектр різноманітних емоцій. Інтенсивність та варіативність емоційного тону стають критичними чинниками у ситуаціях, де ступінь емоційного впливу може істотно впливати на рішення, наприклад, у випадку оцінки інвестиційного клімату. Одна й та сама лексична одиниця може набувати різного емоційного значення залежно від контексту та стилю автора, що ускладнює автоматичну інтерпретацію.

1.1.1 Еволюція поняття “сентимент-аналіз”

Поняття сентимент-аналіз вперше з'явилось в 1990-х роках. На початкових етапах розвитку сентимент-аналізу основною метою було автоматичне визначення емоційного забарвлення текстів. Методи, які застосовувалися, базувалися на статистичних підходах та побудові словників. Словниковий підхід передбачав створення переліку слів із позитивним і негативним значенням, що дозволяло класифікувати текст відповідно до частоти появи цих слів.

Приклади:

— позитивні слова: "успіх", "прибуток", "щастя";

— негативні слова: "збиток", "криза", "невдача".

Попри простоту й ефективність у базових завданнях, такі словники мали значні обмеження. Зокрема, вони ігнорували контекстуальні аспекти, такі як синонімія, полісемія чи зміна значення слова залежно від оточуючих фраз. Наприклад, слово "проблема" могло мати негативний контекст ("велика проблема"), але в іншому випадку використовуватися нейтрально ("виявлення проблеми").

Ранні дослідження того часу, такі як робота М. Херста (1992), зробили важливий внесок у розуміння основ автоматизованого аналізу текстів, хоча їхній підхід був обмежений низькою обчислювальною потужністю та невеликою кількістю доступних даних [3].

Зі стрімким розвитком обчислювальних технологій у 2000-х роках сентимент-аналіз перейшов до використання методів машинного навчання. Це був ключовий момент, коли на зміну статичним словникам прийшли адаптивні алгоритми, здатні "вчитися" на великих наборах даних.

Основні підходи цього періоду включали:

— *методи на основі частотності*, аналіз частоти появи певних слів чи фраз для оцінки їхнього впливу на емоційний тон тексту;

— *моделі класифікації*, такі як SVM та Наївний Байєс.

SVM (Support Vector Machine) — це метод, який розділяє тексти за класами (позитивний, негативний, нейтральний) шляхом побудови гіперплощини, що максимально віддалена від найближчих точок кожного класу. Цей підхід дозволяє ефективно класифікувати тексти, навіть у випадках, коли дані не є лінійно роздільними, завдяки використанню ядерних функцій для відображення даних у вищий вимір, де розділення стає можливим [4].

Наївний Байєс (Naive Bayes) — це статистичний алгоритм класифікації, який базується на теоремі Байєса та припущенні незалежності ознак. У контексті сентимент-аналізу він використовує умовні ймовірності для передбачення настрою тексту, оцінюючи ймовірність того, що текст належить до певного класу (позитивного, негативного чи нейтрального) на основі частоти появи слів у навчальному наборі даних. Хоча припущення незалежності між словами є спрощенням, цей метод часто демонструє високу ефективність при обробці текстів завдяки своїй простоті та швидкості роботи [4].

Дослідження Б. Панга та Л. Лі (2002) були піонерськими у впровадженні цих методів у сентимент-аналіз. Вони показали, що моделі на основі SVM демонструють вищу точність порівняно зі словниковими методами, особливо при роботі з великими текстовими корпусами, такими як відгуки клієнтів чи коментарі [4].

Популярність соціальних мереж, таких як Facebook і Twitter, відкрила нові джерела для аналізу. Це стимулювало створення масштабних навчальних датасетів і привело до використання більш складних моделей, здатних розпізнавати тон навіть у коротких і неструктурованих текстах.

Технологічний прорив у 2010-х роках, зокрема розробка моделей глибокого навчання, здійснив справжню революцію в сентимент-аналізі. Завдяки високій обчислювальній потужності й доступності великих масивів даних з’явилася можливість аналізувати тексти на значно вищому рівні точності.

Серед ключових інновації можна виділити наступні:

— LSTM (Long Short-Term Memory) — архітектура рекурентних нейронних мереж (RNN), здатна враховувати довготривалі залежності між словами. Це дозволило враховувати контекст тексту, що особливо важливо для розуміння складних синтаксичних структур [5].

— BERT (Bidirectional Encoder Representations from Transformers) — модель, яка аналізує двосторонній контекст тексту. Вона одночасно враховує значення слова з позицій його попередніх і наступних слів, що значно покращує точність класифікації [6].

— GPT (Generative Pre-trained Transformer) — глибока трансформерна модель, яка використовується не лише для класифікації настрою, але й для генерації текстів, що відповідають певному емоційному забарвленню [7].

Наукові експерименти підтвердили, що глибокі моделі значно перевершують традиційні методи. Наприклад, застосування BERT у завданнях сентимент-аналізу дозволяє досягати точності понад 90% у великих та складних корпусах, таких як огляди фільмів чи аналіз публікацій у соціальних мережах [8].

Революція глибокого навчання також відкрила можливості для багатомовного сентимент-аналізу. Наприклад, моделі, подібні до XLM-R (Cross Lingual Models), адаптовані для роботи з текстами різними мовами, що особливо корисно для глобальних компаній.

1.1.2 Методи Методичні підходи до вибору інструментів сентимент-аналізу

Сентимент-аналіз є ключовим інструментом для обробки природної мови, що дозволяє визначати емоційну полярність текстів. У контексті цього завдання виділяють два основних підходи: *лексико-орієнтовані* та *машинно-навчальні методи*. Ефективність їх використання залежить від специфіки даних і поставлених завдань, що обумовлює необхідність детального аналізу та обґрунтованого вибору відповідних інструментів.

Лексико-орієнтовані методи базуються на використанні заздалегідь створених словників сентименту, що містять слова із визначеною емоційною полярністю (позитивною, негативною чи нейтральною). Аналіз тексту відбувається шляхом підрахунку загальної полярності на основі вагових значень слів. Серед популярних словників виділяються Sentiment140, WordNet-Affect та SentiWordNet [9].

Машинно-навчальні підходи використовують алгоритми, що навчаються на великих корпусах текстів із заздалегідь маркованими тональностями. Це дозволяє автоматично враховувати контекст тексту і визначати специфічні патерни. До класичних методів належать Наївний Байєс, SVM та логістична регресія [4], а сучасні підходи включають нейронні мережі та трансформерні моделі (BERT, GPT, RoBERTa) [6]. Основні переваги та недоліки даних підходів зібрані в таблиці 1.1.

Таблиця 1.1

Переваги та неоліки лексико-орієнтованих та машинно-навчальних методів

Метод	Переваги	Недоліки
Лексико-орієнтований	1. Простота реалізації, алгоритми легко інтегрувати в системи, і вони не потребують значних обчислювальних ресурсів. 2. Легко інтерпретувати результати завдяки простій структурі словників 3. Не потребують навчальних даних.	1. Нездатність адаптуватися до нових слів чи термінів (терміни, сленг, акроніми). 2. Ігнорування контексту. 3. Обмежена гнучкість, не здатні адаптуватися до динамічних змін у мовленні.

Продовження таблиці 1.1

Машинно-навчальний	1. Висока точність при роботі зі складними текстами (врахування контексту, іронії, специфічної лексики). 2. Можливість адаптації до змін у мовному середовищі шляхом донавчання. 3. Підходять для аналізу великих обсягів даних.	1. Високі обчислювальні витрати. 2. Потребують великих маркованих корпусів для навчання. 3. Складність реалізації та інтеграції в системи..

Окрім вищезазначених підходів існують також гібридні. Вони поєднують переваги лексико-орієнтованих і машинно-навчальних підходів. Наприклад, лексичні словники застосовуються на етапі попередньої обробки тексту, тоді як машинно-навчальні моделі враховують контекст і семантичні зв’язки. Ці методи забезпечують вищу точність при аналізі неоднорідних даних. Унікальні особливості текстів у криптовалютному середовищі, такі як сленг, швидка зміна термінології та високий рівень "шуму", ускладнюють аналіз [10].

При виборі гібридних підходів необхідно враховувати неоднорідність даних, динамічність інформаційного простору, а також вимогу оперативного аналізу в режимі реального часу для прийняття ефективних рішень.

Унікальність криптовалютних ринків у тому, що їх динаміка не залежить виключно від економічних фундаментальних факторів, як у традиційних фінансових ринках. Замість цього криптовалюти, як Bitcoin чи Ethereum, чутливі до інформаційних хвиль, що поширюються через соціальні медіа, форуми та пошукові системи.

1.2 Криптовалютні ринки: особливості та сучасні тенденції

1.2.1 Концептуальні засади функціонування криптовалютних ринків

Криптовалютні ринки є унікальним явищем у сучасній фінансовій системі, оскільки вони об'єднують інноваційні технології, економічні принципи та соціальні аспекти, які взаємодіють у децентралізованому середовищі. Основу їхнього функціонування становлять блокчейн-технології, криптографія та механізми консенсусу, які забезпечують безпеку, прозорість і довіру в системі, незалежній від централізованих установ.

Перший крок до створення криптовалютних ринків було зроблено з появою біткоїна у 2008 році, який описаний у відомій «білій книзі» Сатоші Накамото. Ця концепція запропонувала модель децентралізованої платіжної системи, що не залежить від банків чи урядів, і заснована на використанні криптографічних підписів для підтвердження транзакцій [11]. Біткоїн став першою цифровою валютою, емісія якої контролюється заздалегідь визначеним алгоритмом, що обмежує максимальну кількість монет (21 мільйон). Така архітектура створює ефект дефляційності, що відрізняє криптовалюти від фіатних валют, які емітуються центральними банками без жорстких обмежень [11].

Ключовою характеристикою криптовалют є їх децентралізована структура. Усі транзакції записуються в блокчейн — розподілений реєстр, який підтримується мережею вузлів. Блокчейн забезпечує незмінність даних, адже кожен блок містить криптографічно зв'язаний хеш попереднього блоку. Будь-яка спроба змінити інформацію в одному блоці вимагає змін у всьому ланцюгу, що практично неможливо без значних ресурсних витрат [**Error! Reference source not found.**]. Така архітектура підвищує довіру до системи, адже кожен учасник може перевірити достовірність записів.

Механізми консенсусу, такі як Proof of Work (PoW) та Proof of Stake (PoS), є фундаментальними для функціонування децентралізованих мереж. PoW, що використовується в біткоїні, вимагає від учасників (майнерів) вирішення складних криптографічних задач для додавання нових блоків до ланцюга. Це забезпечує цілісність системи, але вимагає значних енергетичних витрат [12]. Альтернативою є PoS, де учасники створюють нові блоки на основі кількості криптовалют, які вони володіють, що зменшує потребу в енергоресурсах.

Економічна динаміка криптовалютних ринків визначається попитом і пропозицією. Відсутність централізованого регулювання створює умови, за яких ціни на активи формуються винятково ринковими механізмами. Це робить ринки криптовалют дуже чутливими до новин, макроекономічних чинників та інформаційних впливів [13]. Наприклад, заяви впливових осіб, таких як Ілон Маск, можуть суттєво змінювати ринкові настрої, що спричиняє значні коливання цін.

Висока волатильність є ще однією визначальною характеристикою криптовалют. Вона обумовлена порівняно невеликою капіталізацією ринку, а також нерівномірним розподілом ліквідності між торговельними майданчиками. Ліквідність надається централізованими (CEX) і децентралізованими біржами (DEX), які виконують роль посередників для купівлі та продажу активів. Торговельні платформи також є джерелом миттєвих цінових орієнтирів, що впливають на загальну ринкову динаміку [13].

Інноваційні аспекти, такі як смарт-контракти, децентралізовані фінанси (DeFi) та токенізація, є рушійною силою розширення функціональності криптовалютних ринків. Смарт-контракти, впроваджені блокчейнами другого покоління, такими як Ethereum, дозволяють автоматизувати виконання угод, створюючи нові можливості для інтеграції в різні сектори економіки [20]. Токенізація активів дозволяє перенести у цифрове середовище права власності на реальні активи, такі як нерухомість чи мистецтво, що значно розширює потенційний ринок.

Важливим елементом функціонування криптовалютних ринків є саморегулювання через відкритість коду і спільне управління мережами. Учасники можуть вносити пропозиції щодо змін у протоколах та брати участь у їх прийнятті через механізми токенізованого голосування. Такий підхід забезпечує гнучкість і адаптацію системи до нових викликів, не потребуючи втручання зовнішніх регуляторів [20].

Незважаючи на незалежність від державного контролю, криптовалютні ринки не ізольовані від макроекономічних і політичних чинників. Інфляція, зміни процентних ставок або регуляторні заходи можуть суттєво впливати на настрої інвесторів. Наприклад, заборона криптовалют у Китаї спричинила значний відтік майнерів та вплинула на глобальну хешрейт-мережу [13].

Криптовалютні ринки є також об'єктом глибокого дослідження в контексті інформаційних потоків. Інструменти аналізу настроїв використовуються для оцінки впливу новин та соціальних медіа на динаміку цін. Такий підхід дозволяє прогнозувати тренди та приймати обґрунтовані інвестиційні рішення. Це робить крипторинок не лише фінансовим, але й інформаційно-залежним середовищем, де дані відіграють критично важливу роль.

Таким чином, концептуальні засади функціонування криптовалютних ринків поєднують технологічні інновації, економічні закономірності та вплив інформаційних потоків. Це створює унікальну екосистему, яка постійно еволюціонує, формуючи нові виклики та можливості для учасників фінансових ринків.

1.2.2 Моделі поведінки учасників крипторинку

Розвиток криптовалютних ринків обумовлений складною взаємодією різноманітних груп учасників, кожна з яких відіграє специфічну роль у формуванні цін, волатильності та загальних ринкових тенденцій. Криптовалютний ринок відзначається високим рівнем інформаційної асиметрії, нестабільністю нормативно-правового середовища та швидким технологічним прогресом, що стимулює

диференціацію поведінкових моделей його суб’єктів. Серед ключових учасників ринку можна виділити наступних:

— індивідуальні інвестори;

— інституційні учасники;

— професійні трейдери;

— маркет-мейкери;

— майнери;

— валідатори;

— розробники протоколів і проєктів.

Кожна з цих категорій характеризується особливими мотивами, стратегіями прийняття рішень, ступенем раціональності й емоційності, а також рівнем використання технічних та фундаментальних методів аналізу.

Дрібні інвестори

Дрібні інвестори, або індивідуальні роздрібні учасники ринку, відіграли важливу роль на ранніх етапах становлення криптовалют, коли основною рушійною силою зростання цін ставали публічні обговорення, новизна активу та «хайп» у медіа. Їхня поведінка часто базується на емоційних факторах і обмеженій обізнаності щодо ризиків. Серед ключових мотивів можна виділити:

— *Ефект новизни та хайпу*. Для багатьох роздрібних учасників криптовалюти є привабливим активом завдяки інноваційності та потенційно високим прибуткам. Вони схильні купувати або продавати активи під впливом публічного резонансу, думок лідерів думок та засобів масової інформації.

— *Бажання швидкого прибутку*. Дрібні інвестори часто переоцінюють свої можливості та ігнорують ризики, орієнтуючись на історії успіху ранніх власників біткоїна чи інших криптовалют. Це призводить до формування «цінових бульбашок», коли попит штучно роздмухується надмірним оптимізмом.

— *Вплив соціальних мереж*. Telegram, Twitter, Reddit та інші платформи є основним джерелом інформації. Часто інформація неперевірена або емоційно

забарвлена, що спричиняє ефект «стадного інстинкту» та формування інформаційних каскадів. Як наслідок, виникає підвищена волатильність, оскільки панічні продажі (panic selling) або покупки на піку ціни (FOMO) стають типовою реакцією дрібних інвесторів на негативні чи позитивні новини відповідно.

Дослідження Д. Гарсії та ін. (2014) [14] свідчать, що на ринку біткоїна інформаційні каскади та наслідування поведінки інших призводять до циклів гіперболізованого зростання та обвалів. Таким чином, емоційна нестабільність дрібних інвесторів є важливим чинником волатильності крипторинку.

Інституційні інвестори

Поява інституційних гравців — інвестиційних банків, хедж-фондів, страхових компаній та пенсійних фондів — стала новим етапом у розвитку криптовалютних ринків. Їхні поведінкові моделі суттєво відрізняються від підходів дрібних інвесторів через:

— *Фундаментальний аналіз*. Інституційні учасники оцінюють базові чинники, такі як рівень впровадження технологій, регуляторні зміни, глобальні макроекономічні тренди та перспективи масштабування інфраструктури.

— *Диверсифікація портфеля*. Криптовалюти розглядаються як інструмент хеджування або засіб розширення класів активів з метою зниження загальних ризиків.

— *Системність та прагматичність*. Інституційні інвестори активно використовують складні моделі ризик-менеджменту, автоматизовані торговельні алгоритми та регулярний моніторинг ринкових умов, що допомагає уникати панічних реакцій та мінімізує вплив емоцій.

Згідно з дослідженням К. Лопез-Мартіна та ін. (2021) [15], залучення інституційного капіталу сприяє підвищенню ефективності ринку. Інституційні учасники, завдяки значним обсягам капіталу та раціональному прийняттю рішень, часто виступають стабілізаторами ринку, знижуючи його волатильність та сприяючи більш прогнозованому ціноутворенню.

Професійні трейдери та маркет-мейкери

Професійні трейдери займають проміжне положення між інституційними інвесторами та дрібними учасниками. Їхня поведінка базується на:

— *Технічному аналізі*. Професійні гравці широко використовують історичні дані, графіки та індикатори для прогнозування короткострокових змін цін.

— *Алгоритмічній торгівлі та високочастотних стратегіях*. Застосування автоматизованих систем дозволяє отримувати прибуток від локальних коливань та арбітражних можливостей.

— *Арбітражі*. Використання відмінностей у цінах між різними біржами чи ринками зменшує цінові аномалії.

Дослідження Д. Коутмоса (2018) [16] підтверджує роль алгоритмічної торгівлі у формуванні короткострокової динаміки цін. Професійні трейдери менш схильні до емоційних чинників, що допомагає збалансувати локальні відхилення ціни та забезпечує певну стабільність.

Маркет-мейкери виконують роль стабілізаторів ліквідності, виставляючи заявки на купівлю та продаж активів і тим самим зменшуючи спред. Їхня поведінка визначається математичними моделями, які мінімізують вплив інформаційного шуму та суб'єктивних ринкових очікувань. Маркет-мейкери зазвичай дотримуються нейтральної позиції щодо трендів, фокусуючись на стабільному прибутку від різниці цін. Завдяки цьому знижується волатильність та покращуються умови для виконання великих ордерів без значного впливу на ціну.

Майнери та валідатори

Майнери та валідатори є невід'ємною частиною технологічної інфраструктури криптовалютних мереж. Вони не лише забезпечують функціонування блокчейнів, але й суттєво впливають на економічну стабільність та безпеку.

Економічна поведінка майнерів *(Proof of Work)*. залежить від поточної ціни криптовалюти, вартості обладнання, складності майнінгу та енергоспоживання. При зниженні цін частина майнерів може припинити роботу, зменшуючи загальний

хешрейт і, відповідно, безпеку мережі. Дж. Алмедія та Т. К. Гонсалвес (2018) [4] продемонстрували, що флуктуації цін безпосередньо впливають на структуру та стабільність майнінгового середовища.

Валідатори (Proof of Stake та інші механізми) прагнуть довгострокових інтересів, оскільки їхні винагороди залежать від підтримки стабільності мережі. У системах з делегованим стейкінгом (DPoS) репутація валідаторів є ключовим фактором довіри серед користувачів. Їхня мотивація включає отримання пасивного доходу та укріплення позицій у мережі.

Розробники протоколів та проєктів

Команди розробників відіграють центральну роль у підтримці та оновленні блокчейн-протоколів, розвитку функціональності та забезпеченні технічної сумісності між різними екосистемами. Вони впливають на ринкові очікування через:

— *Впровадження технічних оновлень*: Зміни в протоколах, хардфорки та оптимізація мереж можуть підвищувати пропускну здатність, покращувати безпеку та забезпечувати кращу взаємодію з іншими блокчейнами.

— *Формування довіри*: Прозорість процесів прийняття рішень, чітко окреслені дорожні карти розвитку та ефективна комунікація зі спільнотою впливають на довгострокову цінність проєкту.

Згідно з дослідженням Х. Каталіні та Дж. С. Ганса (2016) [18], успіх криптопроєктів значною мірою залежить від прозорості та своєчасності технологічних оновлень, а також від стабільної взаємодії з інвесторами та користувачами.

Децентралізоване управління (DAO) та поведінкова економіка

У рамках децентралізованих автономних організацій (DAO) учасники ринку впливають на розвиток проєктів шляхом голосування за зміни протоколу. Така модель управління моделюється за допомогою теорії ігор та механізмів стимулювання. Ефективно побудована система винагород за участь у прийнятті рішень сприяє підвищенню стійкості та довгостроковій стабільності мережі.

На криптовалютному ринку особливого значення набуває поведінкова економіка. Когнітивні упередження, емоції (FOMO, паніка) та вплив соціальних мереж формують ірраціональну складову прийняття рішень. Н. Барберіс та ін. (2016) [19] вказують, що емоційний аспект впливає на поведінку інвесторів на фінансових ринках, а у випадку крипторинку він посилюється меншою регулятивною базою та високою інформаційною асиметрією.

Взаємодія різних груп та формування ринкової динаміки

Ринкова динаміка виникає з поєднання стратегій та інтересів різних учасників. Інституційні інвестори та професійні трейдери сприяють підвищенню ефективності та прогнозованості ринку, тоді як дрібні інвестори підсилюють волатильність. Арбітражні можливості, що використовуються трейдерами та маркет-мейкерами, допомагають усунути цінові аномалії, знизити інформаційну асиметрію та наблизити ринок до більш збалансованого стану.

Таким чином, моделі поведінки учасників крипторинку охоплюють широкий спектр мотивацій та стратегій — від емоційно забарвлених рішень дрібних інвесторів до раціональних, алгоритмічно оптимізованих стратегій інституційних та професійних гравців. Поєднання поведінкових, економічних та технологічних аспектів створює унікальну структуру, що потребує комплексного аналізу. Розуміння цих моделей є ключовим для прогнозування ринкових трендів, розробки ефективних інвестиційних стратегій і оцінки впливу окремих груп учасників на формування загальної ринкової динаміки.

1.2.3 Сучасний стан та динаміка розвитку криптовалютних ринків

У сучасних умовах формування глобальної економічної архітектури криптовалютні ринки перебувають у стані динамічної трансформації, поступово переходячи від нішевого сегмента, що характеризував початкові етапи їх розвитку, до повноцінної складової світової фінансової екосистеми. Цей процес супроводжується розширенням спектра криптоактивів, зростанням їх ринкової капіталізації та

ліквідності, появою й розвитком інституційної інфраструктури, поширенням децентралізованих фінансових екосистем, а також урізноманітненням регуляторних підходів і зростанням ролі макроекономічних чинників.

Варто відзначити, що загальну тенденцію до збільшення капіталізації та залученості учасників простежено навіть попри періодичні спади та корекції, викликані коливаннями ринкових настроїв, новинними подіями, а також чинниками, пов'язаними з обмеженістю інфраструктури на ранніх етапах розвитку галузі [11]. Наприклад, якщо спочатку біткоїн відігравав домінантну роль, фактично виступаючи синонімом криптовалютного ринку, то сьогодні ситуація значно ускладнилася завдяки появі широкого спектра токенів та цифрових активів з різними функціональними характеристиками. Окрім біткоїна та ефіріуму, важливе значення мають платформи нового покоління (Solana, Cardano, Polkadot), що зосереджені на підвищеній масштабованості та гнучкості смарт-контрактів, а також стейблкоїни, які прагнуть стабілізувати вартість за рахунок прив'язки до фіатних валют чи матеріальних активів [13]. Додатковий імпульс розвитку було надано появою DeFi-токенів, що дали змогу формувати цілісну екосистему децентралізованих фінансів, а також NFT, які, розширюючи можливості для токенізації унікальних об'єктів, створили нові прикладні сценарії використання блокчейну в галузі мистецтва, колекціонування, медіа та розваг.

Інституціоналізація криптовалютних ринків стала можливою завдяки тому, що великі інвестиційні установи, традиційні банки, хедж-фонди, а також пенсійні та страхові фонди поступово визнають криптоактиви як потенційний елемент інвестиційного портфеля. Мотивом їхнього входження у криптовалютний простір стає не лише прагнення до диверсифікації ризиків і пошуку нових джерел прибутковості, але й прагнення захистити активи від інфляційних процесів, які дедалі частіше виявляються у фіатній грошовій системі. Значним каталізатором цього процесу є поява регульованих фінансових інструментів — таких як біткоїн-ETF, ф'ючерси та опціони, — що забезпечують більш прозору інфраструктуру для

інституційних інвесторів [13]. Ці нові продукти сприяють формуванню зрозуміліших правових рамок, знижуючи нормативні та операційні ризики, а також підвищуючи загальний рівень довіри до криптовалют як до повноцінного класу активів.

Одним із ключових драйверів сучасного стану криптовалютних ринків є технологічні інновації. Зокрема, розвиток децентралізованих фінансів (DeFi) фактично створив умови для перенесення традиційних фінансових послуг на інфраструктуру блокчейну: користувачі отримали доступ до децентралізованого кредитування, страхування та торгівлі без участі посередників і з максимально можливим ступенем прозорості [20]. Істотною перевагою тут виступає використання смарт-контрактів, які автоматизують виконання умов угод, забезпечують відкритий доступ до транзакцій та зводять до мінімуму корупційні ризики й фінансові зловживання. Проте широке розповсюдження DeFi зіштовхнулося з обмеженнями масштабованості базових блокчейн-мереж, що спричинило значні транзакційні витрати та затримки. У відповідь на ці виклики активно розробляються й упроваджуються рішення другого рівня (Layer 2), покликані збільшити пропускну здатність мережі, знизити собівартість операцій та відкрити шлях для подальшого розширення користувацької бази.

Не менш важливим фактором впливу на динаміку криптовалютних ринків залишається регуляторне середовище, що досі перебуває у стані пошуку оптимального балансу між заохоченням інновацій і захистом інвесторів та фінансової стабільності. У той час як одні юрисдикції впроваджують заходи, спрямовані на жорстке обмеження або навіть заборону криптовалют, інші створюють сприятливі правові й податкові умови, розробляють стандартизовані підходи до класифікації криптоактивів та механізмів боротьби з відмиванням коштів і фінансуванням тероризму. Хоча відсутність єдиної глобальної регуляторної політики ускладнює міжнародну діяльність та формує додаткові ризики для учасників, зрозумілі та прозорі норми здатні забезпечити підвищення довіри до ринку, сприяти залученню нових

інституційних гравців та знизити загальний рівень шахрайства й правової невизначеності.

Додаткову роль у формуванні стану криптовалютних ринків відіграють макроекономічні та геополітичні чинники. У періоди, коли традиційні економічні системи зазнають суттєвих деформацій, викликаних, наприклад, надмірною грошовою емісією, геополітичними конфліктами, санкціями чи торговельними війнами, криптовалюти часто розглядаються як потенційні засоби хеджування проти інфляції й девальвації національних валют. Таке ставлення підсилюється поширенням ідеї про біткоїн як «цифрове золото», що відіграє роль «тихої гавані» в часи макроекономічної нестабільності. У деяких державах цифрові активи навіть набувають офіційного статусу платіжних засобів, а провідні платіжні системи, такі як Visa та Mastercard, інтегрують криптовалютні розрахунки у свої платформи, відкриваючи двері для практичного повсякденного використання [12].

Процес дорослішання та переходу до більш структурованого ринкового середовища також підтверджується появою розгалуженої інфраструктури аналітичних інструментів, кастодіальних сервісів, рейтингових агентств та індексів криптоактивів. Це свідчить про те, що ринок поступово виходить за межі «дикого» періоду раннього розвитку, коли інвестори покладалися здебільшого на інтуїцію й чутки, а безпекова складова була недостатньо розвиненою. Сьогодні ж учасники ринку можуть користуватися все більш доступними інструментами для оцінювання проєктів, моніторингу ончейн-даних, аналізу ринкових трендів, прогнозування цінових рухів та управління ризиками, що у підсумку закріплює позиції криптовалютних ринків у світовій фінансовій системі та формує підґрунтя для їх подальшого сталого розвитку.

Таким чином, сучасний стан криптовалютних ринків характеризується складною взаємодією між динамічним зростанням капіталізації та диверсифікацією активів, інституціоналізацією та впровадженням регульованих фінансових інструментів, технологічними інноваціями у сфері децентралізованих фінансів,

урізноманітненими регуляторними стратегіями, а також впливом макроекономічних і геополітичних чинників. Ця комплексна взаємодія чинників, попри періодичну волатильність та регуляторну невизначеність, сприяє поступовому формуванню більш зрілої, прозорої та інтегрованої інфраструктури, що у перспективі дозволить криптовалютам набути статусу повноцінного та сталого елемента глобальної фінансової екосистеми.

1.3 Роль інформаційних джерел у формуванні ринкових настроїв

1.3.1 Класифікація інформаційних каналів для ринкових настроїв

Інформаційні канали відіграють ключову роль у формуванні ринкових настроїв, адже вони забезпечують доступ учасників ринку до новин, аналітики, чуток, думок лідерів думок та інших комунікаційних сигналів. Знання про класифікацію цих каналів дозволяє більш цілеспрямовано підходити до збору й аналізу даних, формувати методологічну базу для подальшого застосування мультиканального сентимент-аналізу. Серед основних типів інформаційних каналів, які застосовуються для аналізу ринкових тенденцій, можна виділити наступні:

— традиційні засоби масової інформації (ЗМІ);

— соціальні мережі та мікроблоги;

— аналітичні платформи та агрегатори даних;

— офіційні канали проєктів та компаній;

— аналітичні та дослідницькі центри, університетські дослідження;

— неофіційні, напівформальні канали та альтернативні медіа.

Традиційні ЗМІ, такі як новинні агентства (Bloomberg, Reuters) та фінансові портали (The Wall Street Journal, BBC), охоплюють широкий спектр тем, включаючи новини про регуляцію, технологічні інновації та глобальні економічні тенденції. Інформація з цих джерел зазвичай має високий рівень достовірності, оскільки створюється професійними журналістами та редакційними колективами. Профільні

журнали та бізнес-аналітика надають глибшу аналітику та прогнози, публікуючи оглядові статті, аналітичні звіти та результати досліджень експертів. Телебачення та радіо, зокрема фінансові канали, через прямі ефіри з коментарями аналітиків та інтерв'ю з топ-менеджерами компаній, оперативно реагують на події та впливають на миттєві настрої широкої аудиторії.

Платформи соціальних мереж, такі як Twitter, LinkedIn та Facebook, дозволяють швидко поширювати новини та думки лідерів думок, включаючи CEO компаній, аналітиків та популярних трейдерів. Швидкість поширення інформації тут надзвичайно висока, а контент часто емоційно забарвлений, що може призводити до сплесків волатильності на ринку. Reddit та тематичні форуми, наприклад r/CryptoCurrency чи r/Bitcoin, є платформами, де користувачі обговорюють криптопроєкти, інвестиційні стратегії та технологічні оновлення. Інформація тут менш структурована й перевірена, але відображає настрої мас та може слугувати індикатором поведінки дрібних інвесторів. Telegram-канали, Discord-сервери та інші месенджери пропонують замкнуті чи напівзамкнуті спільноти трейдерів та ентузіастів окремих проєктів. Інформація тут може бути дуже оперативною, інколи інсайдерською, але й високоризиковою з точки зору достовірності.

Сервіси ончейн-аналітики, такі як Glassnode та Nansen, надають дані про транзакції та баланси гаманців, супроводжуючи їх коментарями, інфографікою та аналітичними висновками. Такий контент відображає "настрій" мережі та поведінкові патерни учасників ринку. Агрегатори новин та рейтингів, наприклад CoinMarketCap та CoinGecko, збирають котирування з різних бірж, новини з медіа та думки експертів, пропонуючи цілісний огляд ринку, що спрощує моніторинг настроїв.

Офіційні вебсайти та блоги розробників криптовалютних протоколів та пов'язаних компаній використовуються для публікації дорожніх карт, технічних оновлень та партнерств. Контент тут зазвичай формальний, перевірений та менш емоційно забарвлений, хоча може мати піарну орієнтацію. Активність у репозиторіях

коду, таких як GitHub, може вказувати на прогрес проєкту та динаміку розробки, що впливає на довіру спільноти інвесторів.

Наукові публікації та консалтингові звіти, підготовлені дослідницькими інститутами, університетськими лабораторіями та консалтинговими агенціями, такими як Deloitte, PwC та Ernst & Young, містять ґрунтовні звіти з аналізом ринку, прогнозами розвитку та оцінкою ризиків. Ці джерела часто слугують базою для формування стратегічного та довгострокового бачення настроїв. Спеціалізовані індекси, розроблені на основі багатьох інформаційних каналів, такі як Crypto Fear & Greed Index, служать композитним показником загального настрою ринку, допомагаючи інвесторам оцінити поточний психоемоційний стан ринкових гравців.

Подкасти, відеоблоги та стріми на платформах, таких як YouTube та Twitch, з участю відомих аналітиків можуть формувати громадську думку та задавати "моду" на певні проєкти. Хоча їхній контент менш формалізований та перевірений, харизматичні автори здатні впливати на інвесторські настрої. Меми та інтернет-культура, представлені жартівливими зображеннями, коміксами та короткими відео, також слугують індикатором соціальних настроїв. Меми можуть миттєво поширюватися в соціальних мережах, відображаючи колективну реакцію на події. Хоча вони не є надійним джерелом фактів, меми часто сигналізують про зміну настроїв у спільноті, слугуючи важливим елементом неформальної комунікації.

Класифікація інформаційних каналів дозволяє створити структуру для збору й аналізу даних, забезпечуючи повноту та надійність інформації. Застосування цієї структури у мультиканальному сентимент-аналізі сприяє побудові точніших прогнозів ринкових настроїв. Наприклад, аналіз соціальних мереж може забезпечити розуміння короткострокових коливань настроїв, тоді як офіційні звіти та аналітичні платформи формують основу для довгострокових стратегічних рішень.

Крім того, класифікація допомагає адаптувати методи аналізу до специфіки кожного каналу. Для обробки даних із соціальних мереж можуть використовуватися алгоритми обробки природної мови (NLP), тоді як для аналізу інформації з

аналітичних платформ та індексів ефективно застосовуються методи статистичного аналізу та машинного навчання.

1.3.2 Взаємозв'язок між новинами, соціальними трендами та ринковими рухами

Динаміка цін на криптовалютному ринку значною мірою визначається інформаційним середовищем, у якому новини, повідомлення із соціальних мереж, висловлювання лідерів думок та загальні культурно-соціальні тренди перетворюються на ключові стимули для формування ринкових очікувань і поведінки інвесторів. Інформація різного типу – від офіційних регуляторних заяв до неформальних коментарів у соціальних мережах – може впливати як на короткострокові, так і на довгострокові зміни вартості цифрових активів, створюючи складну динаміку попиту й пропозиції.

Традиційні новинні потоки, зокрема публікації у провідних медіа, часто задають загальний тон ринковим настроям. Повідомлення про запровадження регуляторних обмежень або, навпаки, легалізації криптовалют у певних юрисдикціях нерідко провокують миттєві зміни настрої інвесторів, що є причиною цінових коливань. Зміцнення нормативно-правової бази та повідомлення про технологічні інновації чи партнерства підвищують довіру інвесторів і стимулюють зростання цін, тоді як новини про заборони або посилення контролю з боку великих економік провокують панічні розпродажі та зниження вартості активів.

Не менш важливою є роль соціальних мереж, які завдяки швидкому поширенню інформації здатні миттєво впливати на настрої інвесторів. Платформи на кшталт Twitter (X), Reddit, Telegram та Discord забезпечують надзвичайно високі темпи циркуляції контенту, включно з чутками та непідтвердженими повідомленнями. Емоційний характер таких обговорень зумовлює спекулятивні коливання: позитивні коментарі та прогнози стимулюють активні закупівлі, тоді як негативні оцінки та панічний настрій призводять до масового розпродажу. Авторитетні особи або лідери

думок із широкою аудиторією можуть одним коментарем кардинально змінити ринкову динаміку, підсилюючи або нівелюючи вплив попередніх новинних сигналів.

Ширші культурні та соціальні тренди формують макрофон для сприйняття інформації. Зростання популярності децентралізованих фінансів, NFT чи метавсесвітів створює загальні позитивні очікування щодо блокчейн-індустрії, підвищуючи ймовірність того, що новини про криптовалютні проєкти будуть інтерпретовані як сприятливі для інвестування. Навіть мем-культура, як у випадку з Dogecoin чи іншими «мемними» токенами, здатна стимулювати аномальне зростання цін, оскільки широке онлайн-схвалення та емоційний піднесення інколи переважають раціональні економічні аргументи.

Посилення інформаційних сигналів у різних каналах спілкування – офіційні медіа, соціальні платформи, тематичні форуми – може спричиняти каскадний ефект. Один вагомий імпульс, наприклад, дані про злам великої біржі або запровадження нового податку, породжує лавину коментарів, інтерпретацій та прогнозів. Часте повторення сигналу у різних джерелах підсилює його вплив, трансформуючи короткотермінову реакцію у триваліший тренд, що може змінювати не лише поточні, а й довгострокові очікування учасників ринку.

Сучасні аналітичні підходи, такі як методи обробки природної мови та машинного навчання, дозволяють кількісно оцінити вплив інформаційних чинників на ринкову динаміку. Аналіз текстових даних для виявлення сентименту та ключових тем відкриває можливості створювати нові прогностичні індикатори, що доповнюють традиційні фундаментальні й технічні методи аналізу. Таким чином, комплексний підхід до розуміння інформаційного середовища – включно з аналізом новинних потоків, соціальних трендів та культурних феноменів – надає глибші уявлення про природу ринкових коливань, допомагаючи оптимізувати торговельні стратегії, мінімізувати ризики та підвищити якість інвестиційних рішень.

РОЗДІЛ 2. МУЛЬТИКАНАЛЬНИЙ ПІДХІД ДО СЕНТИМЕНТ-АНАЛІЗУ

2.1 Джерела даних для мультиканального аналізу (соціальні мережі, новини, форуми, блоги)

Мультиканальний підхід до сентимент-аналізу передбачає залучення різноманітних джерел даних, які відображають настрої та думки користувачів в інформаційному просторі. Завдяки цьому вдається отримувати ширше уявлення про ринкові тенденції та більш точно прогнозувати поведінку учасників ринку. Соціальні мережі, новинні ресурси, форуми та блоги відіграють у цьому процесі ключову роль, оскільки дають можливість вивчати громадську думку, відстежувати реакції користувачів майже в реальному часі, а також отримувати глибші аналітичні висновки.

Соціальні мережі (Twitter, Facebook, Reddit, Telegram тощо) формують основу для моніторингу громадської думки в режимі реального часу. Вони пропонують надзвичайно швидкий потік інформації, що має критичне значення для виявлення раптових змін у настроях, зумовлених заявами провідних гравців ринку, неочікуваними регуляторними новинами або значними транзакціями на біржах. Моніторинг кількості згадувань Bitcoin, Ethereum та інших криптоактивів допомагає ідентифікувати короткострокові тренди й дає змогу відстежувати потенційно маніпулятивні або спекулятивні кампанії.

Новинні ресурси, порівняно із соціальними мережами, публікують перевірену й глибоко проаналізовану інформацію. Це редакційно вивірені матеріали, які часто містять експертні коментарі, огляди та прогнози. Публікації на спеціалізованих ресурсах на кшталт CoinDesk чи CoinTelegraph вирізняються детальним описом технологічних новацій, правових питань або інвестиційних аспектів, пов’язаних із криптовалютною сферою. На цих ресурсах часто містять посилання на офіційні документи, коментарі фахівців, дослідження аналітичних компаній і розгорнуті

огляди ринку, що допомагає формувати глибше розуміння нинішньої кон'юнктури. Хоча новини з'являються з певною затримкою, порівняно з соціальними мережами, вони надають контекстуально ширші та ґрунтовніші висновки, які мають істотний вплив на настрої інвесторів і формують довгострокові уявлення про ринок.

Форуми, зокрема Bitcointalk і різні тематичні розділи на Reddit, виступають своєрідними «банками ідей» і експертних думок. Користувачі обговорюють тут надзвичайно специфічні питання, пов'язані з алгоритмами блокчейну, особливостями смарт-контрактів чи безпекою мережі. Учасниками часто є розробники, майнери, засновники стартапів і досвідчені інвестори, що володіють унікальною експертизою. На відміну від динамічних соціальних мереж, де обговорення швидко змінюють напрям, форуми зазвичай підтримують структуровану дискусію, даючи змогу повернутися до раніше піднятих тем і порівняти їх з оновленими даними.

Блоги охоплюють широкий спектр форматів — від коротких авторських колонок до великих аналітичних досліджень, що базуються на інтерв'ю з професіоналами галузі або проведених експериментах. Деякі автори обирають вузьку спеціалізацію (наприклад, DeFi-проєкти, NFT-платформи, алгоритми konsensusy), тоді як інші мають більш загальний погляд, аналізуючи макроекономічні чинники, історію розвитку криптовалют чи загальні тренди діджиталізації. Хоча суб'єктивність є невіддільною рисою блогів, саме вона часом допомагає розширити спектр аналітичних прийомів і поглянути на ті самі події з різних ракурсів.

Таблиця 2.1

Особливості даних основних каналів у межах мультиканального аналізу

Канал	Особливості даних
Соціальні мережі	- Дані в реальному часі, дають змогу оперативно реагувати на події. - Велика кількість реакцій, зручне середовище для швидкого сентимент-аналізу. - Хештеги та тематичні обговорення, які спрощують пошук релевантної інформації.
Новинні ресурси	- Висока достовірність: редакційна перевірка й журналістські стандарти. - Глибокі аналітичні матеріали: дають фундаментальне уявлення про ринок. - Обмежена інтерактивність: повільніший зворотний зв'язок, але ретельно перевірений контент.

Продовження таблиці 2.1

Форуми	- Експертні обговорення: часто містять глибокі технічні та аналітичні інсайти. - Повільніше оновлення: реакції з'являються з відставанням у часі. - Можливі інсайдерські відомості: учасниками можуть бути представники реальних проєктів.дотичності учасників до реальних проєктів і компаній.
Блоги	- Аналітичний підхід: автори подають власне бачення та прогнози. - Суб'єктивність контенту: матеріали часто відображають особисту точку зору. - Прогностична спрямованість: можливість отримати неординарні ідеї та нетривіальні прогнози.вузькоспеціалізованих аспектів можна розробляти прогнози щодо ринкових трендів.

Загалом, мультиканальний збір даних забезпечує надійну основу для застосування алгоритмів машинного навчання та методів обробки природної мови (NLP), які дають змогу класифікувати та оцінювати тональність контенту. Використання кількох каналів зменшує ймовірність похибок, пов'язаних із зависоким рівнем шуму в якомусь одному джерелі, і дає змогу отримати цілісні результати аналізу. Висока оперативність соціальних мереж у сукупності з глибиною й достовірністю новин, експертністю форумів і унікальністю блогового контенту створюють синергію, яка дає можливість дослідникам будувати багатоаспектну картину ринку.

Завдяки цьому підходу можна вчасно виявляти ймовірні сигнали про перегрів ринку, масштабні корекції або потенційні «міхури», пов'язані з надмірним ажіотажем. Крім того, мультиканальний аналіз сприяє кращій валідності прогнозів, адже дозволяє залучити більше структурованих і неформалізованих даних, інтегруючи різні кути зору на проблематику. Це особливо важливо у сфері криптовалют і блокчейну, де інформаційний простір розвивається вкрай динамічно та часто є фрагментованим між різними платформами.

2.2 Методи збору та попередньої обробки даних з різних каналів

2.2.1 Інструменти збору даних

Для збору даних у рамках мультиканального сентимент-аналізу використовується широкий спектр інструментів, що забезпечують інтеграцію з різними джерелами текстової інформації. Цей процес є надзвичайно важливим, оскільки якість зібраних даних визначає точність аналізу. Збір даних охоплює кілька основних напрямків: отримання даних із соціальних мереж, новинних ресурсів, блогів, форумів, а також обробку мультимедійного контенту. Кожен із цих напрямків потребує використання специфічних інструментів та технологій.

Одним із ключових джерел текстових даних є соціальні мережі, які надають доступ до публікацій через програмні інтерфейси (API). Наприклад, Twitter API дозволяє отримувати твіти за ключовими словами, хештегами або від конкретних акаунтів, надаючи не лише текст, але й метаінформацію, таку як час публікації, геолокація та кількість взаємодій (лайків, ретвітів). Подібним чином, Facebook Graph API забезпечує доступ до публічних постів, коментарів і статистики, що дозволяє аналізувати взаємодію користувачів із контентом. У свою чергу, YouTube Data API використовується для отримання інформації про відео, коментарі, кількість переглядів і вподобань, що є критично важливим для оцінки громадської думки щодо відеоконтенту.

Ще одним цінним джерелом даних є новинні ресурси, які можна інтегрувати через RSS-канали. За допомогою RSS-агрегаторів, таких як Feedparser, можливо автоматично отримувати оновлення з новинних сайтів. Цей метод є ефективним для збору даних з великих обсягів публікацій у засобах масової інформації, що важливо для аналізу трендів і суспільних настроїв у контексті новинних подій.

Інформація з форумів та блогів, як правило, є неструктурованою, що ускладнює її обробку. Для збору таких даних використовуються спеціалізовані парсери, як-от BeautifulSoup або Scrapy. Ці інструменти дозволяють витягувати текстові дані зі

сторінок, беручи до уваги унікальні особливості структури HTML кожного ресурсу. Форумні обговорення часто містять унікальні перспективи та відображають суспільні настрої, які складно знайти в інших джерелах.

Коли джерело не надає API, для збору даних використовуються скраппери, які автоматизують процес витягування контенту з вебсторінок. Зокрема, бібліотека Selenium є незамінною для роботи з динамічними сторінками, що використовують JavaScript для відображення контенту. Для великих обсягів даних і високої швидкості обробки краще підходить Scrapy, який дозволяє одночасно витягувати дані з багатьох вебсторінок.

Для джерел, що включають мультимедійний контент, використовуються сервіси перетворення аудіо та відео у текст. Інструменти на кшталт Google Speech-to-Text або AssemblyAI забезпечують автоматичну транскрипцію аудіо чи відеофайлів у текстовий формат. Це дозволяє інтегрувати дані з подкастів, відеороликів чи інших мультимедійних джерел у загальний аналіз, що значно розширює можливості мультиканального підходу.

2.2.2 Процес збору даних для мультиканального підходу

Процес збору даних для мультиканального підходу до сентимент-аналізу передбачає чітко структуровані етапи, кожен з яких виконує конкретну функцію в підготовці текстової інформації до аналізу. Нижче наведено основні етапи цього процесу:

— отримання даних у “сирому” (RAW) форматі;

— фільтрація та попередня обробка даних;

— структуризація та збереження даних.

Отримання даних у RAW-форматі

На першому етапі збору даних для мультиканального сентимент-аналізу забезпечується доступ до текстової інформації з різноманітних джерел, що є основою для подальшого аналізу. Цей процес передбачає використання різних інструментів і

методів для інтеграції даних із соціальних мереж, новинних сайтів, блогів, форумів або платформ із мультимедійним контентом. Вихідні дані, отримані на цьому етапі, зазвичай мають "сирий" формат (неструктуровані або напівструктуровані) і потребують додаткової обробки.

Одним із ключових інструментів для отримання даних є API соціальних мереж, які дозволяють доступ до структурованих даних безпосередньо з платформ. Наприклад, Twitter API надає можливість отримувати твіти, фільтруючи їх за ключовими словами, хештегами або ідентифікаторами акаунтів. Дані зазвичай надаються у форматі JSON і включають не тільки текстове наповнення, але й метаінформацію, таку як час публікації, геолокація та статистика взаємодії (ретвіти, лайки). Подібні функції виконує Facebook Graph API, який надає доступ до публічних постів і коментарів, а також до статистичних даних, наприклад, охоплення та взаємодії з контентом. У випадку відеоконтенту платформа YouTube Data API дозволяє отримувати текст опису, коментарі до відео та інформацію про метрики переглядів і вподобань.

Для вебсайтів, які не надають доступ через API, використовуються методи парсингу HTML-сторінок. Інструменти, такі як BeautifulSoup та Scrapy, дозволяють отримувати текстову інформацію безпосередньо з HTML-коду вебсторінок. Цей підхід вимагає врахування специфіки структури сторінок кожного окремого ресурсу, оскільки розміщення контенту може суттєво відрізнятися. Якщо вебсторінки використовують динамічне генерування контенту за допомогою JavaScript, то стандартні парсери можуть бути неефективними. У таких випадках застосовуються скраппінгові інструменти, такі як Selenium, які емулюють поведінку користувача у веббраузері та дозволяють отримати повний вміст динамічних сторінок.

Іншим важливим джерелом інформації є RSS-канали новинних сайтів, які забезпечують регулярне оновлення контенту у стандартизованому форматі. Використання бібліотек, таких як Feedparser, дозволяє автоматизувати збір даних із

численних новинних ресурсів, забезпечуючи доступ до заголовків статей, основного тексту та публікаційних даних.

Для збору інформації з мультимедійного контенту, такого як відео чи аудіо, використовуються сервіси перетворення аудіо/відео у текст. Наприклад, інструменти на кшталт Google Speech-to-Text або AssemblyAI виконують автоматичну транскрипцію, перетворюючи звукові доріжки чи субтитри у текстовий формат. Це дозволяє інтегрувати дані з подкастів, відеоблогів чи аудіофайлів у загальний аналіз.

На цьому етапі зібрані дані мають "сирий" вигляд — JSON-файли, HTML-код, транскрибований текст або інші формати. Вони ще не придатні для аналітичної обробки й потребують подальшої фільтрації, стандартизації та структурування, що відбувається на наступних етапах. Проте саме цей початковий крок забезпечує доступ до ключової інформації, яка є основою для мультиканального підходу в сентимент-аналізі.

Фільтрація та попередня обробка даних

Попередня обробка та очищення даних є фундаментальним етапом у підготовці текстової інформації для подальшого аналізу, особливо в контексті мультиканального підходу до аналізу крипторинку. Цей процес складається з кількох підетапів, кожен з яких має специфічні цілі та методи реалізації. Мета попередньої обробки полягає у видаленні шуму, стандартизації текстів і забезпеченні їх релевантності до завдань аналізу. У цьому контексті виділяються три основні напрями:

— видалення дублікатів, спаму та реклами;

— мовна нормалізація;

— фільтрація за мовою.

Видалення дублікатів, спаму та рекламних повідомлень є критично важливим етапом попередньої обробки текстових даних, який дозволяє підвищити якість вибірки та забезпечити точність аналізу. У текстових даних, особливо в великих масивах інформації з соціальних мереж, блогів або форумів, часто присутні повторювані повідомлення, спам або реклама, які створюють значний інформаційний

шум і знижують репрезентативність результатів. Для усунення цих проблем застосовуються кілька методологічних підходів, що охоплюють різні аспекти обробки даних.

Першим завданням на цьому етапі є видалення дублікатів. Повторювані повідомлення можуть виникати внаслідок технічних помилок, зловживання алгоритмами автоматичного постингу або масового поширення однакових текстів для підвищення охоплення. Для виявлення дублікатів використовується перевірка хешів повідомлень. Кожне текстове повідомлення конвертується у хеш-код за допомогою алгоритмів, таких як MD5 або SHA-256. Ці алгоритми генерують унікальний код для кожного тексту, що дозволяє швидко ідентифікувати повторювані записи. Повідомлення з однаковими хешами вважаються ідентичними та видаляються. Окрім цього, для виявлення майже однакових текстів, які можуть відрізнятися лише незначними варіаціями (наприклад, додатковими символами чи перестановкою слів), використовуються алгоритми порівняння текстів. Популярними методами є Cosine Similarity і Jaccard Similarity, які аналізують схожість текстів на основі векторних або множинних представлень слів.

Наступним аспектом є видалення спаму. Спам-повідомлення зазвичай характеризуються високою частотою певних ключових слів або фраз, спрямованих на привернення уваги (наприклад, "безкоштовно", "виграй", "знижка"). Для їхньої фільтрації проводиться частотний аналіз тексту: повідомлення, що перевищують встановлений поріг частотності ключових слів, маркуються як спам. Окрім того, використовується аналіз шаблонних повідомлень за допомогою регулярних виразів. Наприклад, регулярні вирази дозволяють виявляти повідомлення, що містять посилання на сумнівні сайти, електронні адреси або типові комерційні заклики на кшталт "Натисніть тут" чи "Зареєструйтеся зараз". Такі повідомлення видаляються, щоб уникнути спотворення результатів аналізу.

Третім важливим завданням є видалення рекламних повідомлень, які не є спамом у класичному сенсі, але мають комерційний характер і можуть знижувати

релевантність аналізу, орієнтованого на настрої користувачів. Для ідентифікації рекламних текстів використовуються такі методи, як аналіз наявності посилань. Повідомлення, що містять URL-адреси або ключові слова, типові для маркетингових текстів (наприклад, "знижка", "акція", "пропозиція"), автоматично маркуються як рекламні. У складніших випадках застосовуються методи контекстного аналізу, які враховують не тільки ключові слова, але й загальний зміст повідомлення. Для цього використовуються алгоритми обробки природної мови (NLP), що аналізують семантичний контекст тексту. Наприклад, алгоритми на основі Word2Vec або BERT можуть ідентифікувати рекламний характер тексту навіть за відсутності явних маркерів, таких як посилання чи типові ключові слова.

Мовна нормалізація є важливим етапом у підготовці текстових даних, який забезпечує їх стандартизацію, зменшує варіативність і покращує якість для подальшого аналізу. Основна мета цього процесу полягає у тому, щоб перетворити текст у формат, який є найбільш зручним для алгоритмів обробки природної мови (NLP). Це включає приведення слів до стандартної форми, видалення зайвих елементів, що не несуть смислового навантаження, та оптимізацію структури тексту. Основні техніки, які використовуються на цьому етапі, охоплюють лематизацію, стемінг, видалення стоп-слів та очищення тексту від зайвих символів.

Однією з ключових технік є лематизація, яка спрямована на приведення слів до їхньої базової або словникової форми. Наприклад, слова "running", "ran" і "runner" будуть приведені до базової форми "run". Цей підхід дозволяє значно зменшити кількість унікальних слів у тексті, зберігаючи при цьому їхній смисл. Лематизація є особливо корисною у випадках, коли слова мають численні граматичні форми, оскільки вона допомагає усунути ці варіації без втрати важливої інформації. Для виконання лематизації використовуються спеціалізовані інструменти, такі як SpaCy або NLTK, які підтримують багатомовну обробку тексту і забезпечують високу точність завдяки використанню мовних моделей і словників.

Альтернативною, хоча і менш точною технікою, є стемінг. На відміну від лематизації, стемінг спрямований на обрізання слова до його кореня шляхом видалення закінчень і суфіксів. Наприклад, слова "running" і "runner" скорочуються до "run". Стемінг є швидшим, ніж лематизація, оскільки не потребує звернення до мовного словника, проте його результати можуть бути менш точними. Наприклад, корінь слова "better" у стемінгу буде "bett", що не відповідає його базовій формі. Для реалізації стемінгу часто використовуються такі інструменти, як Snowball Stemmer або Porter Stemmer, які є ефективними і простими у використанні.

Наступним важливим елементом мовної нормалізації є видалення стоп-слів. Стоп-слова, такі як "the", "and" в англійській мові або "та", "і" в українській, не несуть смислового навантаження і лише збільшують обсяг текстових даних, не додаючи цінності для аналізу. Видалення цих слів допомагає зменшити шум у тексті та покращити результати аналізу. Для цього використовуються вбудовані списки стоп-слів у бібліотеках, таких як NLTK, або створюються спеціалізовані списки відповідно до специфіки завдання. Наприклад, у випадку аналізу текстів криптовалютного ринку можна додати специфічні слова, які часто зустрічаються, але не мають цінності для аналізу, наприклад "crypto", якщо це слово є загальним для всіх повідомлень.

Останньою важливою складовою мовної нормалізації є видалення зайвих символів. Це включає видалення пунктуації, спеціальних символів (наприклад, "!", "@", "#"), URL-адрес і емодзі, які не несуть смислового навантаження для більшості аналітичних завдань. Крім того, текст часто приводиться до нижнього регістру, щоб уникнути розбіжностей між великими та малими літерами (наприклад, "Bitcoin" і "bitcoin" розглядаються як однакові слова).

Криптовалютний ринок, будучи глобальним явищем, характеризується багатомовністю текстового контенту, який генерується користувачами, аналітиками, медіа та іншими зацікавленими сторонами. Для забезпечення точності та релевантності аналізу критично важливо виконати фільтрацію текстів за мовою. Цей процес дозволяє ефективно розподілити дані на мовні групи, що враховують

специфіку кожної мови, і забезпечити їх коректну обробку з використанням спеціалізованих алгоритмів.

На першому етапі фільтрації виконується визначення мови тексту. Це завдання вирішується за допомогою автоматизованих алгоритмів, які оцінюють мову тексту на основі частотного аналізу символів і слів. Одним із найбільш поширених інструментів є бібліотека langdetect, яка забезпечує швидкий і простий спосіб розпізнавання мови тексту. Цей підхід ґрунтується на статистичному аналізі граматичних структур і частотності символів, притаманних конкретним мовам. Наприклад, тексти англійською мовою характеризуються високою частотністю букв "e" та "t", тоді як українські тексти часто включають букви "і" та "ї".

Альтернативним і більш точним інструментом є Google Translator API, який, окрім визначення мови тексту, пропонує можливість автоматичного перекладу. Це може бути особливо корисним для аналізу текстів менш поширеними мовами або для інтеграції текстів різних мов у єдину мовну площину. Наприклад, автоматичний переклад текстів на англійську мову може значно спростити подальший аналіз за допомогою моделей, оптимізованих для англомовного контенту.

Після визначення мови тексту виконується розподіл текстів за мовними групами. Це дозволяє створити окремі набори даних для кожної мови, наприклад, англійської, української або будь-якої іншої мови, представленої у вибірці. Такий підхід забезпечує адаптацію методів обробки та аналізу до специфіки конкретної мови. Наприклад, для текстів англійською мовою можна використовувати спеціалізовані моделі обробки природної мови (NLP), такі як BERT або RoBERTa, натреновані на англомовних корпусах. Для текстів українською мовою можна застосовувати україномовні моделі, такі як lang-uk або багатомовні моделі, налаштовані на підтримку української мови.

Розподіл за мовами також дозволяє уникнути можливих помилок в аналізі, які виникають через змішування текстів різними мовами. Наприклад, багатомовні моделі часто можуть давати некоректні результати, якщо текст містить кодову мову

(змішування кількох мов у межах одного повідомлення). Розподіл текстів за мовами допомагає ізолювати такі випадки або попередньо обробляти їх для коректної обробки.

Структуризація та збереження даних

Формати зберігання даних можуть значно варіюватися залежно від характеру та структури інформації, а також від цілей подальшого аналізу чи використання. Основною метою правильного вибору формату є зручність доступу до даних, ефективна організація зберігання та максимально можлива оптимізація процесів обробки. Наприклад, коли дані мають чітку табличну структуру, доцільно звернутися до реляційних баз даних. У таких системах інформація впорядковується у вигляді таблиць із чітко визначеними полями, що забезпечує можливість виконання SQL-запитів та встановлення складних зв’язків між окремими таблицями. Це, у свою чергу, полегшує виконання операцій читання та запису, а також дає змогу з високою продуктивністю обслуговувати велику кількість користувачів. Найпоширенішими системами керування реляційними базами даних є, зокрема, PostgreSQL та MySQL, які пропонують широкий набір інструментів для організації, супроводу та масштабування сховища.

У випадках, коли структура даних змінюється або не є чітко визначеною від початку, а також тоді, коли мова йде про зберігання документів чи складних об’єктів із потенційно різноманітними полями, перевагу можуть мати нереляційні бази даних. До таких СУБД належать MongoDB та Cassandra, орієнтовані на зберігання напівструктурованих або навіть неструктурованих даних. Формат JSON у цьому контексті є одним із найпоширеніших, адже його гнучкість дає змогу зберігати як типові, так і специфічні поля без жорсткого задання схеми. Це стає особливо актуальним у сценаріях, коли дані швидко змінюють структуру, або коли потрібна масштабованість та висока пропускна здатність для операцій запису та читання, характерні для застосувань із великою кількістю одночасних з’єднань чи запитів.

Окрім баз даних, у практиці обробки та аналітики великих обсягів даних широко використовуються файлові формати. Найпростішим форматом для передачі структурованих даних між різними системами чи середовищами є CSV (Comma-Separated Values), де кожен рядок являє собою один запис, а поля відокремлені символом коми або іншим роздільником. Такий підхід полегшує обмін даними між різними програмними платформами та мовами програмування. Водночас для зберігання більших масивів інформації, які потребують частих операцій зчитування і вибірки, зручнішим є формат Parquet. Він забезпечує ефективне стиснення та пришвидшує виконання аналітичних запитів завдяки своїй колонковій структурі, що дає змогу завантажувати лише ті стовпці, які дійсно потрібні для певної операції. Це особливо важливо під час побудови масштабованих аналітичних рішень у середовищах, які працюють із розподіленими обчисленнями на зразок Apache Spark чи Dask.

Окрему увагу слід приділити також датафреймам, оскільки вони часто використовуються у процесі дослідницької аналітики, побудови та навчання моделей машинного навчання чи під час виконання різноманітних маніпуляцій із даними безпосередньо в оперативній пам’яті. Наприклад, у Python бібліотека Pandas пропонує потужний функціонал для зчитування, фільтрації, трансформації та агрегування даних у форматі датафреймів, що суттєво спрощує попередню обробку перед подальшим використанням у алгоритмах машинного навчання або статистичного аналізу. У підсумку, вибір конкретного формату зберігання завжди продиктований тим, як і з якою метою дані будуть використовуватися, які вимоги існують до обсягів, швидкості доступу та масштабованості, а також наскільки часто змінюється структура записів.

Нижче наведено порівняльну таблицю з основними методами збереження даних, їхнім застосуванням, перевагами та недоліками (Таблиця 2.2).

Таблиця 2.2

Порівняння основних методів збереження даних

Метод збереження	Коли застосовувати	Приклади	Переваги	Недоліки
Реляційні БД	Коли дані мають чітку табличну структуру і важлива цілісність та зв’язки	PostgreSQL, MySQL	- Підтримка SQL - Надійні транзакції - Швидкий пошук	- Жорстка схема - Складність масштабування
Нереляційні БД (NoSQL)	Для гнучких, змінних структур та великих обсягів неструктурованих даних	MongoDB, Cassandra	- Гнучка модель даних - Висока масштабованість	- Немає уніфікованої SQL - Складніша аналітика
Файлові формати (CSV, Parquet)	Для обміну та зберігання великих наборів даних у файловому вигляді	CSV: простий текст Parquet: колонковий формат	- Простота імпорту/експорту (CSV) - Ефективне стиснення (Parquet)	- Необхідне додаткове ПЗ для Parquet - CSV займає більше місця
Датафрейми (in-memory)	Для інтерактивного аналізу та підготовки даних безпосередньо в оперативній пам’яті	Pandas/Dask (Python), DataFrame (R)	- Зручні для швидкої фільтрації та агрегування - Легке поєднання з ML-бібліотеками	- Обмеження за розміром оперативної пам’яті - Не підходять для великих виробничих систем

Зведені таблиці

Зведені таблиці дають змогу узагальнити та оцінити результати попередніх етапів обробки і фільтрації даних, наочно демонструючи кількість повідомлень, які залишаються у вибірці після кожного кроку. Такий формат представлення інформації дає змогу швидко визначити обсяги потенційно корисних даних і порівняти їх з початковою вибіркою (Таблиця 2.3).

Таблиця 2.3

Приклад зведеної таблиці

Етап обробки	**Кількість записів**	**Частка від початкового обсягу (%)**
Початкова вибірка	100,000	100
Після видалення дублікатів	85,000	85
Після видалення спаму	60,000	60
Релевантні дані	50,000	50

Окрім кількісної оцінки, зведені таблиці забезпечують прозорість процесу: аналітик може бачити, скільки даних було втрачено на кожному етапі очищення і з яких причин (наприклад, видалення дублікатів, вилучення спаму тощо). Завдяки цьому легше відстежувати вплив фільтрації на репрезентативність вибірки та приймати рішення щодо доцільності проведених змін. Якщо частка відсіяних даних виявляється надто великою, фахівець має підстави переглянути критерії видалення, а якщо навпаки — занадто мало даних вважаються нерелевантними, можливо, слід застосувати жорсткіші правила очищення. У такий спосіб зведена таблиця постає інструментом і для контролю якості, і для аналізу ефективності застосованих фільтрів. Це особливо важливо, коли йдеться про побудову моделей машинного навчання або проведення комплексних статистичних досліджень, де розмір та якість вихідної вибірки прямо впливають на результати аналізу.

2.3 Виклики та обмеження мультиканального аналізу

Мультиканальний аналіз даних, який передбачає одночасне використання різноманітних джерел інформації (соціальних мереж, офіційних новин, платформ для обміну повідомленнями тощо), стикається з низкою викликів, пов’язаних із гетерогенністю та неоднорідністю цих джерел. Однією з найістотніших проблем є

відмінності у тональності та стилі спілкування. У соціальних мережах комунікація зазвичай ведеться неформальним чином, із широким використанням скорочень, сленгу та емодзі, тоді як офіційні новинні ресурси дотримуються усталених журналістських стандартів, уникають сленгу та дотримуються чітких правил побудови тексту. Для мультиканальної системи аналізу це означає необхідність створювати універсальні алгоритми, здатні коректно інтерпретувати обидва види даних, розпізнаючи значущі нюанси тексту. Крім того, неоднорідність мовних ресурсів посилюється через наявність жаргонів, технічних термінів і професійного сленгу, прикладом чого є специфічні криптовалютні вирази на кшталт “moon” або “bear”. У різному контексті ці слова можуть мати відмінне значення, що вимагає ретельного контекстного аналізу та належної адаптації алгоритмів обробки природної мови до кожної конкретної платформи.

2.3.1 Гетерогенність та неоднорідність джерел

Однією з головних проблем мультиканального аналізу є гетерогенність джерел даних. Це включає відмінності у тональності, стилі спілкування, а також неоднорідність мовних ресурсів.

Відмінності у тональності та стилі спілкування

Соціальні мережі та офіційні новини характеризуються різними підходами до подання інформації. У соціальних мережах комунікація зазвичай неформальна, з використанням скорочень, сленгу, емодзі, що створює труднощі для автоматичних систем аналізу. Наприклад, в офіційних новинах застосовуються чіткіші формулювання, а стиль подання більш стандартизований, що спрощує обробку тексту. Для системи мультиканального аналізу це створює виклик у забезпеченні універсальності алгоритмів для обробки даних з таких різних джерел.

Неоднорідність мовних ресурсів

Важливим аспектом є наявність різних мовних варіантів: жаргону, технічних термінів або сленгу. Наприклад, слова на кшталт “moon” чи “bear” у контексті

криптовалют мають специфічне значення, яке залежить від контексту. Умови використання подібних термінів змінюються залежно від платформи, що вимагає складних алгоритмів контекстного аналізу.

2.3.2 Технічні та обчислювальні проблеми

Мультиканальний аналіз вимагає значних ресурсів, що створює низку технічних викликів, особливо при роботі з великими обсягами даних.

Витрати ресурсів на обробку великих даних

Обробка великих даних (“Big Data”) потребує високопродуктивного обладнання та значного часу для обчислень. Для аналізу потоків даних у реальному часі необхідні ефективні алгоритми, здатні працювати з високою швидкістю, зберігаючи точність. Наприклад, під час збору даних із Twitter у періоди підвищеної волатильності фінансових ринків обсяг твітів може зрости в кілька разів, створюючи навантаження на систему.

Необхідність масштабованості системи збору та аналізу

Масштабованість системи є важливою умовою для її ефективної роботи. Системи повинні адаптуватися до збільшення обсягів даних без втрати продуктивності. Це вимагає впровадження сучасних хмарних технологій, таких як AWS або Google Cloud, які дозволяють масштабувати обчислювальні потужності залежно від потреб.

Водночас низька якість даних, поширення ботів та спам-контенту становлять одну з найбільших загроз для достовірності результатів. У соціальних мережах під час політичних криз або різких коливань фінансових ринків відзначається зростання кількості ботів, які штучно формують потрібний наратив чи поширюють недостовірні дані. Зафіксовано випадки, коли в моменти раптового зростання ціни криптовалют кількість спам-твітів збільшувалася на 40%. Це істотно знижує відношення «сигналу до шуму» в загальному масиві зібраної інформації. Тому важливим завданням мультиканальних систем є впровадження механізмів фільтрації та пріоритезації

даних. Традиційні методи, що ґрунтуються на лінгвістичних і текстових ознаках, можуть бути доповнені сучасними моделями машинного навчання та глибинного аналізу тексту, однак такі моделі потребують великих навчальних вибірок і регулярного оновлення через динамічний характер мовлення.

Дані низької якості та шум

Одним із ключових обмежень мультиканального аналізу є присутність даних низької якості та шуму, які спотворюють результати.

Наявність ботів, фейкових новин та маніпуляцій

Соціальні мережі є основним джерелом дезінформації. Наприклад, у періоди політичних криз значно зростає кількість ботів, які поширюють фейкові новини або маніпулятивний контент. Це знижує точність аналізу, оскільки сентимент може бути навмисно спотворений.

Виклики у відфільтровуванні контенту низької якості

Фільтрування даних низької якості — складна задача, що потребує комплексного підходу. Використання методів машинного навчання, таких як класифікація контенту на основі текстових ознак, допомагає частково вирішити цю проблему. Однак моделі потребують навчання на великих вибірках, що ускладнює їхнє впровадження.

Правові та етичні аспекти

Мультиканальний аналіз також стикається з численними правовими та етичними обмеженнями, які необхідно враховувати під час збору та обробки даних.

Конфіденційність даних

Використання персональних даних із соціальних мереж викликає серйозні питання щодо конфіденційності. Необхідно забезпечити анонімізацію даних, аби уникнути порушення прав користувачів.

Відповідність регуляторним вимогам

Правила використання API платформ, таких як Twitter або Facebook, можуть обмежувати обсяг даних, доступних для аналізу. Крім того, відповідність політикам GDPR або інших регуляторних вимог є критично важливою для легітимності аналізу.

2.4 Порівняння мультиканального та одноканального підходів у сентимент-аналізі

У цьому розділі здійснено порівняння мультиканального та традиційного одноканального підходів до сентимент-аналізу на основі емпіричних даних та аналітичних висновків. Детально розглядаються методологічні відмінності, кількісні показники точності аналізу, якісні приклади застосування, а також формулюються рекомендації щодо практичного використання кожного підходу.

2.4.1 Методологічні відмінності між мультиканальним та одноканальним підходами

Широта охоплення джерел

Мультиканальний підхід базується на зборі даних одночасно з кількох каналів, таких як соціальні мережі, новинні портали, форуми, блоги тощо. Це дозволяє аналізувати широкий спектр точок зору, тем і настроїв, що значно перевершує можливості одноканального підходу, який зазвичай обмежується одним джерелом, наприклад, Twitter. У контексті вивчення суспільної думки або динаміки ринку, широта охоплення забезпечує більш репрезентативні результати.

Гнучкість у визначенні контексту

Одна з ключових переваг мультиканального підходу — можливість аналізу контекстуальних змін у тональності повідомлень. Різні платформи мають свою аудиторію, лексичні особливості та стилі комунікації. Наприклад, обговорення у форумах може бути технічно насиченим, тоді як у соціальних мережах превалює

емоційна складова. Мультиканальний аналіз дозволяє враховувати ці специфічні особливості, що сприяє точнішому визначенню настроїв та тенденцій.

Оперативність збору великих обсягів

Завдяки паралельному збору даних із різних платформ мультиканальний підхід забезпечує оперативний доступ до великих обсягів інформації. Це особливо важливо в умовах динамічних ринків або соціальних криз, коли зміна настроїв аудиторії відбувається швидко. Наприклад, під час несподіваних політичних подій або значних змін на ринку криптовалют швидкість збору даних може бути вирішальним фактором для аналізу та прогнозування.

Адаптивність до різних форматів

Мультиканальний підхід включає використання різноманітних інструментів збору даних, таких як API, веб-скрапінг, підписка на RSS-стрічки тощо. Це дозволяє оперативно додавати нові джерела інформації або адаптуватися до змін у вже існуючих. Наприклад, якщо одна з платформ змінює правила доступу до даних, мультиканальна система може швидко переключитися на альтернативні канали, мінімізуючи втрати в якості або обсязі зібраної інформації.

2.4.2 Якісні аспекти порівняння

Порівняльний аналіз різних джерел даних відіграє визначальну роль у формуванні об’єктивних висновків щодо ринкових настроїв, інвестиційних тенденцій та загального інформаційного фону. Однак у процесі дослідження слід ураховувати ряд якісних аспектів, які впливають на релевантність і точність отриманих результатів. Передусім важливо розглянути спотворення картини при одноканальному аналізі, що виникає через обмеження збору інформації лише з однієї платформи. Яскравим прикладом є використання Twitter як єдиного джерела: ця соціальна мережа часто містить значну кількість бот-акаунтів, а також велику питому вагу низькоякісного або відверто спамного контенту. Такий контент може цілеспрямовано накручувати негативну тональність чи поширювати надмір

критичних і провокативних твітів, які не відображають реальних настроїв цільової аудиторії. Унаслідок цього формується хибне уявлення щодо рівня скептицизму або негативу, що може вводити в оману дослідників та інвесторів і спотворювати оцінку реальних ринкових трендів. Окрім того, зосередження виключно на одній платформі звужує коло респондентів, оскільки користувачі, котрі зазвичай віддають перевагу іншим соціальним мережам, професійним форумам або спеціалізованим інформаційним ресурсам, залишаються поза увагою аналізу. Таким чином, одноканальний підхід не тільки підвищує ризик отримання необ'єктивних результатів, а й ставить під сумнів валідність висновків, оскільки реальна структура аудиторії та її інформаційні вподобання залишаються непредставленими.

З метою мінімізації вищезгаданих ризиків доцільно застосовувати мультиканальний аналіз, що передбачає збір, обробку та порівняння даних із різних соціальних мереж, профільних ЗМІ й аналітичних платформ. Така інтеграція різнорідних джерел дає змогу виявити закономірності та тенденції, які можуть залишитися непомітними за умови обмеження одним-єдиним каналом збору інформації. Наприклад, якщо на Twitter існує надлишок негативних висловлювань, спричинених штучним "накручуванням" тональності, водночас на професійному форумі чи в LinkedIn може домінувати радше нейтральна або позитивна оцінка тих самих ринкових подій. Отже, врахування даних із декількох джерел дозволяє "вирівняти" цей інформаційний фон і сформувати більш зважений, наближений до реального висновок про настрої та очікування ринку. Крім того, залучення мультиканальних даних підвищує репрезентативність вибірки, оскільки різні групи користувачів (від звичайних споживачів до аналітиків та експертів) мають власні звичні канали комунікації та способи поширення інформації. Такий комплексний погляд уможливлює поєднання кількісних і якісних вимірів: від вивчення загальних статистичних показників (кількість публікацій, частота згадувань) до глибинного аналізу змісту, тональності й контексту дописів, що формують колективну думку.

Таким чином, для отримання максимально об’єктивних результатів дослідникам варто застосовувати мультиканальний підхід, спираючись на принципи інтеграції різних джерел та компенсації потенційних похибок. Використання лише одного джерела, хоч і може бути спокусливо простим, несе в собі високу ймовірність викривлення реальної картини, пов’язану з впливом бот-активності, спаму та характерних особливостей саме цієї платформи. Натомість збалансований аналіз із використанням сукупності соціальних мереж, професійних та тематичних ресурсів забезпечує ширшу та надійнішу емпіричну базу для оцінювання настроїв і подальших прогнозів. У такий спосіб дослідник отримує змогу не лише скоригувати імовірні перекоси у висновках, а й ширше охопити різні сегменти аудиторії, що є важливою передумовою для формування достовірних рекомендацій, стратегій та інтерпретацій динаміки ринкових показників.

2.4.3 Висновки щодо надійності мультиканального аналізу при довгострокових прогнозах

Мультиканальний підхід до аналізу даних є важливим інструментом для довгострокового прогнозування ринкових трендів завдяки здатності інтегрувати інформацію з різних джерел і забезпечувати багатовимірний погляд на ринкову динаміку. Водночас, його ефективність не позбавлена певних обмежень і викликів, які необхідно враховувати для повного розуміння можливостей і ризиків такого підходу. Нижче наведено основні переваги та недоліки мультиканального аналізу в контексті довгострокових прогнозів.

Переваги мультиканального аналізу

1. Стабільність у мінливому інформаційному середовищі

Мультиканальний аналіз дозволяє адаптувати прогнози до швидких змін у зовнішньому середовищі. Ринки реагують на численні чинники, такі як економічні кризи, соціальні рухи або політичні конфлікти, які відображаються в різних інформаційних каналах. Завдяки інтеграції даних із соціальних медіа, фінансових

звітів, новинних джерел та інших платформ цей підхід забезпечує гнучкість і актуальність прогнозів. Це особливо важливо в умовах, коли довгострокові тенденції часто формуються на основі слабких, але стійких сигналів, які легко втратити, використовуючи лише одне джерело.

2. Зниження ризику однобічного аналізу

Різні канали інформації пропонують доповнюючі перспективи: соціальні медіа можуть відображати емоційний фон, новинні джерела — ключові події, а аналітичні звіти — експертні оцінки. Ця різнобічність дозволяє уникнути ситуацій, коли окремий канал формує викривлену картину. Наприклад, надмірна залежність від соціальних мереж може призводити до упередженості через інформаційний шум, але зважена інтеграція компенсує такі недоліки.

3. Аналіз складних явищ та сегментів

У складних і динамічних секторах, таких як криптовалюти або високотехнологічні ринки, мультиканальний аналіз забезпечує здатність враховувати широкий спектр взаємопов'язаних факторів. Це дозволяє врахувати не лише економічні показники, але й соціальні, технічні й навіть культурні аспекти, що є важливими для формування стійких довгострокових прогнозів.

Недоліки та виклики мультиканального аналізу

1. Висока залежність від якості даних

Незважаючи на переваги, мультиканальний аналіз може стати ненадійним, якщо якість даних є низькою. Наприклад, інформація з соціальних мереж часто містить багато шуму, спаму або свідомо викривлених даних. Аналіз таких джерел вимагає застосування складних методів фільтрації, які не завжди гарантують абсолютну точність. Крім того, відсутність стандартизації даних з різних джерел може ускладнювати їх інтеграцію та інтерпретацію.

2. Підвищена складність обробки даних

Інтеграція великих обсягів даних з багатьох джерел вимагає значних ресурсів, як обчислювальних, так і людських. Потрібні спеціалізовані інструменти для збору,

обробки й аналізу інформації, а також команди фахівців, здатних працювати з великими даними. Це може зробити мультиканальний підхід недоступним для організацій із обмеженим бюджетом чи ресурсами.

3. Ризик упередженості при інтеграції

Хоча мультиканальний аналіз знижує ризик викривлень від одного джерела, існує ймовірність упередженості на етапі інтеграції даних. Наприклад, неправильна побудова вагових коефіцієнтів між джерелами або суб'єктивний вибір каналів може призвести до отримання спотворених результатів.

4. Тривалість процесу аналізу

Довгостроковий прогноз вимагає не лише збору та обробки даних, але й постійного моніторингу, оновлення моделей та врахування нових факторів. Це збільшує час виконання аналізу, що може бути критичним у швидкоплинних ринкових умовах.

5. Складність інтерпретації результатів

Різноспрямовані сигнали з різних каналів можуть створювати суперечливі висновки. Це вимагає розробки складних методів інтеграції й прийняття рішень, які не завжди забезпечують однозначність. Наприклад, одночасна позитивна динаміка в соціальних мережах і негативні економічні показники можуть створювати двозначні прогнози.

РОЗДІЛ 3. ЕМПІРИЧНИЙ АНАЛІЗ ТА РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

3.1 Вибір інструментів та технологій для проведення сентимент-аналізу

У контексті досягнення поставленої мети – розробки підходу до оцінки сентименту ринкових новин і його впливу на прогнозування цін активів – критично важливим є вибір інструментів, які забезпечують високу точність, швидкість і масштабованість аналізу. Ретельно обґрунтований вибір програмних засобів і технологій є необхідним для створення комплексного, гнучкого й ефективного рішення.

Проведення якісного сентимент-аналізу потребує використання ефективних програмних інструментів та технологій, здатних обробляти великі обсяги даних та забезпечувати надійні результати. У межах даного дослідження було прийнято рішення про використання низки засобів і технологій, зокрема мови програмування Python із відповідними бібліотеками, моделей LSTM та xml RoBERTa, словника VADER, а також WSL для гнучкої обробки даних у форматі CSV. Такий вибір обумовлений результатами аналітичного та теоретико-методичного розділів кваліфікаційної роботи та спрямований на досягнення основної дослідницької мети – виявлення та інтерпретації настроїв користувачів щодо певних активів (цінних паперів, криптовалют тощо).

Мова програмування Python та її бібліотеки

Python обрано як основний інструмент для розробки завдяки його універсальності, потужності в аналізі даних і зручності у використанні. Однією з ключових переваг є широка екосистема бібліотек, яка дозволяє ефективно вирішувати задачі на всіх етапах обробки даних — від збирання до моделювання. Його зрозумілий синтаксис і гнучкість створюють оптимальні умови для швидкої розробки і тестування моделей.

Бібліотеки для обробки та аналізу даних

Python надає різноманітні інструменти для роботи з даними, які адаптуються до потреб великих обсягів і різних форматів.

Pandas є основним інструментом для обробки табличних даних. Він забезпечує зручне зчитування, фільтрацію, групування й об'єднання наборів даних, що особливо корисно для етапу підготовки даних. Pandas дозволяє легко маніпулювати структурованими даними, підтримуючи роботу зі складними таблицями й забезпечуючи простоту інтеграції з іншими бібліотеками.

Dask розширює можливості Pandas, дозволяючи працювати з даними, обсяги яких перевищують оперативну пам'ять комп'ютера. Dask забезпечує паралельну обробку і оптимізує час виконання операцій, що є важливим для масштабних досліджень.

NumPy використовується для ефективних числових обчислень. Завдяки своїй архітектурі, яка побудована на масивах даних, NumPy оптимізує математичні й статистичні операції, зменшуючи час обробки великих масивів.

Matplotlib і Seaborn надають широкі можливості для візуалізації даних. Matplotlib забезпечує побудову графіків будь-якої складності, тоді як Seaborn додає інтуїтивно зрозумілі інструменти для створення привабливих і зрозумілих діаграм.

Бібліотеки для інтеграції з платформами соціальних мереж

Python також має потужні інструменти для інтеграції з API соціальних платформ, що важливо для збору й аналізу текстових даних у реальному часі.

Twikit — це спеціалізована бібліотека, яка забезпечує автоматизований збір даних із веб-ресурсів і API. Вона допомагає формувати готові набори даних для обробки, зменшуючи час і складність інтеграції з зовнішніми джерелами.

Tweepy є ідеальним інструментом для роботи з Twitter. Вона дозволяє отримувати доступ до твітів за певними параметрами, зберігати дані у зручному форматі й обробляти їх для подальшого аналізу.

Beautiful Soup — бібліотека для парсингу HTML і XML-документів. Вона забезпечує можливість зручного вилучення даних із веб-сторінок, які не надають API для доступу. Beautiful Soup автоматично обробляє структуру веб-сторінок, дозволяючи швидко знаходити й вилучати потрібну інформацію. Ця бібліотека є корисною для збору текстів із новинних порталів, блогів чи інших сайтів, які публікують матеріали, що стосуються досліджуваних тем.

Використання Twikit, Tweepy і Beautiful Soup у комплексі дозволяє забезпечити всебічне покриття джерел даних:

— Twikit автоматизує збір даних із веб-ресурсів, які надають структуровану інформацію через API.

— Tweepy концентрується на спеціалізованій роботі з Twitter, що є важливим джерелом текстів для аналізу настроїв.

— Beautiful Soup заповнює прогалини, обробляючи дані з ресурсів, які не надають прямого доступу через API.

Такий підхід забезпечує гнучкість і масштабованість у зборі даних, що є критично важливим для якісного сентимент-аналізу та моделювання впливу настроїв на фінансові ринки.

Бібліотеки для машинного навчання

Python має багатий набір бібліотек для машинного навчання, які дозволяють розробляти моделі будь-якої складності.

Scikit-learn забезпечує основний набір інструментів для задач класифікації, регресії й кластеризації. Її простий інтерфейс дозволяє швидко тестувати класичні алгоритми, такі як логістична регресія чи метод опорних векторів, забезпечуючи високу швидкість і надійність.

TensorFlow та Keras використовуються для побудови й навчання складних нейронних мереж, включаючи LSTM і трансформери. TensorFlow є потужною платформою для розробки моделей глибокого навчання, тоді як Keras надає високорівневий API для швидкого прототипування.

PyTorch є гнучким інструментом для досліджень і розробки, що пропонує динамічний підхід до побудови графів обчислень. Завдяки простоті в роботі та активній підтримці спільноти, PyTorch часто використовується для розробки сучасних моделей.

NLTK (Natural Language Toolkit) і spaCy є базовими бібліотеками для обробки текстів. NLTK пропонує широкий спектр інструментів для роботи з текстами, включаючи токенізацію, розбір тексту й визначення частин мови. SpaCy, у свою чергу, забезпечує високу продуктивність і точність у задачах аналізу природної мови.

Використання моделі LSTM для прогнозування цін активів

Long Short-Term Memory (LSTM) є різновидом рекурентних нейронних мереж (RNN), розробленим для ефективної роботи з послідовними даними. Основна унікальність LSTM полягає в її здатності враховувати довгострокові залежності у часових рядах. Це досягається завдяки спеціальній архітектурі, яка дозволяє "пам'ятати" інформацію протягом тривалих проміжків часу.

LSTM складається з комірок пам’яті (memory cells), кожна з яких містить три основні механізми:

— Вхідний шлюз (input gate): визначає, яка частина нової інформації має бути збережена.

— Шлюз забуття (forget gate): регулює, яку частину попередньої інформації необхідно видалити.

— Вихідний шлюз (output gate): обирає, яка інформація має бути передана на наступний етап.

Ця багатошарова структура дозволяє LSTM уникати проблеми "зникнення градієнта", що є поширеною у звичайних RNN, та ефективно працювати з довгими часовими послідовностями.

Причини вибору LSTM

Обробка часових рядів у фінансовому контексті є складним завданням, оскільки ціни активів часто демонструють складні й нелінійні залежності, що змінюються у

часі. Наприклад, важливі ринкові новини, такі як оголошення про зміну облікової ставки чи події політичного характеру, можуть впливати на ціни активів не тільки в момент їхньої публікації, але й протягом тривалого періоду. Ці залежності потребують врахування історичної інформації, що охоплює як короткотермінові реакції, так і довгострокові тренди. У цьому контексті модель LSTM є ідеальним інструментом, оскільки вона здатна зберігати й аналізувати історичні дані, використовуючи їх для точнішого прогнозування майбутніх значень. Завдяки своїй архітектурі LSTM може ефективно "запам'ятовувати" важливу інформацію, яка впливає на ринкові тенденції, та "забувати" другорядні або застарілі дані.

Ключовою перевагою LSTM є механізми пам'яті, реалізовані через спеціальні комірки пам’яті та шлюзи. Шлюзи регулюють потік інформації на різних етапах її обробки: шлюз забуття дозволяє видаляти непотрібну інформацію, вхідний шлюз вирішує, яку частину нових даних зберегти, а вихідний шлюз визначає, яка інформація буде передана на наступний етап. Така структура забезпечує збереження довготривалих залежностей, що є критично важливим для фінансових задач. Традиційні рекурентні нейронні мережі (RNN) часто стикаються з проблемою зникнення градієнта, через що їхня здатність враховувати тривалі часові залежності є обмеженою. LSTM успішно долає ці недоліки, дозволяючи зберігати релевантну інформацію протягом тривалих проміжків часу.

Гнучкість і універсальність LSTM є ще однією важливою характеристикою, яка робить її придатною для широкого спектру задач обробки часових рядів. Модель може бути адаптована для роботи з різними типами даних, включаючи числові показники, такі як ціни чи обсяги фінансових інструментів, і текстові дані, що відображають настрої ринку. У фінансовій сфері це означає можливість комбінованого використання кількісних та якісних даних, що значно підвищує точність аналізу й прогнозування. Наприклад, інтеграція цінових показників з результатами сентимент-аналізу новин дозволяє створити більш повну картину ринкових тенденцій.

Стійкість до шуму є ще однією суттєвою перевагою LSTM, особливо у випадках роботи з фінансовими даними, які часто містять значну кількість випадкових коливань. Шум у даних може спричинити неправильні висновки при використанні традиційних статистичних методів, таких як ARIMA або регресія. Натомість LSTM здатна фокусуватися на суттєвих трендах і залежностях, ігноруючи дрібні й несистематичні відхилення. Це дозволяє отримувати прогнози, які є більш точними та надійними, особливо в умовах високої волатильності ринку.

Таким чином, LSTM поєднує в собі здатність враховувати складні залежності у часових рядах, механізми для роботи з довгостроковою пам'яттю, універсальність у застосуванні до різних типів даних і стійкість до шуму. Усе це робить її одним із найкращих інструментів для задач прогнозування в умовах фінансових ринків. Модель дозволяє використовувати як історичні дані про ціни активів, так і результати аналізу настроїв, що забезпечує комплексний підхід до вивчення ринкової динаміки та прийняття інформованих рішень.

Роль LSTM у дослідженні

У межах цього дослідження модель LSTM використовується для прогнозування динаміки цін фінансових активів, таких як акції, криптовалюти чи облігації. Вона працює у зв'язці з результатами сентимент-аналізу, який дозволяє оцінити емоційне забарвлення ринкових новин.

Прогнозування відбувається у декілька етапів:

— Попередня обробка даних. Дані очищуються від шумів, нормалізуються та формуються у вигляді послідовностей, які подаються на вхід моделі.

— Навчання моделі. LSTM навчається на історичних даних про ціни активів, враховуючи вплив новинних подій і ринкових тенденцій. Навчання включає оптимізацію ваг нейронів для мінімізації похибки прогнозування.

— Прогнозування. На основі аналізу історичних тенденцій і нових даних модель генерує прогнози майбутніх цін активів.

— Інтеграція з сентимент-аналізом. Результати прогнозування комбінуються з оцінками емоційного забарвлення ринкових новин, щоб створити більш комплексний погляд на стан ринку.

Модель xml RoBERTa для аналізу текстів

Модель xml RoBERTa є однією з найсучасніших реалізацій архітектури трансформерів, яка застосовується для обробки природної мови (NLP). Вона побудована на основі моделі RoBERTa (Robustly Optimized BERT Pretraining Approach), що є вдосконаленням моделі BERT, і розроблена для більш ефективного навчання та точнішого розуміння текстів різного рівня складності. Xml RoBERTa інтегрує додаткові оптимізації, які роблять її особливо корисною для аналізу текстів у різних форматах і доменах, зокрема у фінансових даних.

Сутність і переваги xml RoBERTa

Однією з ключових особливостей моделі xml RoBERTa є її здатність до глибокого контекстного аналізу тексту. Архітектура трансформерів, на основі якої побудована модель, дозволяє враховувати не тільки значення кожного слова, але й його оточення в межах усього тексту. Це досягається завдяки використанню механізму self-attention, який оцінює зв'язок між словами незалежно від їхньої віддаленості у тексті. У фінансових текстах, де ключове значення можуть мати навіть віддалені частини повідомлення, така здатність є критичною. Наприклад, заголовок новини про "підвищення прибутку" може бути уточнений у тексті фразою "нижче очікувань", що повністю змінює тональність повідомлення.

Xml RoBERTa демонструє високу точність у задачах класифікації текстів і визначення настроїв завдяки своїй вдосконаленій архітектурі. На відміну від базової моделі BERT, яка використовує статичні параметри для обробки текстів, xml RoBERTa дозволяє обробляти текстові дані у більших обсягах і з урахуванням багатомовних контекстів. Це особливо важливо для фінансових аналізів, де тексти можуть бути складними та містити специфічну термінологію.

Модель також підтримує можливість донавчання (fine-tuning) на специфічних наборах даних. Це означає, що xml RoBERTa може бути адаптована для роботи із текстами певного домену, наприклад, новинами про ринок криптовалют або коментарями в соціальних мережах. Донавчання дозволяє моделі не лише загально розпізнавати настрої, а й враховувати специфічні лексеми чи шаблони, які часто зустрічаються в аналізованих текстах. У результаті це значно підвищує точність класифікації тональності.

Ще однією перевагою xml RoBERTa є її гнучкість. Вона підтримує обробку текстів різного формату, що робить її універсальним інструментом для роботи із новинами, постами в соціальних мережах чи коментарями. У фінансовій сфері ця здатність є особливо важливою, оскільки ринкові настрої часто формуються на основі різнорідних джерел інформації.

Роль xml RoBERTa у дослідженні

У цьому дослідженні xml RoBERTa використовується як ключовий інструмент для аналізу текстових даних з метою оцінки настроїв у ринкових повідомленнях. Її завдання полягає в аналізі новин, постів і коментарів, пов’язаних із фінансовими ринками, для визначення емоційного забарвлення кожного повідомлення. Результати цього аналізу є основою для виявлення зв’язків між настроями ринку та динамікою цін активів.

Процес роботи з моделлю xml RoBERTa включає кілька етапів:

— попередня обробка текстів: тексти очищуються від зайвих символів, токенізуються та формуються у формат, який підтримує модель.

— донавчання моделі: на специфічному наборі даних (наприклад, новинах із фінансових порталів) модель адаптується для врахування специфіки фінансової лексики.

— класифікація тональності: модель визначає, чи є текст позитивним, негативним або нейтральним, враховуючи весь контекст повідомлення.

— інтеграція результатів: результати аналізу використовуються для побудови моделей прогнозування динаміки ринку, у поєднанні з іншими джерелами даних (наприклад, історичними ціновими рядами).

Переваги використання xml RoBERTa

Поєднання глибокого розуміння контексту, високої точності й можливості адаптації робить xml RoBERTa незамінним інструментом для аналізу текстів у фінансовій сфері. Вона забезпечує:

— Точність у складних випадках, аналізує тексти з багатозначними формулюваннями, враховуючи всі деталі.

— Універсальність, працює з різними джерелами текстових даних, що дозволяє отримати повну картину ринкових настроїв.

— Масштабованість, обробляє великі обсяги текстів завдяки оптимізованій архітектурі трансформерів.

Словник VADER для аналізу сентименту

VADER (Valence Aware Dictionary and sEntiment Reasoner) є одним із найпопулярніших інструментів для лексичного аналізу настроїв текстів. Його розроблено спеціально для роботи із сучасними текстовими даними, зокрема тими, які зустрічаються в соціальних мережах. Ця бібліотека базується на словниковому підході, що дозволяє швидко й ефективно класифікувати тональність текстів як позитивну, негативну або нейтральну, враховуючи специфіку сучасної мови спілкування.

Причини вибору VADER

Однією з ключових причин використання VADER є його простота та швидкість. На відміну від нейронних мереж чи трансформерів, лексичний підхід не потребує великих обчислювальних ресурсів. VADER використовує заздалегідь визначений словник емоційно забарвлених слів, фраз і символів, що дозволяє обробляти великі обсяги текстів у режимі реального часу. Це особливо важливо для задач, де необхідно

швидко отримати базову оцінку тональності, наприклад, при аналізі соціальних мереж чи новинних потоків.

Ще однією значною перевагою є адаптованість VADER до сучасної мови. Словник розроблений з урахуванням:

— сленгових виразів: наприклад, слова на кшталт "awesome" або "meh" автоматично отримують відповідну емоційну оцінку.

— скорочень і абревіатур: такі як "LOL" чи "OMG" також враховуються під час аналізу.

— емодзі та пунктуації: наявність знаків оклику, питань чи смайликів у тексті впливає на тональність. Наприклад, фраза "Great!!!" буде мати сильніше позитивне значення, ніж просто "Great".

Застосування VADER

У рамках дослідження VADER використовується для попередньої оцінки тональності текстів. Основною його роллю є базова класифікація текстів перед їхньою передачею на більш складні моделі, такі як xml RoBERTa чи LSTM. Наприклад, VADER може швидко оцінити, чи є текст позитивним, негативним чи нейтральним, що дозволяє ефективно фільтрувати дані або використовувати отримані результати для валідації результатів нейронних моделей.

VADER також ефективно інтегрується з Python-бібліотеками, такими як Pandas, що дозволяє легко застосовувати його до великих текстових корпусів. Завдяки цьому він є зручним і надійним інструментом для швидкого аналізу текстів у різних форматах і з різних джерел.

Обробка даних за допомогою WSL

WSL (Windows Subsystem for Linux) забезпечує зручне середовище для роботи з великими масивами даних у форматі CSV, що часто використовуються для зберігання текстів, новин і фінансових даних. Використання WSL дозволяє уникати обмежень стандартних Windows-застосунків, таких як Excel, які можуть некоректно обробляти великі файли або змінювати їх структуру.

Основні інструменти для обробки даних у WSL включають:

— Csvkit, набір утиліт для роботи з CSV-файлами, який дозволяє перевіряти їх структуру, сортувати, фільтрувати й конвертувати в інші формати;

— Vim, текстовий редактор, який є зручним для швидкого перегляду й редагування великих файлів безпосередньо в терміналі.

Також були використані команди різноманітні shell-терміналу, такі grep, awk чи sed для обробки текстових даних без потреби у спеціалізованих додатках.

Переваги та роль WSL

Windows Subsystem for Linux (WSL) забезпечує потужні можливості для роботи з великими наборами даних у форматі CSV, об'єднуючи переваги Linux-інструментів із доступністю Windows-середовища. Серед ключових переваг WSL виділяється його висока швидкість обробки даних. Інструменти Linux, оптимізовані для великих файлів, дозволяють виконувати складні пакетні операції, такі як фільтрація, сортування й об'єднання, значно швидше, ніж традиційні Windows-додатки.

Гнучкість WSL полягає в можливості поєднувати кілька утиліт, таких як grep (для пошуку текстових шаблонів), awk (для обробки даних) і csvkit (для роботи з файлами CSV), що дозволяє зручно виконувати навіть складні операції. Це забезпечує стабільність і надійність у роботі з великими наборами даних, знижуючи ризик пошкодження файлів чи втрати інформації, які часто виникають через обмеження стандартних Windows-засобів.

Інтеграція з Python є ще однією важливою перевагою WSL. Дані, попередньо оброблені за допомогою Linux-утиліт, легко передаються в Python-скрипти для

подальшого аналізу, моделювання чи візуалізації. Це робить WSL універсальним інструментом, який підвищує ефективність всього аналітичного процесу.

WSL також підтримує роботу з великими файлами через такі команди, як split (для поділу файлів) і cat (для їх об'єднання), що полегшує обробку текстових даних значного обсягу. Додаткова можливість створення Bash-скриптів для автоматизації типових завдань, таких як очищення, сортування та конвертація даних, забезпечує значну економію часу й ресурсів.

У межах дослідження WSL відіграє ключову роль на етапах попередньої обробки даних. Вона дозволяє структурувати та очищати інформацію, зберігаючи її у форматі, придатному для подальшого аналізу. Завдяки своїй швидкості, стабільності й інтеграційним можливостям WSL сприяє ефективному використанню великих текстових і фінансових даних, необхідних для побудови моделей прогнозування та аналізу настроїв ринкових повідомлень.

Середня абсолютна похибка (MAE)

Середня абсолютна похибка (Mean Absolute Error, MAE) — це одна з основних статистичних метрик, яка використовується для оцінки точності моделей прогнозування часових рядів. Вона вимірює середню абсолютну величину різниці між фактичними та прогнозованими значеннями.

Формула для обчислення MAE:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|\,; \tag{3.1}$$

де n — кількість спостережень;

y_i — фактичне значення тимчасового ряду в момент часу i;

$\hat{y}_i$ — прогнозоване значення тимчасового ряду в момент часу i.

MAE дозволяє оцінити середню величину помилки, яка є інтуїтивно зрозумілою та легко інтерпретується. Вона враховує лише абсолютні значення відхилень, що робить її нечутливою до великих помилок. Менше значення MAE вказує на вищу точність моделі. Ця метрика широко використовується завдяки своїй простоті,

особливо у випадках, коли потрібно отримати загальне уявлення про точність моделі без акценту на великі відхилення.

Середньоквадратична похибка (MSE)

Середньоквадратична похибка (Mean Squared Error, MSE) є ще одним важливим показником для оцінки ефективності моделей прогнозування часових рядів. Вона вимірює середню величину квадратів різниць між фактичними та прогнозованими значеннями.

Формула для обчислення MSE:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\,; \tag{3.2}$$

де n — кількість спостережень;

y_i — фактичне значення тимчасового ряду в момент часу i;

$\hat{y}_i$ — прогнозоване значення тимчасового ряду в момент часу i.

Менше значення MSE свідчить про вищу точність моделі, оскільки воно відображає менший рівень середньої помилки. MSE зазвичай використовується разом із такими метриками, як MAE та RMSE, для комплексної оцінки ефективності моделей.

Завдяки чутливості до великих відхилень, MSE дозволяє ідентифікувати проблемні аспекти прогнозування, які потребують покращення, і є важливою метрикою для оцінки моделей у задачах прогнозування часових рядів.

Корінь середньоквадратичної помилки (RMSE)

Корінь середньоквадратичної помилки (Root Mean Squared Error, RMSE) є однією з найпоширеніших метрик для оцінки точності прогнозування часових рядів. Ця метрика визначає ступінь відхилення прогнозованих значень від реальних, обчислюючи квадратний корінь із середнього значення квадратів різниць між фактичними та прогнозованими значеннями.

Формула обчислення RMSE:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)}{n}}; \quad (3.3)$$

де n — кількість спостережень;

y_i — фактичне значення тимчасового ряду в момент часу i;

$\hat{y}_i$ — прогнозоване значення тимчасового ряду в момент часу i.

Менші значення RMSE свідчать про вищу точність моделі, оскільки вони вказують на менші розбіжності між прогнозованими та реальними значеннями. RMSE є особливо корисною метрикою для задач прогнозування, оскільки вона чутлива до великих похибок, що робить її ефективною для виявлення значних відхилень. Ця метрика зазвичай застосовується разом із іншими показниками, такими як MAE (Mean Absolute Error) і MAPE (Mean Absolute Percentage Error), для комплексної оцінки ефективності моделей прогнозування часових рядів.

Середній відсоток абсолютних помилок (MAPE)

Середній відсоток абсолютних помилок (Mean Absolute Percentage Error, MAPE) є універсальною метрикою, яка використовується для оцінки точності моделей прогнозування у часових рядах. Вона вимірює середнє значення відсотків абсолютних похибок між прогнозними й фактичними значеннями, що дозволяє легко інтерпретувати результати у відносних одиницях.

Формула обчислення MAPE:

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\%; \quad (3.4)$$

де n — загальна кількість спостережень;

y_i — фактичне значення тимчасового ряду в момент часу i;

$\hat{y}_i$ — прогнозоване значення тимчасового ряду в момент часу i.

MAPE надає чітке уявлення про те, наскільки ефективно модель здатна прогнозувати значення, порівнюючи відносні відхилення від реальних даних. Чим нижчий показник MAPE, тим точнішою є модель. Як і RMSE, MAPE є важливою

частиною набору метрик для оцінки продуктивності моделей і використовується разом з іншими показниками для комплексного аналізу ефективності прогнозування.

MAPE має особливу перевагу у тому, що результати представлені у відсотках, що спрощує їх інтерпретацію для практичного застосування у фінансових, економічних або інженерних задачах.

3.2 Проведення мультиканального сентимент-аналізу

У межах цього дослідження було прийнято рішення збирати та аналізувати текстові дані переважно з платформ X.com (раніше відомої як Twitter) і Reddit, оскільки саме ці ресурси є одними з найпопулярніших для обговорення криптовалют, що дає змогу одержувати репрезентативну вибірку публікацій із тематичних спільнот. Дані сентименту охоплюють період із 1 серпня 2022 року до 31 грудня 2022 року, причому такий часовий інтервал було зумовлено низкою технічних і ресурсних причин. Зокрема, платформи блокують надмірні спроби скрепінгу й установлюють високі ціни на API-доступ, що ускладнює безперервне вилучення великих обсягів даних; додаткові перешкоди також спричиняли перебої з електропостачанням і високі навантаження на персональний комп’ютер під час масового збору та первинної обробки інформації.

В результаті, завдяки скріптам, було зібрано дані і створено дві таблиці (Рисунок 3.1 та Рисунок 3.2), з сентиментом на платформах X та Reddit, пов'язаних з криптовалютою Bitcoin.

Рис. 3.1 — Зібрані дані X #BTC

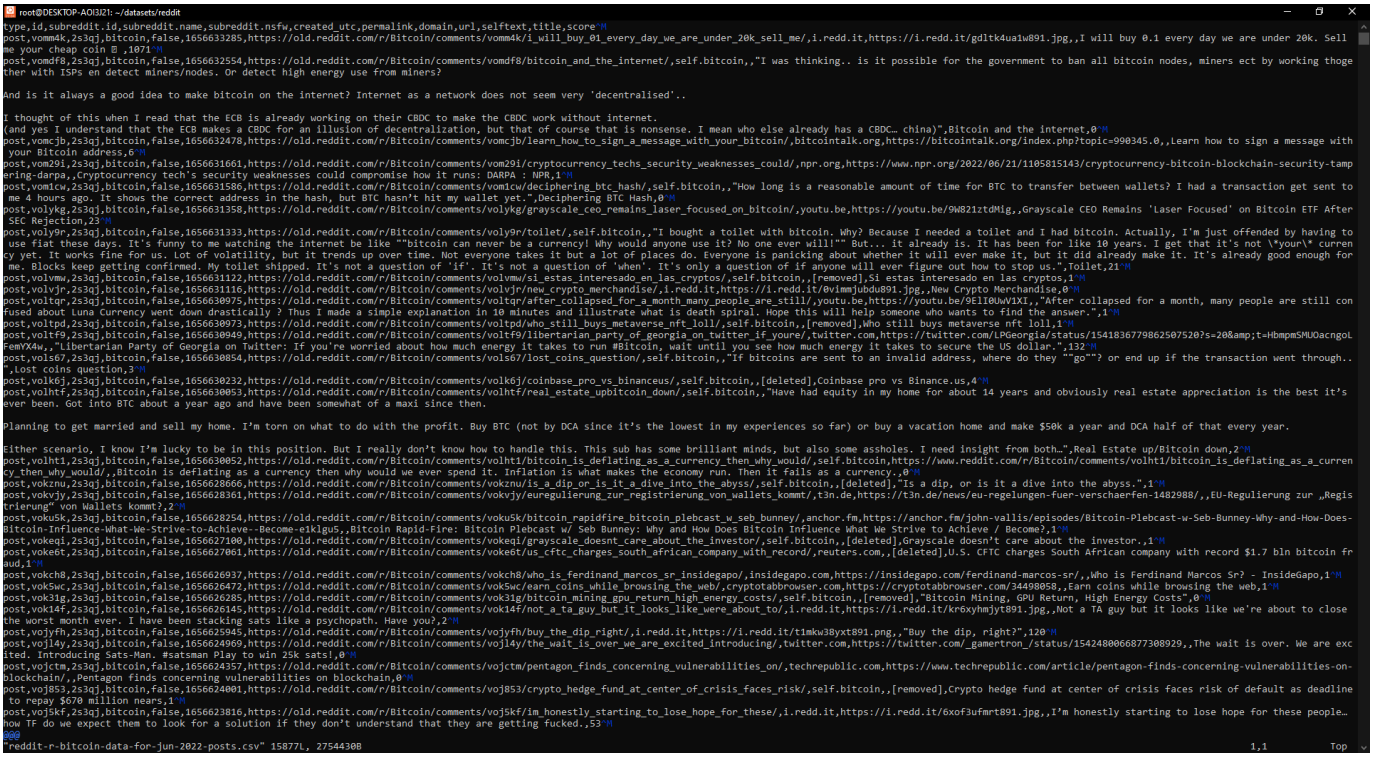


Рис. 3.2 — Зібрані дані Reddit/Bitcoin

Після завершення процесу збирання, початкові дані було належним чином відформатовано за допомогою утиліти csvkit, що дозволило вилучити надлишкову інформацію та виокремити релевантний діапазон. Після цього підготовлені дані було передано на оброблення моделям RoBERTa та VADER для подальшого проведення сентимент-аналізу. Результати обчислень (у вигляді нових CSV-файлів) наведено нижче на відповідних рисунках.

У межах цього дослідження було прийнято рішення здійснювати збір та аналіз текстових даних переважно із платформ X.com (раніше відомої як Twitter) і Reddit. Такий вибір зумовлений тим, що ці ресурси є одними з найпопулярніших для обговорення криптовалют, що дає змогу отримати репрезентативну вибірку публікацій із тематичних спільнот. Одержані дані щодо сентименту охоплюють період із 1 серпня 2022 року до 31 грудня 2022 року, що обумовлено низкою технічних і ресурсних чинників. Зокрема, блокування надмірних спроб скрепінгу та високі ціни на API-доступ ускладнюють безперервне вилучення значних обсягів даних із цих платформ. Крім того, перебої з електропостачанням та високе навантаження на персональний комп’ютер під час масового збору та первинної обробки інформації додатково обмежували розмір вибірки.

Незважаючи на зазначені труднощі, упроваджені скрипти дали змогу зібрати достатню кількість публікацій, що стосуються криптовалюти Bitcoin. У результаті було побудовано дві зведені таблиці (Рисунок 3.3 та Рисунок 3.4), які демонструють результати сентимент-аналізу в межах платформ X і Reddit. Ці таблиці містять числові оцінки емоційної спрямованості контенту та можуть бути використані для подальшого детального аналізу тенденцій у криптовалютному середовищі.

Загалом, отримані результати підтверджують можливість масштабованого підходу до дослідження динаміки суспільних настроїв навколо криптовалют. Запропонована методика в подальшому може бути поширена на більший часовий діапазон або інші соціальні платформи, що дозволить розширити розуміння патернів

поведінки криптоспільноти та впливу різних зовнішніх чинників на формування настроїв.

Рис 3.3 — Дані мережі X після проведення сентимент-аналізу

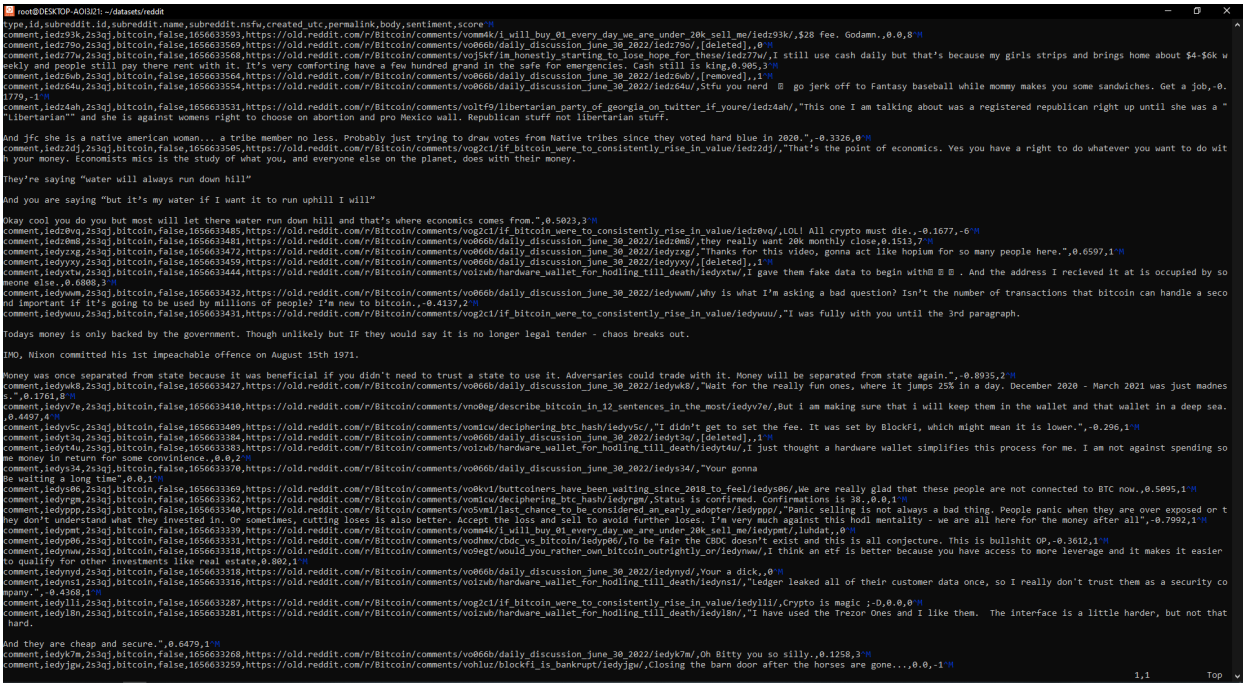


Рис 3.3 — Дані мережі Reddit після проведення сентимент-аналізу

3.3 Прогнозування ринкових тенденцій на основі результатів аналізу настроїв

Для прогнозування динаміки вартості криптоактивів, у тому числі Bitcoin, було використано історичний графік цін за період із 1 серпня 2022 року по 31 грудня 2022 року. Дані для аналізу отримано із сайту CoinMarketCap, що є одним із найвідоміших ресурсів для відстеження ринкових показників криптовалют. Візуалізацію історичних цін за вказаний період наведено нижче, на Рисунку 3.5.

Рис. 3.5 — Історична ціна криптовалюти BTC

Згаданий часовий інтервал відповідає тому ж періоду, у межах якого було здійснено збирання та аналіз текстових публікацій із соціальних мереж. Такий підхід надає можливість комплексного зіставлення результатів сентимент-аналізу з фактичними змінами ринкової вартості активу, що, у свою чергу, дозволяє виявити потенційні кореляційні зв’язки та прогнозувати подальші цінові тенденції з урахуванням соціального й інформаційного контексту.

У наступному етапі дослідження історичні дані щодо ціни Bitcoin та зібрані раніше показники сентименту було завантажено в нейронну мережу на основі LSTM. Модель розроблено таким чином, щоб поєднувати часові ряди ринкових показників із результатами мультиканального та одноканального сентимент-аналізу. Для цього

було введено ваговий коефіцієнт α=0.65, що визначає ступінь впливу сентименту на загальну прогнозну модель.

Задана архітектура LSTM забезпечує ефективне опрацювання послідовних даних завдяки механізмам «короткочасної» та «довготривалої» пам'яті, що дозволяє моделі враховувати як недавні, так і більш віддалені в часі коливання на ринку криптовалют. Після навчання нейронної мережі було отримано дві серії прогнозів:

— Мультиканальний (багатовимірний) прогноз, що інтегрує декілька джерел сентименту (X.com, Reddit) одночасно.

— Одноканальний (одновимірний) прогноз, у якому використовується лише один канал (наприклад, Reddit або виключно цінові дані).

Для оцінки точності передбачення застосовано низку загальноприйнятих метрик: MAE, RMSE, MSE, MAPE.

Отримані результати свідчать про задовільну якість прогнозування (Рисунок 3.6 та Рисунок 3.7), а також підкреслюють доцільність урахування результатів сентимент-аналізу для більш точної оцінки майбутніх коливань ціни криптовалют.

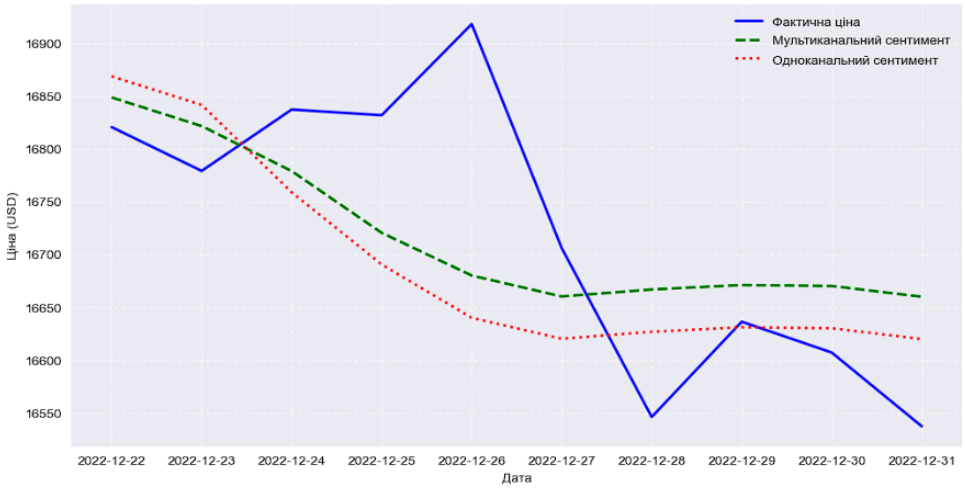


Рис. 3.6 – Графік прогнозу для криптовалюти BTC

```
Метрики ефективності:

Мультиканальний сентимент:
MAE: 86.49
MSE: 11186.82
RMSE: 105.77
MAPE: 0.52%

Одноканальний сентимент:
MAE: 88.54
MSE: 13082.00
RMSE: 114.38
MAPE: 0.53%
```

Рис. 3.6 – Метрики відповідності прогнозу для криптовалюти BTC

ВИСНОВОК

У процесі дослідження було здійснено комплексний аналіз можливостей мультиканального сентимент-аналізу та обґрунтовано його ефективність для прогнозування ринкових тенденцій криптовалют. Розроблено інтегровану концепцію, що передбачає поєднання даних із соціальних мереж, новинних ресурсів та форумів із використанням сучасних методів машинного навчання (LSTM, трансформери) та принципів об’єктно-орієнтованого програмування (ООП) для створення гнучкої програмної архітектури. Удосконалено методологію збору й попередньої обробки великих даних завдяки застосуванню автоматизованих скриптів та API для виявлення релевантних записів. Впроваджено алгоритми мультимодального аналізу, що поєднують текстові дані та метадані, дозволяючи оперативно оцінювати настрої користувачів у реальному часі. Застосування технологій глибокого навчання дало змогу виявляти приховані семантичні зв’язки в масивах даних, підвищуючи точність аналізу настроїв.

Практичний результат роботи полягає у створенні автоматизованої системи (прототипу програмного комплексу), здатної приймати великі обсяги даних із різних джерел, виконувати їхню стандартизовану обробку та формувати коротко- і середньострокові прогнози ринкових коливань. Запропоновано кількісну методику оцінювання впливу настроїв за допомогою кореляційного і регресійного аналізу, яка підтвердила статистично значущу залежність між рівнем «позитивного» чи «негативного» сентименту та змінами цінового тренду. Проведені експерименти засвідчили підвищення точності прогнозу.

Разом із тим, запропоновані рішення мають певні обмеження, пов’язані з необхідністю регулярного оновлення та перевірки моделей для збереження їхньої ефективності, а також із труднощами в точному визначенні іронічних чи контекстуально насичених повідомлень. Незважаючи на ці недоліки, розроблена методологія мультиканального сентимент-аналізу забезпечує комплексний підхід до раннього виявлення ринкових сигналів і може бути впроваджена в діяльність

приватних трейдерів, інвестиційних фондів, бірж і державних установ, а також адаптована для інших високо-вольатильних ринків.

СПИСОК ЛІТЕРАТУРИ

1. Jain, S., Johari, S., & Delhibabu, R. (2023). Analyzing Cryptocurrency trends using Tweet Sentiment Data and User Meta-Data. arXiv preprint arXiv:2307.15956.

2. Jeleskovic, V., & Mackay, S. (2023). *Intraday Trading Algorithm for Predicting Cryptocurrency Price Movements Using Twitter Big Data Analysis*. arXiv preprint arXiv:2401.00603.

3. Hearst, M. Direction-Based Text Interpretation as an Information Access Refinement // Text-Based Intelligent Systems / Paul Jacobs (Ed.) – Lawrence Erlbaum, 1992.

4. Pang B., Lee L., Vaithyanathan S. Thumbs up?. the ACL-02 conference, Not Known, 6 July 2002. Morristown, NJ, USA, 2002. URL: https://doi.org/10.3115/1118693.1118704.

5. Hochreiter S., Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997. Vol. 9, no. 8. P. 1735–1780. URL: https://doi.org/10.1162/neco.1997.9.8.1735

6. 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. 2019. Vol. 1. P. 4171–4186. URL: https://doi.org/10.18653/v1/N19-1423.

7. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. Improving Language Understanding by Generative Pre-training. *OpenAI*. 2018. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

8. 5. Sun C., Huang L., Qiu X. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019. Vol. 1. P. 380–385. URL: https://doi.org/10.18653/v1/N19-1035.

9. Esuli A., Sebastiani F. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). 2006. P. 417–422. URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/384_pdf.pdf.

10. Medhat W., Hassan A., Korashy H. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal. 2014. Vol. 5, no. 4. P. 1093–1113. URL: https://doi.org/10.1016/j.asej.2014.04.011.

11. Nakamoto S. Bitcoin: A Peer-to-Peer Electronic Cash System. URL: https://bitcoin.org/bitcoin.pdf.

12. An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends / Z. Zheng et al. 2017 IEEE International Congress on Big Data (BigData Congress). 2017. Vol. 6. P. 557–564. URL: https://doi.org/10.1109/BigDataCongress.2017.85.

13. Antonopoulos A. M. Mastering Bitcoin: Unlocking Digital Cryptocurrencies. O'Reilly Media, Inc., 2014. 298 p.

14. The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy / D. Garcia et al. Journal of The Royal Society Interface. 2014. Vol. 11, no. 99. P. 20140623. URL: https://doi.org/10.1098/rsif.2014.0623

15. López-Martín C., Benito Muela S., Arguedas R. Efficiency in cryptocurrency markets: new evidence. Eurasian Economic Review. 2021. Vol. 11, no. 3. P. 403–431. URL: https://doi.org/10.1007/s40822-021-00182-5

16. Koutmos D. Return and volatility spillovers among cryptocurrencies. Economics Letters. 2018. Vol. 173. P. 122–127. URL: https://doi.org/10.1016/j.econlet.2018.10.004

17. Almeida J., Gonçalves T. C. A Decade of Cryptocurrency Investment Literature: A Cluster-Based Systematic Analysis. International Journal of Financial Studies. 2023. Vol. 11, no. 2. P. 71. URL: https://doi.org/10.3390/ijfs11020071

18. Catalini C., Gans J. S. Some Simple Economics of the Blockchain. SSRN Electronic Journal. 2016. URL: https://doi.org/10.2139/ssrn.2874598

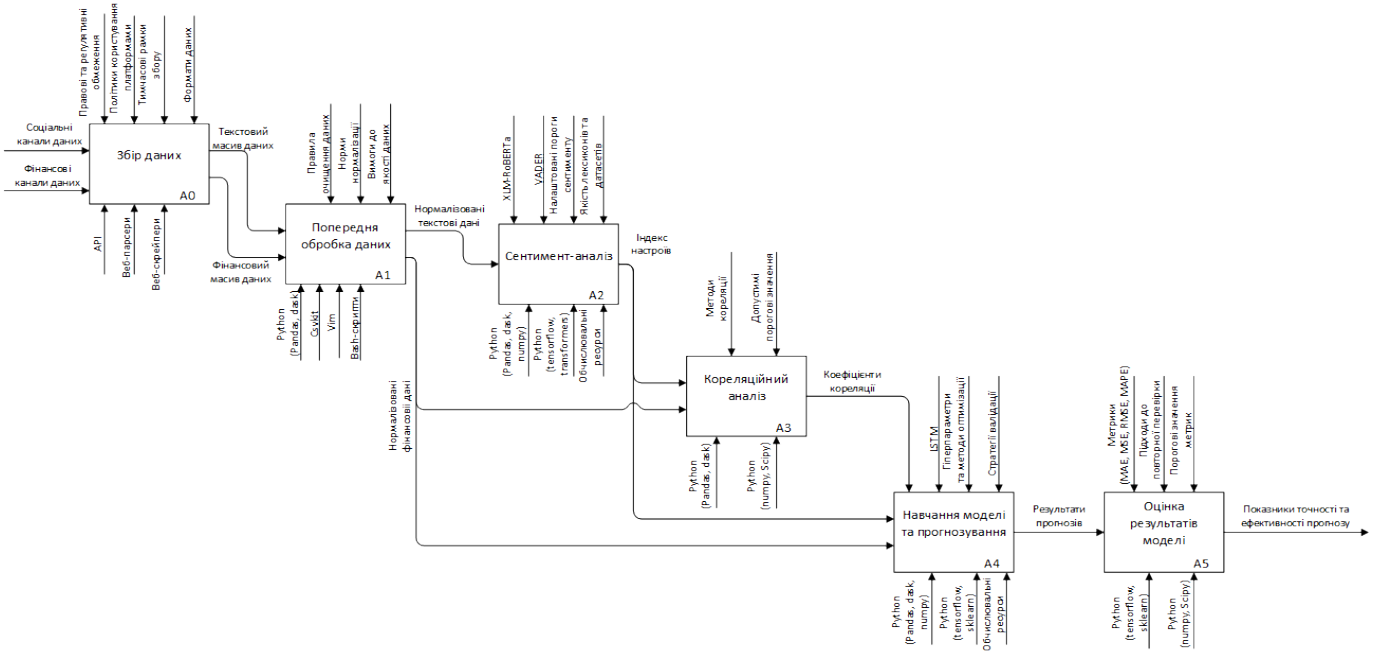
19. Barberis N., Mukherjee A., Wang B. Prospect Theory and Stock Returns: An Empirical Test. SSRN Electronic Journal. 2014. URL: https://doi.org/10.2139/ssrn.2528149

20. Buterin V. A Next-Generation Smart Contract and Decentralized Application Platform. Ethereum Whitepaper. 2013. URL: https://ethereum.org/content/whitepaper/whitepaper-pdf/Ethereum_Whitepaper_-_Buterin_2014.pdf.

ДОТАТКИ

Додаток А

IDEF0 діаграма методики сентимент-аналізу та прогнозування



Додаток Б

Скрапер для збору даних

```
import asyncio
import pandas as pd
from datetime import datetime, timedelta
from twikit import Client
import nest_asyncio
import json
import time
import random
import os
from typing import List, Dict, Optional, Set
from rich.console import Console
from rich.live import Live
from rich.table import Table
```

```
# Enable nest_asyncio for Jupyter
nest_asyncio.apply()

class EnhancedTwitterEtherScraper:
    def __init__(self):
        self.client = None
        self.base_delay = 2.0
        self.max_delay = 15.0
        self.current_delay = self.base_delay
        self.tweets_limit = 5000
        self.tweets_per_request = 20
        self.console = Console()
        self.start_time = None
        self.seen_cursors: Set[str] = set()
        self.seen_tweet_hashes: Set[str] = set()
        self.seen_tweet_ids: Set[str] = set()
        self._csv_lock = asyncio.Lock()  # Правильна ініціалізація блокування
        self.stats = {
            'requests_made': 0,
            'tweets_collected': 0,
            'unique_tweets': 0,
            'duplicates_skipped': 0,
            'duplicate_ids': 0,
            'errors': 0,
            'rate_limits': 0,
            'retries': 0
        }

    def calculate_tweet_hash(self, tweet_data: dict) -> str:
        """Створює ключ для порівняння твітів на основі тексту, дати та автора"""
        return f"{tweet_data['text']}|{tweet_data['created_at']}|{tweet_data['user_screen_name']}"
        self.csv_lock = asyncio.Lock()
```

```
def create_stats_table(self) -> Table:
    """Creates a rich table with current scraping statistics"""
    table = Table(title="Scraping Progress")
    table.add_column("Metric", style="cyan")
    table.add_column("Value", justify="right", style="green")

    elapsed = time.time() - self.start_time if self.start_time else 0
    tweets_per_minute = (self.stats['tweets_collected'] / elapsed * 60) if elapsed > 0 else 0

    table.add_row("Runtime", str(timedelta(seconds=int(elapsed))))
    table.add_row("Requests Made", str(self.stats['requests_made']))
    table.add_row("Tweets Collected", str(self.stats['tweets_collected']))
    table.add_row("Unique Tweets", str(self.stats['unique_tweets']))
    table.add_row("Duplicates Skipped", str(self.stats['duplicates_skipped']))
    table.add_row("Different IDs for Same Tweet", str(self.stats['duplicate_ids']))
    table.add_row("Tweets/Minute", f"{tweets_per_minute:.1f}")
    table.add_row("Rate Limits Hit", str(self.stats['rate_limits']))
    table.add_row("Errors", str(self.stats['errors']))
    table.add_row("Current Delay", f"{self.current_delay:.1f}s")

    return table

async def load_cookies(self) -> Dict[str, str]:
    """Enhanced cookie loading with validation"""
    try:
        with open('twitter_cookies.json', 'r') as f:
            cookies_list = json.load(f)

        required_cookies = {'auth_token', 'ct0'}
        cookie_dict = {}

        for cookie in cookies_list:
```

```
            if isinstance(cookie, dict) and cookie.get('name') in required_cookies:
                cookie_dict[cookie['name']] = cookie['value']

        if not all(cookie in cookie_dict for cookie in required_cookies):
            raise ValueError("Missing required cookies")

        return cookie_dict

    except Exception as e:
        self.console.print(f"[red]Error loading cookies: {str(e)}[/]")
        raise

async def initialize_csv(self, filename: str) -> None:
    """Initialize CSV and load existing tweet data"""
    if os.path.exists(filename):
        # Load existing tweets
        df = pd.read_csv(filename)
        existing_tweets = df.to_dict('records')
        for tweet in existing_tweets:
            try:
                tweet_hash = self.calculate_tweet_hash(tweet)
                self.seen_tweet_hashes.add(tweet_hash)
                self.seen_tweet_ids.add(str(tweet['id']))
            except Exception as e:
                self.console.print(f"[yellow]Warning: Could not process existing tweet: {str(e)}[/]")
                continue

        self.stats['unique_tweets'] = len(self.seen_tweet_hashes)
        self.console.print(f"[cyan]Loaded {len(self.seen_tweet_hashes)} existing unique tweets[/]")
    else:
        headers = [
            'id', 'created_at', 'text', 'retweet_count', 'favorite_count',
            'reply_count', 'quote_count', 'lang', 'user_id', 'user_name',
```

```
                'user_screen_name', 'user_followers', 'user_verified',
                'collected_at'
            ]
            async with self._csv_lock:  # Використання _csv_lock
                with open(filename, 'w', newline='', encoding='utf-8') as f:
                    writer = csv.writer(f)
                    writer.writerow(headers)

        async def update_csv(self, tweets_data: List[Dict], filename: str) -> tuple:
            """Optimized CSV update with batching"""
            async with self.csv_lock:
                try:
                    new_df = pd.DataFrame(tweets_data)
                    new_df['collected_at'] = datetime.now().isoformat()

                    if os.path.exists(filename):
                        existing_df = pd.read_csv(filename, usecols=['id'])
                        new_df = new_df[~new_df['id'].isin(existing_df['id'])]

                        if not new_df.empty:
                            new_df.to_csv(filename, mode='a', header=False, index=False, encoding='utf-8')
                    else:
                        new_df.to_csv(filename, index=False, encoding='utf-8')

                    return len(new_df), len(new_df) + len(existing_df) if 'existing_df' in locals() else
len(new_df)

                except Exception as e:
                    self.console.print(f"[red]Error updating CSV: {str(e)}[/]")
                    raise

        async def make_request(self, search_query: str, cursor: Optional[str] = None) -> Optional[dict]:
            """Enhanced request handling with better diagnostics"""
```

```
if cursor and cursor in self.seen_cursors:
    return None

max_retries = 3
retry_count = 0

while retry_count < max_retries:
    try:
        delay = random.uniform(2.0, 3.0)
        await asyncio.sleep(delay)

        self.stats['requests_made'] += 1
        self.console.print(f"[cyan]Making request with cursor: {cursor if cursor else 'initial'}[/]")
        result = await self.client.search_tweet(
            query=search_query,
            product='Latest',
            count=20,
            cursor=cursor
        )
        self.console.print(f"[green]Request successful[/]")

        if hasattr(result, '__len__'):
            tweet_count = len(result)
            self.console.print(f"[green]Retrieved {tweet_count} tweets[/]")

        if result and result.next_cursor:
            self.seen_cursors.add(result.next_cursor)

        return result

    except Exception as e:
        retry_count += 1
        self.stats['errors'] += 1
```

```
            if 'rate limit' in str(e).lower():
                await asyncio.sleep(60 * retry_count)  # Wait longer for rate limits
            else:
                await asyncio.sleep(5 * retry_count)

            if retry_count >= max_retries:
                self.console.print("[red]Max retries reached[/]")
                return None

    def create_search_query(self, start_date: str, end_date: str) -> str:
        """Creates optimized search query for specific date range"""
        core_terms = [
            "ethereum",
            "ETH",
            "$ETH"
        ]

        filters = [
            "-is:retweet",
            "lang:en",
            "min_faves:5"
        ]

        # Додаємо точний час для кращого охоплення
        start_time = f"{start_date}T00:00:00Z"
        end_time = f"{end_date}T23:59:59Z"

        query = f"({' OR '.join(core_terms)})"
        query += f" since:{start_time} until:{end_time}"
        query += f" {' '.join(filters)}"

        self.console.print(f"[cyan]Search query:[/] {query}")
```

```
        return query.strip()

    async def scrape(self, start_date: str, end_date: str):
        """Main scraping function with improved control flow and diagnostics"""
        self.start_time = time.time()

        # Validate and fix dates if needed
        try:
            start = pd.to_datetime(start_date)
            end = pd.to_datetime(end_date)
            if start > end:
                start_date, end_date = end_date, start_date
                self.console.print("[yellow]Warning: Dates were in wrong order, swapped automatically[/]")
        except Exception as e:
            self.console.print(f"[red]Invalid dates: {str(e)}[/]")
            return None

        # Initialize client
        try:
            cookies = await self.load_cookies()
            self.client = Client(language='en-US')
            self.client.set_cookies(cookies)
            user_id = await self.client.user_id()
            self.console.print(f"[green]Successfully authenticated as user ID: {user_id}[/]")
        except Exception as e:
            self.console.print(f"[red]Failed to initialize: {str(e)}[/]")
            return None

        filename = f'ether_tweets_{start_date}_{end_date}.csv'
        await self.initialize_csv(filename)

        search_query = self.create_search_query(start_date, end_date)
```

```
    cursor = None
    empty_results_streak = 0
    max_empty_streak = 3

    with Live(self.create_stats_table(), refresh_per_second=1) as live:
        while self.stats['tweets_collected'] < self.tweets_limit:
            result = await self.make_request(search_query, cursor)

            if not result:
                empty_results_streak += 1
                if empty_results_streak >= max_empty_streak:
                    self.console.print("[yellow]No more results available[/]")
                    break
                continue

            tweets_processed = await self.process_tweets(result, filename)

            if tweets_processed > 0:
                empty_results_streak = 0
            else:
                empty_results_streak += 1

            cursor = result.next_cursor
            if not cursor:
                self.console.print("[yellow]Reached end of results[/]")
                break

            live.update(self.create_stats_table())

    self.print_final_stats(filename)
    return filename

async def process_tweets(self, tweets, filename: str) -> int:
```

```
"""Process and save tweets with simple duplicate detection"""
if not tweets:
    return 0

new_tweets = []
for tweet in tweets:
    try:
        if hasattr(tweet, 'id'):
            tweet_data = {
                'id': tweet.id,
                'created_at': tweet.created_at,
                'text': tweet.text,
                'retweet_count': getattr(tweet, 'retweet_count', 0),
                'favorite_count': getattr(tweet, 'favorite_count', 0),
                'reply_count': getattr(tweet, 'reply_count', 0),
                'quote_count': getattr(tweet, 'quote_count', 0),
                'lang': tweet.lang,
                'user_id': tweet.user.id,
                'user_name': tweet.user.name,
                'user_screen_name': tweet.user.screen_name,
                'user_followers': tweet.user.followers_count,
                'user_verified': tweet.user.verified
            }

            # Спрощена перевірка на дублікати
            tweet_key = self.calculate_tweet_hash(tweet_data)

            if tweet_key in self.seen_tweet_hashes:
                self.stats['duplicates_skipped'] += 1
                # Виводимо інформацію про знайдений дублікат для відлагодження
                self.console.print(f"[yellow]Skipping duplicate tweet:[/]")
                self.console.print(f"Text: {tweet_data['text'][:100]}...")
                self.console.print(f"Author: {tweet_data['user_screen_name']}")
```

```
                    self.console.print(f"Date: {tweet_data['created_at']}")
                    continue

                self.seen_tweet_hashes.add(tweet_key)
                new_tweets.append(tweet_data)

            except AttributeError as e:
                self.console.print(f"[yellow]Warning: Skipping malformed tweet: {str(e)}[/]")
                continue

        if new_tweets:
            try:
                async with self._csv_lock:
                    new_df = pd.DataFrame(new_tweets)
                    new_df['collected_at'] = datetime.now().isoformat()
                    new_df.to_csv(filename, mode='a', header=False, index=False, encoding='utf-8')

                self.stats['tweets_collected'] += len(new_tweets)
                self.stats['unique_tweets'] = len(self.seen_tweet_hashes)
                return len(new_tweets)
            except Exception as e:
                self.console.print(f"[red]Error saving tweets: {str(e)}[/]")
                return 0

        return 0

    def print_final_stats(self, filename: str):
        """Print detailed final statistics"""
        elapsed = time.time() - self.start_time
        tweets_per_minute = (self.stats['tweets_collected'] / elapsed * 60) if elapsed > 0 else 0

        self.console.print("\n[bold green]Scraping completed![/]")
        self.console.print(f"Total runtime: {str(timedelta(seconds=int(elapsed)))}")
```

```
        self.console.print(f"Total tweets collected: {self.stats['tweets_collected']}")
        self.console.print(f"Unique tweets: {self.stats['unique_tweets']}")
        self.console.print(f"Average rate: {tweets_per_minute:.1f} tweets/minute")
        self.console.print(f"Results saved to: {filename}")
        success_rate = ((self.stats['requests_made'] - self.stats['errors']) /
                        self.stats['requests_made'] * 100 if self.stats['requests_made'] > 0 else 0)
        self.console.print(f"Success rate: {success_rate:.1f}%")

# Function for running in Jupyter Notebook
async def run_enhanced_scraper(start_date: str, end_date: str):
    scraper = EnhancedTwitterEtherScraper()
    return await scraper.scrape(start_date, end_date)
```

Додаток В

Модель сентимент-аналізу з використанням RoBERTa та VADER

```
import pandas as pd
import numpy as np
from transformers import AutoTokenizer, AutoModelForSequenceClassification
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
import torch
from datetime import datetime
from tqdm import tqdm

class SentimentAnalyzer:
    def __init__(self):
        # Ініціалізація RoBERTa
        self.tokenizer = AutoTokenizer.from_pretrained("cardiffnlp/twitter-roberta-base-sentiment")
```

```
            self.model = AutoModelForSequenceClassification.from_pretrained("cardiffnlp/twitter-roberta-base-sentiment")

            # Ініціалізація VADER
            self.vader = SentimentIntensityAnalyzer()

        def get_roberta_sentiment(self, text):
            try:
                inputs = self.tokenizer(str(text), return_tensors="pt", truncation=True, max_length=512)
                outputs = self.model(**inputs)
                probabilities = torch.nn.functional.softmax(outputs.logits, dim=1)
                sentiment_score = probabilities[0].tolist()
                # Повертаємо нормалізований скор від -1 до 1
                return (sentiment_score[2] - sentiment_score[0])
            except Exception as e:
                print(f"Помилка при аналізі тексту з RoBERTa: {e}")
                return 0

        def get_vader_sentiment(self, text):
            try:
                return self.vader.polarity_scores(str(text))['compound']
            except Exception as e:
                print(f"Помилка при аналізі тексту з VADER: {e}")
                return 0

        def analyze_data(self, price_data_path, text_data_path, date_col, price_col, text_col):
            """
            Аналізує дані з двох CSV файлів: один з цінами, інший з текстовими даними
```

```
Parameters:
price_data_path (str): шлях до CSV файлу з цінами
text_data_path (str): шлях до CSV файлу з текстовими даними
date_col (str): назва колонки з датами
price_col (str): назва колонки з цінами
text_col (str): назва колонки з текстом для аналізу
"""
try:
    # Завантаження даних
    price_df = pd.read_csv(price_data_path, parse_dates=[date_col])
    text_df = pd.read_csv(text_data_path, parse_dates=[date_col])

    # Об'єднання даних по даті
    df = pd.merge(price_df, text_df, on=date_col, how='inner')

    results = []

    for index, row in tqdm(df.iterrows(), total=len(df)):
        date = row[date_col]
        price = row[price_col]
        text = row[text_col]

        # Розрахунок зміни ціни у відсотках
        if index > 0:
            price_change = ((price - df.iloc[index-1][price_col]) /
                            df.iloc[index-1][price_col]) * 100
        else:
            price_change = 0

        # Аналіз тексту
        vader_sentiment = self.get_vader_sentiment(text)
        roberta_sentiment = self.get_roberta_sentiment(text)
```

```
            # Комбінований сентимент (середнє значення)
            combined_sentiment = np.mean([vader_sentiment, roberta_sentiment])

            results.append({
                'Дата': date,
                'Фактична ціна': price,
                'Зміна ціни (у %)': round(price_change, 2),
                'Сентимент (X)': round(roberta_sentiment, 3),
                'Сентимент (Reddit)': round(vader_sentiment, 3),
                'Сентимент (комбінований)': round(combined_sentiment, 3)
            })

        # Створення результуючого DataFrame
        results_df = pd.DataFrame(results)

        # Сортування за датою
        results_df = results_df.sort_values('Дата')

        return results_df

    except Exception as e:
        print(f"Помилка при обробці даних: {e}")
        return None

def save_results(results_df, output_path):
    """
    Зберігає результати аналізу у CSV файл
    """
    try:
        results_df.to_csv(output_path, index=False, encoding='utf-8')
        print(f"Результати збережено у файл: {output_path}")
    except Exception as e:
```

```
        print(f"Помилка при збереженні результатів: {e}")

# Приклад використання
if __name__ == "__main__":
    # Шляхи до файлів
    price_data_path = "price_data.csv"
    text_data_path = "text_data.csv"
    output_path = "sentiment_analysis_results.csv"

    # Назви колонок
    date_column = "date"
    price_column = "price"
    text_column = "text"

    # Ініціалізація аналізатора
    analyzer = SentimentAnalyzer()

    # Аналіз даних
    results = analyzer.analyze_data(
        price_data_path=price_data_path,
        text_data_path=text_data_path,
        date_col=date_column,
        price_col=price_column,
        text_col=text_column
    )

    if results is not None:
        # Виведення перших рядків результатів
        print("\nПерші рядки результатів:")
        print(results.head().to_string())

        # Збереження результатів
        save_results(results, output_path)
```

Додаток Г

Код для проведення кореляційного аналізу критерієм Пірсона

```
import pandas as pd
import numpy as np

# 1) ЗАВАНТАЖЕННЯ CSV-ФАЙЛУ
# Припустимо, що у поточній директорії є файл "btc_data.csv"
# зі стовпцями Date, Price, Change%, Sentiment_X, Sentiment_Reddit (тощо).
df = pd.read_csv("btc_data.csv")

# Приклад структури файлу (умовна):
# Date,Price,Change%,Sentiment_X,Sentiment_Reddit
# 2022-07-01,19240,3.32%,0.45,0.40
# 2022-07-02,19275,0.18%,0.30,0.28
# ...

# 2) ПЕРЕВЕДЕННЯ КОЛОНКИ DATE У ТИП DATETIME (якщо потрібно)
df['Date'] = pd.to_datetime(df['Date'], errors='coerce')

# 3) ПАРСИНГ ВІДСОТКА ЗМІНИ ЦІНИ (Change%)
# Якщо у стовпці "Change%" є значення з символом %, наприклад "3.32%",
# видалимо '%' і перетворимо на float. 3.32% → 3.32 → 0.0332.
def parse_change_percent(value):
    return float(value.replace('%', '')) / 100.0

df['Change%'] = df['Change%'].apply(parse_change_percent)

# 4) ЯКЩО НЕМАЄ РЕАЛЬНОГО СЕНТИМЕНТУ — ЗГЕНЕРУЄМО ШТУЧНИЙ (демо-варіант)
# Якщо у вас уже є колонки Sentiment_X і Sentiment_Reddit - пропустіть цей крок.
if 'Sentiment_X' not in df.columns:
    np.random.seed(42)
```

```
    # Наприклад, створюємо лінійну залежність від Change% + випадковий шум
    df['Sentiment_X'] = df['Change%'] * 10 + np.random.normal(0, 0.5, len(df))
    # нормалізуємо в діапазон [-1; 1]
    df['Sentiment_X'] = df['Sentiment_X'].clip(-1, 1)

if 'Sentiment_Reddit' not in df.columns:
    np.random.seed(24)
    df['Sentiment_Reddit'] = df['Change%'] * 8 + np.random.normal(0, 0.6, len(df))
    df['Sentiment_Reddit'] = df['Sentiment_Reddit'].clip(-1, 1)

# 5) ОБЧИСЛЕННЯ КОМБІНОВАНОГО СЕНТИМЕНТУ (як середнє двох)
df['Sentiment_Combined'] = (df['Sentiment_X'] + df['Sentiment_Reddit']) / 2

# 6) ФОРМУВАННЯ ФІНАЛЬНИХ КОЛОНОК З ВІДПОВІДНИМИ НАЗВАМИ УКРАЇНСЬКОЮ
# Створимо нову DataFrame для виводу результату, щоб точно контролювати порядок колонок.
result_df = pd.DataFrame()
result_df['Дата'] = df['Date']
result_df['Фактична ціна'] = df['Price']
# Зміна ціни (у %) - виведемо саме у відсотках, тобто помножимо на 100
result_df['Зміна ціни (у %)'] = df['Change%'] * 100
result_df['Сентимент (X)'] = df['Sentiment_X']
result_df['Сентимент (Reddit)'] = df['Sentiment_Reddit']
result_df['Сентимент (комбінований)'] = df['Sentiment_Combined']

# 7) КОРЕЛЯЦІЙНИЙ АНАЛІЗ (МЕТОДОМ ПІРСОНА)
# Для коректної кореляції беремо лише числові стовпці з result_df:
corr_cols = [
    'Фактична ціна',
    'Зміна ціни (у %)',
    'Сентимент (X)',
    'Сентимент (Reddit)',
```

```
    'Сентимент (комбінований)'
]
corr_matrix = result_df[corr_cols].corr(method='pearson')

# 8) ВИВІД РЕЗУЛЬТАТІВ
# 8.1) Виводимо перші 5 рядків отриманої таблиці (або всю, якщо невелика)
print("Таблиця з усіма потрібними стовпцями:")
print(result_df.head())

# 8.2) Виводимо кореляційну матрицю
print("\nКореляційна матриця (метод Пірсона):")
print(corr_matrix)
```

Додаток Ґ

Модель прогнозування ринкових цін

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense, Dropout
from tensorflow.keras.callbacks import EarlyStopping
from tensorflow.keras.optimizers import Adam
from sklearn.metrics import mean_absolute_error, mean_squared_error
import seaborn as sns

class BitcoinPricePredictor:
    def __init__(self, look_back=10, features=['Close', 'Volume', 'High', 'Low']):
        self.look_back = look_back
        self.features = features
        self.scalers = {}
        self.model = None

    def preprocess_data(self, data):
        """
        Попередня обробка даних з використанням декількох функцій
        """
        processed_data = np.zeros((len(data), len(self.features)))

        for i, feature in enumerate(self.features):
            # Створення окремого скейлера для кожної функції
            self.scalers[feature] = MinMaxScaler(feature_range=(0, 1))
            processed_data[:, i] = self.scalers[feature].fit_transform(
                data[feature].values.reshape(-1, 1)
```

```
        ).ravel()

    return processed_data

def create_sequences(self, data):
    """
    Створення послідовностей для LSTM з множинними функціями
    """
    X, y = [], []
    for i in range(len(data) - self.look_back):
        X.append(data[i:(i + self.look_back)])
        y.append(data[i + self.look_back, 0])  # Прогнозуємо лише ціну закриття
    return np.array(X), np.array(y)

def build_model(self):
    """
    Створення вдосконаленої архітектури LSTM
    """
    model = Sequential([
        LSTM(100, input_shape=(self.look_back, len(self.features)),
             return_sequences=True),
        Dropout(0.2),
        LSTM(50, return_sequences=False),
        Dropout(0.2),
        Dense(25, activation='relu'),
        Dense(1)
    ])

    optimizer = Adam(learning_rate=0.001)
    model.compile(optimizer=optimizer, loss='mse', metrics=['mae'])
    self.model = model
    return model
```

```
def train(self, train_X, train_y, validation_split=0.1, epochs=50,
          batch_size=32, verbose=1):
    """
    Навчання моделі з раннім зупиненням
    """
    early_stopping = EarlyStopping(
        monitor='val_loss',
        patience=10,
        restore_best_weights=True
    )

    history = self.model.fit(
        train_X, train_y,
        validation_split=validation_split,
        epochs=epochs,
        batch_size=batch_size,
        callbacks=[early_stopping],
        verbose=verbose
    )
    return history

def predict(self, X):
    """
    Прогнозування цін
    """
    return self.model.predict(X)

def inverse_transform_predictions(self, predictions):
    """
    Зворотне перетворення прогнозованих значень
    """
    return self.scalers['Close'].inverse_transform(predictions.reshape(-1, 1))
```

```
def calculate_metrics(self, y_true, y_pred):
    """
    Розрахунок метрик ефективності
    """
    mae = mean_absolute_error(y_true, y_pred)
    mse = mean_squared_error(y_true, y_pred)
    rmse = np.sqrt(mse)
    mape = np.mean(np.abs((y_true - y_pred) / y_true)) * 100

    return {
        'MAE': mae,
        'MSE': mse,
        'RMSE': rmse,
        'MAPE': mape
    }

def plot_results(self, dates, actual_values, predicted_values, title='Bitcoin Price Prediction'):
    """
    Візуалізація результатів з покращеним форматуванням
    """
    plt.figure(figsize=(15, 8))
    plt.plot(dates, actual_values, label='Фактична ціна', linewidth=2)
    plt.plot(dates, predicted_values, label='Прогнозована ціна', linewidth=2)

    plt.title(title, fontsize=16, pad=20)
    plt.xlabel('Дата', fontsize=12)
    plt.ylabel('Ціна (USD)', fontsize=12)
    plt.legend(fontsize=12)

    plt.grid(True, alpha=0.3)
    plt.xticks(rotation=45)

    # Додавання анотацій для останньої ціни
```

```
        last_actual = actual_values[-1]
        last_pred = predicted_values[-1]
        plt.annotate(f'Остання фактична: ${last_actual:,.2f}',
                     xy=(dates[-1], last_actual),
                     xytext=(10, 10),
                     textcoords='offset points')
        plt.annotate(f'Останній прогноз: ${last_pred:,.2f}',
                     xy=(dates[-1], last_pred),
                     xytext=(10, -10),
                     textcoords='offset points')

        plt.tight_layout()
        return plt

def main():
    # Завантаження даних
    data = pd.read_csv('BTC-USD.csv')

    # Ініціалізація предиктора
    predictor = BitcoinPricePredictor(look_back=20)

    # Попередня обробка даних
    processed_data = predictor.preprocess_data(data)

    # Розділення на тренувальний та тестовий набори
    train_size = int(len(processed_data) * 0.8)
    train_data = processed_data[:train_size]
    test_data = processed_data[train_size:]

    # Створення послідовностей
    train_X, train_y = predictor.create_sequences(train_data)
    test_X, test_y = predictor.create_sequences(test_data)
```

```
    # Побудова та навчання моделі
    predictor.build_model()
    history = predictor.train(train_X, train_y, epochs=50)

    # Прогнозування
    predictions = predictor.predict(test_X)

    # Зворотне перетворення
    actual_values = predictor.inverse_transform_predictions(test_y)
    predicted_values = predictor.inverse_transform_predictions(predictions)

    # Розрахунок метрик
    metrics = predictor.calculate_metrics(actual_values, predicted_values)
    print("\nМетрики ефективності:")
    for metric, value in metrics.items():
        print(f"{metric}: {value:.2f}")

    # Візуалізація результатів
    test_dates = data['Date'].values[train_size + predictor.look_back:]
    plt = predictor.plot_results(test_dates, actual_values, predicted_values)
    plt.show()

if __name__ == "__main__":
    main()
```

ДЕМОНСТРАЦІЙНІ МАТЕРІАЛИ (ПРЕЗЕНТАЦІЯ)

Міністерство освіти і науки України
Державний університет інформаційно-телекомунікаційних технологій
Навчально-науковий інститут телекомунікацій
Кафедра системного аналізу

Кваліфікаційна магістерська робота

за темою
«МУЛЬТИКАНАЛЬНИЙ СЕНТИМЕНТ-АНАЛІЗ ТА ЙОГО ЗАСТОСУВАННЯ У ПРОГНОЗУВАННІ РИНКОВИХ ТЕНДЕНЦІЙ КРИПТОВАЛЮТ»

Виконав: ***студент групи САДМ-61,***
Прокопенко Д. О.

Науковий керівник:
Доц. кафедри ІСТ, Доктор філософії (PhD),
Кузміч М. Ю.

Київ-2024

Актуальність, мета, об’єкт та предмет дослідження, практичне значення

- **Актуальність** обумовлена високою волатильністю ринку криптовалют та залежністю від настроїв учасників, що вимагає застосування інноваційних методів аналізу. Мультиканальний сентимент-аналіз дозволяє враховувати інформацію з різних джерел для точнішого прогнозування ринкових тенденцій.
- **Мета**: розробка методики мультиканального сентимент-аналізу для прогнозування цінових коливань криптовалют.
- **Об'єкт дослідження**: процес прогнозування ринкових тенденцій у криптоіндустрії з урахуванням впливу сентименту.
- **Предмет дослідження**: мультиканальний сентимент-аналіз для оцінки настроїв і впливу на ціни.
- **Практичне значення:** застосування сентимент аналізу виступає одним з інструментів для трейдерів і аналітиків, що дозволяють прогнозувати ринкові тенденції та приймати більш обґрунтовані рішення.

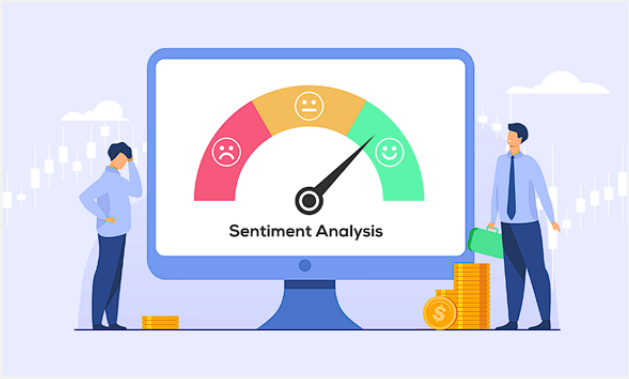
Завдання роботи

- Дослідити існуючі методи збору та аналізу даних із соціальних мереж, форумів та новинних платформ.
- Розробити підхї проведення мультиканального сентимент аналізу для подальшого прогнозування.
- Інтегрувати моделі машинного навчання для аналізу настроїв учасників ринку.
- Провести кореляційний аналіз настроїв і ринкових змін для виявлення взаємозв'язків.
- Перевірити ефективність запропонованої методології на реальних даних криптовалютного ринку.
- Оцінити практичну значимість розглянутого підходу для учасників криптовалютних ринків, включаючи інвесторів, трейдерів та аналітиків.

Завдання роботи

- Дослідити існуючі методи збору та аналізу даних із соціальних мереж, форумів та новинних платформ.
- Розробити методику мультиканального сентимент аналізу для подальшого прогнозування.
- Інтегрувати моделі машинного навчання для аналізу настроїв учасників ринку.
- Провести кореляційний аналіз настроїв і ринкових змін для виявлення взаємозв'язків.
- Перевірити ефективність запропонованої методології на реальних даних криптовалютного ринку.
- Оцінити практичну значимість розглянутого підходу для учасників криптовалютних ринків, включаючи інвесторів, трейдерів та аналітиків.

Основні поняття, застосування та виклики

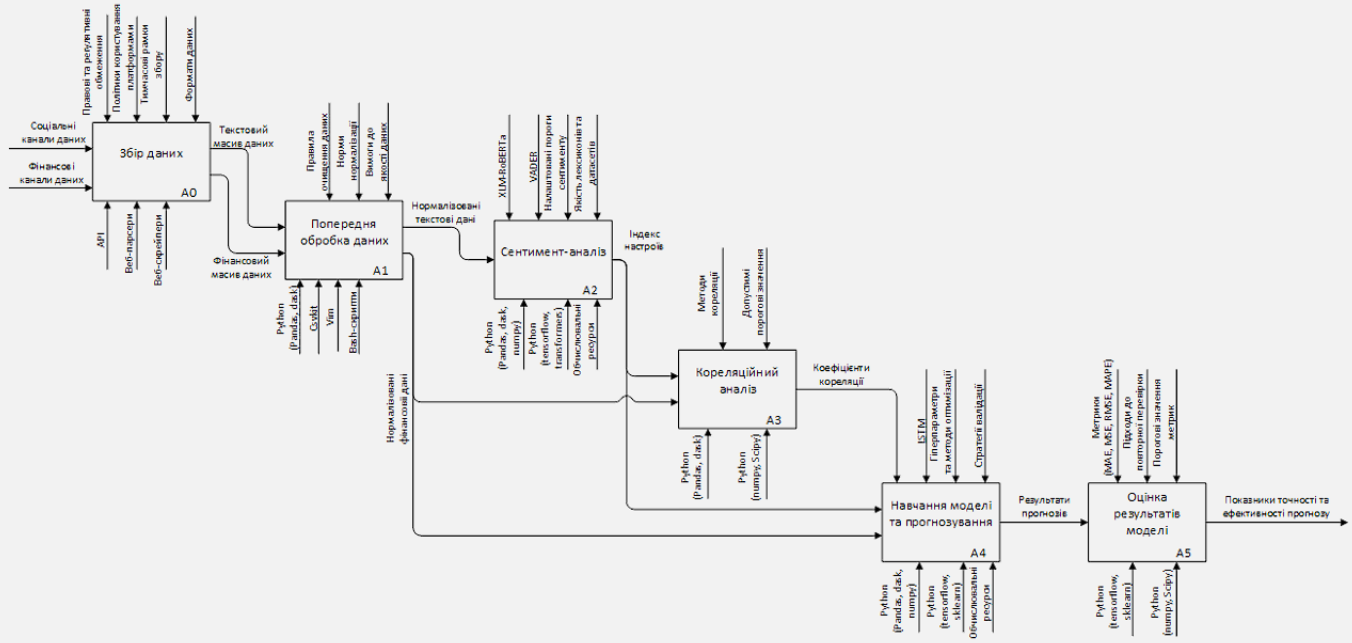


- **Сентимент-аналіз** — це метод обробки текстової інформації, що визначає емоційне забарвлення тексту (позитивне, нейтральне, негативне).
- **Об’єкти аналізу:** текстові повідомлення, коментарі, новини, публікації в соціальних мережах, блоги.
- **Мультиканальний підхід** — обробка даних з різних джерел для отримання ширшого контексту.
- **Застосування в криптовалютах:** аналіз настроїв трейдерів і інвесторів, прогнозування ринкових тенденцій на основі текстових сигналів, Виявлення потенційних ризиків та можливостей.
- **Виклики:** мова специфічних термінів криптовалют, швидкість змін настроїв у реальному часі, великий обсяг мультиканальних даних.

Методика сентимент-аналізу та прогнозування

1. **Збір даних.** Визначення ключових джерел та отримання релевантної інформації для подальшого прогнозу.
2. **Попередня обробка даних.** Очищення та нормалізація текстових і ринкових часових рядів для узгодження форматів і часових міток.
3. **Сентимент-аналіз.** Визначення полярності та об’єднання результатів із різних джерел.
4. **Кореляційний аналіз.** Перевірка взаємозв’язку між індексом настроїв і ринковими показниками для виявлення ступеня впливу настроїв на ринкову динаміку.
5. **Вибір та навчання моделі.** Обрання та налаштування статистичної чи ML-моделі з урахуванням кореляційних результатів і ринкових індикаторів.
6. **Оцінка моделі.** Використання показників для оцінки точності, а також порівняння з кореляційними оцінками.

Методика сентимент-аналізу та прогнозування



Інструменти, технології та методи мультиканального сентимент-аналізу

Етап	Методи/Процеси	Інструменти/Технології
Збір даних	Соцмережі (Twitter, Telegram), новини, форуми	API, веб-скрапінг (BeautifulSoup, Scrapy), веб-парсинг
Попередня обробка	Очищення тексту, токенізація, нормалізація	Python (nltk, spacy, pandas, dask, NumPy, SciPy), csvkit, vim
Формування індексу	Агрегація тональності, зважування джерел	Python (NumPy, Pandas)
Сентимент-аналіз	Словниковий (VADER), ML (LSTM, Transformers, xml-RoBERTa)	Python (TensorFlow, PyTorch), Hugging Face, NLTK, SpaCy
Кореляційний аналіз	Кореляція Пірсона, Спірмена	Python (SciPy, Statsmodels, NumPy)
Навчання моделі	ARIMA, LSTM, Transformer	Python (Scikit-learn, TensorFlow, PyTorch)
Оцінка моделі	Метрики: MAE, MSE, RMSE, MAPE	Python (Scikit-learn), Excel, BI-інструменти

Кореляційний аналіз сентименту та ринкових тенденцій

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

n — к-сть спостережень (парних значень x_i, y_i);
x_i — i-те значення змінної X;
$\bar{x}$ — середнє арифметичне всіх значень x_i;
y_i — i-те значення змінної Y;
$\bar{y}$ — середнє арифметичне всіх значень y_i.

Дата	Фактична ціна	Зміна ціни (у %)	Сентимент (X)	Сентимент (Reddit)	Сентимент (комбінований)
01.12.2022	16972		-0.7	-0.76	-0.73
02.12.2022	17093.6	0.72	0.41	0.45	0.43
03.12.2022	16884.5	-1.22	-0.24	-0.5	-0.37
04.12.2022	17112.6	1.35	0.35	0.39	0.37
05.12.2022	16966.5	-0.85	-0.37	-0.38	-0.37
06.12.2022	17089.3	0.72	0.37	0.35	0.36
07.12.2022	16835.2	-1.49	-0.34	-0.22	-0.28
08.12.2022	17225.7	2.32	0.61	0.65	0.63
09.12.2022	17125.7	-0.58	-0.21	-0.42	-0.32
10.12.2022	17127.2	0.01	-0.1	-0.11	-0.11
...	...	...	...	...	...

Порівняльні таблиця результатів прогнозів одноканального та мультиканального підходів

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|;$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)}{n}};$$

Дата	Фактична ціна	Прогнозована ціна (мультикан.)	Відхилення % (мультикан.)	Прогнозована ціна (одноканал.)	Відхилення % (мультикан.)
22.12.22	16820.6	16846.88	0.17	16868.66	0.29
23.12.22	16831.8	16873.65	0.25	16879.75	0.37
24.12.22	16837.2	16778.58	-0.35	16795.77	-0.46
25.12.22	16796.8	16689.25	-0.66	16654.82	-0.84
26.12.22	16706.1	16614.65	-0.55	16610.18	-0.57
27.12.22	16643.4	16619.77	-0.27	16669.5	0.49
28.12.22	16630	16607.58	-0.73	16600.8	-1.64
29.12.22	16646	16658.44	0.21	16664.9	0.14
30.12.22	16640.2	16652.31	0.38	16662.91	-0.03
31.12.22	16537.4	16659.99	0.74	16620.99	0.5

n — к-сть спостережень (парних значень $y_i, \hat{y}_i$);
y_i — фактичне значення для i-го спостереження;
$\hat{y}_i$ — передбачене значення для i-го спостереження.

Підхід	Показники			
	MAE	MSE	RMSE	MAPE
Однокан. підхід	88.54	13082.0	114.38	0.53%
Мультикан. підхід	86.49	11186.82	105.77	0.52%

Порівняння графіків прогнозованих та фактичних даних

Особливості, які слід враховувати під час аналізу та прогнозування

- Результати моделі не ставлять на меті точкового прогнозу цін, а формують одне із джерел для комплексного аналізу при прийнятті рішень.
- Аналіз настроїв не є універсальним інструментом, але він слугує важливим елементом для формування комплексного підходу до прогнозування цінових змін.
- Вибір набору моделей для аналізу прогнозування цінових тенденцій є індивідуальним рішенням трейдера, що виходить із його професійної підготовки.
- Дані для прийняття кінцевого рішення базуються на комплексі інформації, в якому кожна інформація поодинці необхідна, а в сукупності достатня для прийняття рішення.

Висновки

Під час роботи були виконані такі завдання:

- Досліджено методи збору та аналізу даних із соціальних мереж, форумів і новинних платформ.
- Розроблено методику мультиканальний сентимент-аналізу для точного прогнозування ринкових тенденцій.
- Інтегровано моделі машинного навчання для аналізу настроїв учасників крипторинку.
- Встановлено взаємозв'язок між настроями ринку та ціновими змінами криптовалют.
- Перевірено ефективність запропонованої методології на реальних даних.

Таким чином, результати проведеного дослідження підтвердили важливість мультиканального підходу до сентимент-аналізу для прогнозування ринкових тенденцій. Запропонована методологія дозволяє ефективно аналізувати ринкові настрої, що може слугувати базою для прийняття інформованих рішень та подальших досліджень у цій сфері.

Подальший розвиток може включати адаптацію методології до інших фінансових ринків, розширення джерел даних та впровадження нових моделей машинного навчання для підвищення точності аналізу.

Дякую за увагу!