МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет «Києво-Могилянська академія»
Факультет економічних наук
Кафедра фінансів

Магістерська робота
ОСВІТНІЙ СТУПІНЬ - МАГІСТР
на тему: « ФОРМУВАННЯ ПОРТФЕЛЯ ЦІННИХ ПАПЕРІВ З УРАХУВАННЯМ УПЕРЕДЖЕНОСТЕЙ ЗА ДОПОМОГОЮ МЕТОДІВ ШТУЧНОГО ІНТЕЛЕКТУ»

Виконала: студентка 2-го року навчання,
спеціальності 072 «Фінанси, банківська
справа та страхування»

Набок Валерія Дмитрівна

Керівник: Семко Р. Б.
кандидат економічних наук, доцент

Рецензент: Зубченко В. П.
кандидат фізико-математичних наук,
доцент кафедри теорії ймовірностей,
статистики та актуарної математики
КНУ ім. Тараса Шевченка

Магістерська робота захищена
з оцінкою «_____»

Секретар ЕК _____
«____» _____ 2023 р.

Київ - 2023

**ЗМІСТ**

# ВСТУП

Пост-індустріальна епоха – тобто нинішня– найбільше характеризується з-поміж попередніх історичних епох двома ключовими рисами: діджиталізацією та глобалізацією. Останні два чинники значно вплинули й на фінансовий сектор, а особливо на торгівлю на фондових біржах. Сучасні технології започаткували новий вид трейдингу, що зветься алгоритмічним трейдингом (англ. high-frequency trading). HFT змінив спосіб мислення трейдерів, значно скоротивши час прийняття рішення, проте вагомо підвищивши волатильність цінних паперів, тим самим кинувши виклик інвесторам шукати нових способів збільшити прибутковість своїх портфелів.

Завдяки високій волатильності та прибутковості фондовий ринок завжди був надзвичайно привабливим для інвесторів.  Серед фаворитів у сучасних інвесторів має місце торгівля на фондових біржах США та Великої Британії. Нью-Йоркську фондову біржу не даремно називають “імперією фінансових магнатів”. Сукупна ринкова капіталізація компаній, акції яких торгуються на Нью-Йоркській біржі вимірюється десятками трильйонів доларів [1]. Заснована у 1971 р. система електронного котирування акцій NASDAQ є однією найбільших у світі електронних бірж, лістинг на якій мають близько 3300 компаній різних галузей економіки. Цей ринок може здійснювати оборот до 4 млрд акцій за день. [2]. Проте людям притаманно приймати забарвлені поведінковими упередженнями фінансові рішення які змушують їх діяти емоційно або помилятись у процесі обробки інформації.

Технічний прогрес та поява інтернету вплинув на формування переконань людей, розширюючи вплив окремих знаменитостей настільки, що це було неможливо уявити раніше. Наприклад, акції компанії Tesla

втратили майже 9% вартості на тлі публікації її засновником Ілоном Маском свого бачення «умов» завершення війни між Україною та росією, серед яких пропозиція віддати рф Крим.

Зазвичай, люди покладаються у своїх рішенням у формуванні потрфелю цінних паперів на одну з двох парадигм – фундаментальний або технічний аналіз. У першому випадку, інвестори зважають на вартість акцій та репутацію компанії, стан економіки, політичний клімат тощо. У другому ж випадку,  використовують технічний аналіз – набір статистичних методів, що націлені на дослідження ключевих метрик, як-от історичні дані цін та обсягів продажів цінних паперів. Зважаючи на суттєве збільшення даних, покладатись на рішення, прийняте на основі технічного аналізу стає дедалі складніше. У зв’язку з цим також зростає ризик прийняття упереджених рішень. Для вирішення цієї проблеми у нагоді стають методи штучного інтелекту.

**Актуальність теми** полягає у дослідженні впливу поведінкових упередженостей на формування інвестиційного портфелю, аналізу фондового ринку та його оцінку, а також використанні найновіших статистичних методів у вирішенні задачі формування найбільш прибуткових інвестиційних портфелів шляхом прогнозування прибутковостей цінних паперів.

**Метою дослідження** є дослідити та поглибити знання з алгоритмів ШІ, що працюють з множиною даних з фондових бірж, вирішуючи задачу утворення портфелю цінних паперів в умовах недостатньої інформованості, та доповнюють свої прогнози за допомогою даних соціальних мереж.

Досягнення мети було визначене наступними необхідними завданнями:

- дослідити сучасний стан фондових ринків
- визначити основні типи гравців на фондових ринках
- дослідити вплив поведінкових упередженостей на формування інвестиційного портфелю
- опанувати алгоритми обробки та аналізу текстових даних та фінансових часових рядів
- сформувати інвестиційні портфелі на основі алгоритмів ШІ
- проаналізувати результати

**Об’єктом дослідження** є поведінкові упередженості та їхній вплив на формування інвестиційних портфелів.

**Предметом дослідження** є підходи до формування інвестиційних портфелів за допомогою методів штучного інтелекту з урахуванням поведінкових упередженостей .

**Методи дослідження,** використані в даній роботі включають в себе: метод аналізу і синтезу , метод гіпотез та припущень, метод логічного узагальнення, методи діалектики: індукція й дедукція, описово-аналітичний з метою аналізу досліджуваних явищ та процесів.

Для реалізації статистичних моделей формування інвестиційних портфелів було використано наступне програмне забезпечення на основі мови програмування Python :

- фреймворк Tensorflow для побудови моделей прогнозування часових рядів
- інструментарій аналізу текстової інформації NLTK
- бібліотека PyfolioOpt для обчислення ваг інвестиційних портфелей

- бібліотека Word2Vec для синтезу кількісних характеристик досліджуваних текстів

**Наукова новизна одержаних результатів** полягає в проведенні ґрунтовного аналізу наявних алгоритмів ШІ для формування інвестиційних портфелів.

**Інформаційну базу** даного дослідження становлять роботи провідних іноземних авторів на тематику аналізу фондових ринків та теорії портфелю Марковіца, звіти компаній лідерів галузей, що входять до синтезованого інвестиційного портфелю, бази даних фінансових ресурсів, дані соціальної мережі Twitter та платформи Kaggle.

**Структура роботи відповідає поставленим завданням**. Магістерська робота складається з вступу, трьох розділів, висновків, списку використаних джерел та додатків.

У першому розділі розтлумачено поняття інвестиційного портфелю та поведінкових упереджень, наведено їх таксономію та наданий огляд існуючих методів формування інвестиційного портфелю. У другому розділі були описані алгоритми машинного навчання, що працюють з фінансовими та текстовими даними, а також запропоновано алгоритми машинного навчання, метою якого є синтез оптимального інвестиційного портфелю з урахування упередженості. У третьому розділі було продемонстровано роботу запропонованого алгоритму та сформовано низку інвестиційних портфелів з різним степенем упередженості та проаналізовано результати.

# РОЗДІЛ 1 ПОВЕДІНКОВІ ФІНАНСИ В ЗАДАЧІ ПОБУДОВИ ІНВЕСТИЦІЙНИХ ПОРТФЕЛІВ

## 1.1 Основні поняття теорії інвестиційних портфелів

Історія теорії ІП починається з часів великої депресії, а саме 1930их років. У 1938 році Джон Берр Вільямс написав книгу під назвою «Теорія інвестиційної вартості», яка відобразила мислення того часу: модель дисконтування дивідендів. Тодішньою метою інвесторів була купівля найбільш прибуткових акцій серед стабільних. Але так як інвестори не були всебічно обізнані щодо стану компаній та щодо ринку загалом, задача пошуку найоптимальнішого портфелю є більш складною, ніж задача пошуку найбільш прибуткових акцій, оскільки доля невизначеності ринку продукує стохастичну природу прибутковості цінних паперів.

При формуванні ІП інвестор має заздалегідь визначити для себе мету та стратегію ведення даного портфелю. Важливим фактором також є оцінка ступені ризику, на який готовий піти інвестор, адже від цього залежить швидкість отримання прибутку, а також його розмір. З усього вищесказаного, можна зробити висновок, що процес створення і керування портфелем цінних паперів повний компромісів, а отже вимагає вміння зважувати сильні та слабкі сторони, можливості та загрози в усьому спектрі інвестицій.

За часовим проміжком стратегії можуть бути:

- Довгострокові
- Короткостроковими.

Важливо розуміти, що для довгострокових стратегій більш притаманна менш агресивна поведінка над ринком цінних паперів, у той час як для короткострокових стратегій припускається більш високі ризики.

## 1.2 Види інвестиційних портфелів

По відношенню до вразливості інвестора до ризиків, виділяють такі типи інвесторських портфелей:

- Консервативний ІП
- Агресивний ІП
- Збалансований ІП
- Помірний ІП

### 1.2.1. Консервативний ІП

Консервативний тип інвестицій - це інвестиційна стратегія, при якій пріоритетом є збереження капіталу над зростанням чи ринковою прибутковістю. Формування такого портфелю властиве  схильним до мінімізації ризиків інвесторам. При обранні даної стратегії, ІП матиме більшу частку низькоризикових інвестицій з фіксованим доходом і меншу кількість високоякісних акцій або фондів. Консервативна стратегія вимагає інвестування в найбезпечніші короткострокові інструменти, такі як казначейські векселі та депозитні сертифікати.[3]

При консервативній стратегії інвестування більше половини портфеля, як правило, зберігається в боргових цінних паперах та грошових еквівалентах, а не в акціях чи інших ризикованих активах. І хоча консервативна стратегія інвестування може захистити від інфляції, вона може не принести значних прибутків з часом порівняно з більш агресивними стратегіями.

Загалом, консервативний ІП За типом портфеля можна віднести до високонадійних, але низькоприбуткових інструментів. Варто також відзначити, що середньорічна прибутковість консервативної стратегії становитиме 5-6% річних.

### 1.2.2. Агресивний ІП

Основною задачею при використанні стратегії агресивного інвестування є максимізація потенційних прибутків від портфеля, при цьому із високим рівнем ризику. Стратегії досягнення доходів, вищих за середні, зазвичай наголошують на прирості капіталу як на головній інвестиційній меті, а не на доході чи безпеці основного капіталу. Таким чином, така стратегія передбачатиме розподіл активів із суттєвою вагою в акціях і, можливо, невеликий розподіл або відсутність розподілу на облігації чи готівку.

Стратегії агресивного інвестування можуть також включати стратегію високого обороту, яка має на меті перекуповування акцій, які показують високу відносну ефективність за короткий період часу. Високий оборот може створити більший прибуток, але також може призвести до вищих транзакційних витрат, збільшуючи таким чином ризик поганої роботи(poor performance).

Агресивна стратегія потребує більш активного управління, ніж консервативна стратегія, оскільки вона, швидше за все, буде набагато нестабільнішою та потребуватиме частих коригувань залежно від ринкових умов. Крім того, знадобиться більше ребалансування, щоб повернути розподіл портфеля до цільового рівня. Волатильність активів може призвести до суттєвого відхилення розподілу від початкової ваги. Ця додаткова робота також призводить до вищих гонорарів, оскільки

менеджеру портфоліо може знадобитися більше персоналу для керування всіма такими позиціями.

У портфель із агресивною стратегією часто входять :

- цінні папери невеликих, новостворених емітентів з переважанням венчурних інвестицій, перспективні стартапи та інноваційні проекти.
- акції зростання

Незважаючи на те, що дохідність при агресивній стратегії становитиме 12-14% річних, зважуючи на нестабільну та непрогнозовану ситуацію на глобальному ринку, а також поганий перформанс менеджерів останнім часом спостерігається значний відступ проти стратегій активного інвестування. Багато інвесторів вивели свої активи з хедж-фондів, наприклад, через неефективність менеджерів. Натомість деякі вирішили розмістити свої гроші пасивним менеджерам. Ці менеджери дотримуються стилів інвестування, які часто використовують управління індексними фондами для стратегічної ротації. У цих випадках портфелі часто відображають ринковий індекс, такий як S&P 500 [4].

### 1.2.3. Збалансований ІП

Збалансована інвестиційна стратегія поєднує різні види активів у портфелі, намагаючись при цьому збалансувати ризик і прибутковість. На практиці, як правило, збалансовані портфелі складаються переважно з (майже) рівних долей акцій і облігацій. Наприклад, 60% в акціях і 40% в облігаціях. Збалансовані портфелі також можуть мати долю готівки або інших високоліквідних фінансових інструментів для забезпечення ліквідності.

При формуванні збалансований портфелю інвестор ставить перед собою мету прибутковості даного портфелю за будь-яких умов: інфляція, падіння ринку чи настання рецесії.

Дана ціль може бути досягнена шляхом поєднання агресивної стратегії з консервативною, тобто додавання до портфеля інструментів, що працюють у тривалій перспективі.Слід зазначити, що дані активи мають по-різному реагувати на ринкові зміни, а також кожен актив, що входить у збалансований портфель, має бути по-своєму волатильний.

Даний тип інвестиційного портфелю притаманний для інвесторів, які ставлять за мету поступове та помірковане, але гарантоване збільшення капіталу. Як правило, до такого типу інвесторів належать інвестори з низьким рівнем ризику.

До цінних паперів, які найчастіше входять у збалансований портфель належать:

- державні цінні папери, у тому числі великі і надійні емітенти
- Акції чи ETF фонди.

Середньорічна дохідність збалансованих портфелів зазвичай становить е 8-10% річних.

## 1.3 Задача розрахунку ваг оптимального ІП

Типи портфелю, наведені вище є більш інтуітивно зрозумілими, аніж формально означеними. Для подальших цілей даної роботи пропонується опиратися на кількісну характеристику прибутковості та ризикованості портфелю, згідно теорії портфелів Марковіца.

У березні 1952 року Гаррі Марковіц опублікував знамениту статтю «Вибір оптимального портфелю», в якій він вперше представив теорію оптимізації портфеля, засновану на компромісі середнього - дисперсії.

Марковіц заявляв, що інвестор «бажає (або має бажати) збільшення очікуваного прибутку, при цьому не бажаючи збільшення волатильності»[7]. Зі слів американського економіста, дотримуючись певних припущень і умов, рішення інвесторів можна звести лише до очікуваної прибутковості та дисперсії портфеля.

На думку Марковіца, ефективними портфелями є ті, які для будь-якої визначеної величини дисперсії мають найвищий очікуваний прибуток.

Набір усіх цих портфелів створює так звану «ефективну межу», де інвестори мають обрати свою стратегію відповідно до своїх конкретних упереджень щодо ризику та прибутку. Згідно Марковіца, на загальний ризик портфеля впливає не лише дисперсія доходності окремих активів, а й набір коваріацій усіх цінних паперів. Тому важливо враховувати всі можливі взаємодії між різними інвестиціями. Роблячи це, інвестор створює портфель із таким же очікуваним прибутком, але з нижчим ризиком, ніж портфель, який не враховує ці сумісну залежність активів.

Поставлена задача оптимізації інвестиційного портфелю може бути формалізована наступним чином: нехай $w_i$ – вага і-го активу в дослідженому портфелі, $r_i \sim N(\mu_i, \sigma_i)$ – прибутковість даного активу. Тоді Очікувана прибутковість портфелю дорівнює :

$$E(r_p) = \left( \sum_i w_i \mathbb{E}(r_i) \right) \tag{1.1}$$

Волатильність портфелю дорівнює:

$$\sigma_p^2 = \sum_i w_i^2 \sigma_i^2 + \sum_i \sum_j w_i w_j \sigma_{ij} \tag{1.2}$$

У рівнянні (1.2) ефект диверсифікації визначається першим членом правої частини, який дійсно залежить лише від значень дисперсій цінних паперів. Чим більше активів включено в портфель, тим більше ця частина

рівняння наближається до 0. Отже, ті інвестори, які розподіляють свої загальні статки в різні активи, побачать, що загальна дисперсія портфеля зменшується. Однак не всю волатильність портфеля можна диверсифікувати. Правий член рівняння (1.2) справді залежить лише від коваріації активів і не залежить від кількості цінних паперів у портфелі[8].

Sharpe ratio є одним із найвідоміших методів оцінки ефективності портфеля. Цей коефіцієнт допомагає інвесторам зрозуміти прибутковість їхніх інвестицій з поправкою на ризик.

Він був розроблений у 1966 році Вільямом Шарпом для проведення аналізу продуктивності менеджерів взаємних фондів[9]. Наприклад, менеджер міг забезпечити дуже високий рівень прибутку протягом певного періоду часу. Питання, на яке намагається відповісти sharpe ratio, полягає в тому, який ризик прийняв менеджер, щоб отримати ці прибутки. Це може бути дуже важливо в умовах спадаючого ринку.

$$\text{Sharpe Ratio(SR)} = \frac{\mathrm{E}(\mathrm{r_p})}{\sqrt{\sigma_p^2}} \rightarrow \max \tag{1.3}$$

З формули (1.3) видно, що SR вимірює додатковий прибуток на одиницю збільшення волатильності. Вищий коефіцієнт портфеля означає кращу прибутковість з поправкою на ризик; отже, це фундаментальний інструмент для критичного аналізу інвестиційних рішень. Однак цей коефіцієнт базується на деяких важливих припущеннях, які можуть обмежити його надійність як хорошого показника ефективності ризику. Наприклад, стандартне відхилення (волатильність) використовується як проксі для ризику, навіть якщо прибутковість на фінансових ринках виявилася відхиленою від середнього через вагомі зміни на ринку [10].

## 1.4 Таксономія видів інвесторів по відношенню до набутих поведінкових упереджень

Поведінкові фінанси представляють сучасний підхід у фінансовій науці, який спрямований на розуміння унікальних аспектів прийняття фінансових рішень, шляхом поєднання теоретичних принципів психології, економічної теорії та фінансового аналізу. Вони займають все дедалі більш помітне місце у світі фінансових консультацій з моменту краху доткомів в березні 2000 року. Дане твердження проаналізовано та задокументовано в дослідженні Майкла Помпіана під назвою «Підхід до кінця: «Знай свого клієнта» використання поведінкових фінансів для утримання та примноження багатства клієнтів». У рамках цього дослідження було опитано 290 досвідчених фінансових консультантів у 30 країнах щодо їхньої власної зацікавленості та використання поведінкових фінансів під час комунікацій зі своїми клієнтами. Результати засвідчили, що 93% консультантів вважали, що індивідуальні інвестори приймають нераціональні інвестиційні рішення, і при цьому 96% із них успішно використовували поведінкові фінанси для покращення відносин зі своїми клієнтами.

Типи поведінкових інвесторів були розроблені з метою поліпшення роботи консультантів та ідентифікації потрібної ніші клієнтів та дозволяють швидко та всебічно оцінити, з яким типом інвестора вони мають справу, перш ніж рекомендувати інвестиційний план. Перевага попереднього визначення типу інвестора, з яким має справу консультант, полягає в тому, що це пом’якшить небажані «сюрпризи» від клієнта, який бажає змінити розподіл портфеля через зміну характеру ринку. У такому випадку консультант здатний надати найбільш компромісну альтернативу на ринку, мінімізуючи ймовірність великих втрат і при цьому задовільняючи потреби клієнта в темпах та об’ємах надбання додаткових

прибутків.  Варто зазначити, що кожен інвестор не є строгим втіленням одного з типу поведінкових інвесторів (ТПІ), а радше демонструє схильність до поведінки одного з запропонованих типів у більшості випадків. Наприклад, консультант може визначити коректність класифікації клієнта як певний ТПІ і при цьому виявити, що клієнт також має риси (упередження) іншого.

Кожен ТПІ характеризується певним рівнем толерантності до ризику та основним типом упередженості — або когнітивним (керованим неправильним міркуванням), або емоційним (керованим імпульсами чи почуттями). Однією з найважливіших концепцій являється те, що найменш толерантний до ризику ТПІ і найбільш толерантний до ризику ТПІ обумовлюються емоційними упередженнями, тоді як два проміжних типи в основному характеризуються когнітивними упередженнями. Варто зазначити, що з емоційними клієнтами, як правило, важче працювати. Консультанти, які можуть розпізнати тип клієнта, перш ніж давати рекомендації щодо інвестицій, будуть набагато краще підготовлені до боротьби з нераціональною поведінкою, у разі її виникнення.

З огляду на вищесказане, таксономія ТПІ характеризується 4 категоріями:

1. Пасивний консерватор
2. Дружній послідовник
3. Незалежний індивідуаліст
4. Активний накопичувач

### 1.4.1. Пасивний консерватор

Даний тип інвестора характеризується пасивністю, низькою толерантністю до ризику і характеризується емоційним упередженням.

Пасивні консерватори (ПК)— це, як випливає з назви, інвестори, які приділяють велику увагу фінансовій безпеці та збереженню багатства, при цьому не йдуть на ризик заради високих прибутків. Як правило, до ПК належать люди, що здобули власний капітал внаслідок хорошої роботи або отримавши його у спадок. Оскільки вони розбагатіли, не ризикуючи власним капіталом, пасивні консерватори можуть не демонструвати фінансову обізнанність. Зазвичай, пасивні консерватори повільно приймають рішення щодо інвестицій, оскільки не люблять змін. Це може бути обумовлено тим, яким чином вони підходили до свого професійного життя, намагаючись не ризикувати. Деякі пасивні консерватори, які успадковують багатство, можуть відчувати сильне почуття провини або низької самооцінки, оскільки вони не заробили своїх грошей, і можуть боятися невдачі або відсутності мотивації.

Більшість пасивних консерваторів зосереджені на збереженні прибутків на достатньому рівні, щоб дбати про сім'ю і майбутні покоління, особливо того, що стосується фінансування таких життєво важливих витрат, як витрати на освіту та купівлю житла. Оскільки увага зосереджена на сім'ї та безпеці, упередження Passive Preserver, як правило, емоційні, а не когнітивні. Із зростанням віку та рівня добробуту цей ТПІ стає все більш поширеним. Хоча це не завжди так, багато пасивних консерваторів не втручаються в процес управління своїми інвестиціями, а радше сконцентровані на ідеї збереження якості рівня свого життя, і тому, як правило, є зручними клієнтами. Поведінкові упередження ПК, як правило, є емоційними, орієнтованими на безпеку власної грошової «подушки» упередженнями, як endowment bias, loss aversion, status quo, and regret. Вони також демонструють когнітивні упередження, такі як anchoring і mental accounting.

*Endowment bias.* Ця емоційна упередженість виникає, коли людина надає більшу цінність об’єкту, коли володіє ним і при цьому стикається з його втратою, ніж коли не володіє об’єктом і має потенціал отримати його. Класичним прикладом даного типу упередженості є клієнт, який тримає інвестиції, якими володіли попередні покоління, зокрема левову частку акцій або нерухомості, які переходили у спадок між поколіннями, без обґрунтування того, чому ці активи зберігаються.

*Loss aversion bias.* Більшість пасивних консерваторів відчувають «біль» від втрат більше, ніж задоволення від виграшів — суть неприйняття втрат. Ця емоційна упередженість заважає людям позбуватися завідомо збиткових інвестицій, навіть якщо вони не бачать перспективи зростання. Як правило, зберігання подібного типу інвестицій нівелює прибутковість портфелю.

Проста діагностика упередження неприйняття втрат: надайте клієнту сценарій, за яким він купує цінний папір, і він впаде на 25 відсотків без передбачуваного відскоку. Запитайте, чи вони, ймовірно, утримають його, доки він не вирівняється, або продайте його і придбайте щось із кращими перспективами. Якщо вони тримаються на інвестиціях, вони, швидше за все, будуть упереджені щодо неприйняття втрат.

*Status quo bias.* Ця емоційна упередженість спонукає людей, стикаючись із безліччю варіантів, обирати той варіант, який зберігає умови незмінними. Пасивні консерватори часто схильні дотримуватись того, що «все завжди було так» і їм зручніше дотримуватись старого порядку речей. Упередженість статус-кво демонструє інвестор, який багато років робив речі певним чином, а потім наймає нового фінансового консультанта. Останній, в свою чергу, може запропонувати практичні зміни лише для того, щоб виявити, що інвестор бере участь або не приймає поради. Справа не в тому, що клієнту не потрібна хороша порада — він просто застряг у статус-кво.

*Regret aversion bias*. Люди, які демонструють цю емоційну упередженість, уникають рішучих дій, бо бояться, що в перспективі, який би альтернативу вони не вибрали, їхнє рішення виявиться менш оптимальним. Regret aversion може призвести до того, що інвестори будуть занадто консервативними у своєму інвестиційному виборі. Зазнавши збитків у минулому, вони можуть ухилятися від нових розумних інвестицій. Така поведінка може призвести до довгострокової недостатньої ефективності та поставити під загрозу інвестиційні цілі. Визначити схильність клієнта до даного типу упередженості можна через запитання, чи він робив інвестиції в минулому, про що шкодує, і чи впливає це на поточні чи майбутні інвестиційні рішення.

*Anchoring bias.* Ця когнітивна упередженість виникає, коли на інвесторів впливають ціна купівлі активу або довільні рівні цін, і вони схильні враховувати неактуальні ціни при прийнятті рішення щодо продажі активу.

Один із найпоширеніших прикладів даної упередженості відбувається під час впровадження нового розподілу активів. Наприклад, припустимо, що клієнт звертається до консультанта, маючи у своєму портфелю 30% в одній акції, і консультант рекомендує його диверсифікувати. Далі припустимо, що акція впала на 25 відсотків від максимуму, який він досяг п’ять місяців тому (75 доларів на акцію проти 100 доларів за акцію). Для простоти припустимо, що податки на продаж не є проблемою. Часто клієнти чинитимуть опір новому розподілу, оскільки вважають, що повинні продати акції лише тоді, коли їх ціна підскочить до 100 доларів за акцію, яку вона досягла п’ять місяців тому. У даному випадку можна казати про наявність anchoring bias.

*Mental accounting bias*. Останнє упередження пасивного консеватора— це mental accounting bias, когнітивне упередження, яке виникає, коли люди по-різному ставляться до різних сум грошей залежно

від того, як ці суми категоризовані самим інвестором. Пасивні консерватори не схильні до ризиків і люблять розділяти свої активи на безпечні категорії. Класичним прикладом такого обліку є розділення грошей на навчання, гроші на пенсію та гроші на відпустку. Якщо всі ці активи розглядати як «безпечні гроші», результатом зазвичай є неоптимальна прибутковість.

Отже, пасивних консерваторів вирізняє несхильність до ризику, емоційна упередженість, нетолерантність до втрат і пасивність.

### 1.4.2. Дружній послідовник

Даний тип інвестора характеризується пасивністю, низькою чи середньою толерантністю до ризику і характеризується когнітивним упередженням.

Friendly Followers — це пасивні інвестори, які зазвичай не мають власних уявлень про інвестування. Вони часто слідують прикладу своїх друзів і колег у прийнятті інвестиційних рішень і хочуть брати участь у останніх, найпопулярніших інвестиціях, не звертаючи уваги на довгостроковий план. Однією з ключових проблем роботи з дружніми послідовниками є те, що вони часто переоцінюють свою толерантність до ризику. Консультанти мають бути обережними, щоб не запропонувати занадто багато «гарячих» інвестиційних ідей — Дружні послідовники, швидше за все, захочуть зробити їх усі. Деякі являються необізнаними в основних принципах інвестування, багато з них також відкладає час звернення до фінансового консультанта заради прийняття інвестиційних рішень, що може стати причиною скупчення великої кількості незаінвестованих грошей. Дружні послідовники зазвичай дотримуються

професійних порад, коли вони їх отримують, і вони навчаються фінансово, але іноді це може бути важко, оскільки вони не люблять або не мають здібностей до процесу інвестування. Зазвичай, упередження дружніх послідовників когнітивні: recency, hindsight, framing, cognitive dissonance, і ambiguity aversion.

*Recency bias.* Це схильність людей більш помітно згадувати й наголошувати на нещодавних подіях чи спостереженнях, а також потенційно екстраполювати патерни там, де їх немає. Дана упередженість активно впливала на стан ринку в період економічної рецесії, відомий як «bull market», між 1995 і 1999 роками, коли багато інвесторів помилково припускали, що ринок продовжуватиме зростати вічно. Дружні послідовники часто створюють створюють свій інвестиційний портфель з активів, що демонструють зростання протягом тривалого періоду часу, ризикуючи зайти на ринок в піковий момент перед зниженням ринку.

*Hindsight bias.* Дружні послідовники, яким часто не вистачає незалежної думки щодо інвестицій, схильні до даного типу упередженості, що може виникати, коли інвестор проектує результати власних інвестицій так, ніби їхня поведінка є завчасно, навіть якщо дане твердження є помилковим. Якравим прикладом такого упередженого є реакція інвесторів на «бульбашку» технологічних акцій, коли спочатку багато хто вважав роботу ринку нормальним (не є симптомом бульбашки), а проаналізувавши ситуацію постфактум дійшли правильних висновків про стан ринку. Результатом цієї упередженості є те, що це дає інвесторам хибне відчуття безпеки під час прийняття інвестиційних рішень, і таким чином береться надмірний ризик.

*Framing bias*. Це тенденція дружніх послідовників по-різному реагувати на різні ситуації залежно від контексту, в якому присутній певний вибір. Інвестори часто зосереджуються занадто обмежено на одному або двох аспектах ситуації, виключаючи інші міркування.

Хорошим прикладом є використання результатів опитувальників щодо толерантності до ризику. Залежно від того, як задаються питання, дана упередженість може змусити інвесторів реагувати на питання, що стосуються толерантністі до ризику як надмірно консервативним, так і ризикованим способом. Наприклад, коли в запитанні ставиться акцент на ймовірні прибутки, тоді інвестор схильний до більшого ризику. У той же час, коли предмет ризику зміщається до питання уникнення трат, то інвестор демонструє меншу толератність до ризикованих активів.

Проста діагностика для визначення упередженості полягає в тому, щоб вибрати одне запитання із типового опитувальника про толерантність до ризику та перефразувати його, щоб перевірити, чи відповість клієнт на те саме запитання по-різному, залежно від того, як його задають. Якщо вони відповідають на запитання по-різному, вони, ймовірно, схильні до упередженості.

*Cognitive dissonance bias*. У психології когніції представляють собою установки, емоції, переконання або цінності. Коли перетинаються кілька когніцій — наприклад, виникає ситуація, коли людина шукає лише докази власного упередження щодо предмету дискусії тільки для того, щоб виявити, що це неправда, — вона намагається полегшити свій дискомфорт, ігноруючи правду та обґрунтовуючи своє рішення ігнорувати істину. Дружні послідовники, які схильні до цієї упередженості, можуть продовжувати інвестувати в цінний папір або фонд, яким вони вже володіють, після того, як він знецінився чи продовжує знецінюватись, навіть якщо вони знають, що їм слід об’єктивно оцінювати нові інвестиції.

*Ambiguity aversion bias.* Дана упередженість описує поширене когнітивне упередження, яке виникає, коли нам потрібно зробити інвестиційне рішення в умовах недостатньої інформованості. Особа, яка не схильна до неоднозначності, скоріше вибере альтернативу, де розподіл ймовірностей результатів відомий, а не альтернативу, де ймовірності

невідомі. Ця поведінка вперше була представлена через парадокс Еллсберга (люди вважають за краще робити ставки на результат урни з 50 червоними та 50 чорними кулями, а не на урну із 100 кульками, але для якої кількість чорних чи червоних куль невідома) . Ризиковані події мають відомий розподіл ймовірностей за результатами, тоді як для неоднозначних подій розподіл ймовірностей невідомий. Різниця між неприйняттям двозначності(ambiguity aversion bias) та уникненням ризику(risk-aversion) важлива, але непомітна. Уникнення ризику виникає в ситуації, коли ймовірність може бути приписана кожному можливому результату ситуації, і вона визначається перевагою між ризикованою альтернативою та її очікуваною цінністю. Відмова від неоднозначності стосується ситуації, коли ймовірності результатів невідомі, і вона визначається через перевагу між ризикованими та неоднозначними альтернативами, після контролю переваг над ризиком. Варто зазначити, що навіть якщо інвестори відчувають себе професіоналами, вони можуть не захотіти вкладатися «неоднозначні» інвестиції, як-от акції, навіть якщо вони вважають, що можуть передбачити ці результати на основі власного судження.

Дружні послідовники часто переоцінюють свою толерантність до ризику. Ризикована поведінка, що слідкує за тенденцією, частково виникає через те, що дружні послідовники не люблять ситуацій двозначності. Вони також можуть переконати себе, що «знали про це весь час», що також підвищує ризикову поведінку. Консультанти повинні обережно поводитися з дружелюбними послідовниками, тому що вони, швидше за все, скажуть «Так» на пораду, яка для них має сенс. Консультанти повинні направляти їх, щоб вони уважно поглянули на поведінкові тенденції, щоб переоцінити свою толерантність до ризику. Оскільки упередження дружних послідовників мають переважно когнітивний характер, демонстрація ефективної диверсифікації портфеля зазвичай є найкрощою

порадою. Консультанти повинні закликати клієнтів дружних послідовників бути інтроспективними та надавати рекомендації на основі даних.

### 1.4.3. Незалежний індивідуаліст

Даний тип інвестора характеризується активністю, середньою чи високою толерантністю до ризику і характеризується когнітивним упередженням.

Незалежний індивідуаліст – це активний інвестор із толерантністю до середнього та високого рівнів ризику. Незалежні індивідуалісти схильні покладатися на минулий досвід, який іменують «інтуїцією» під час прийняття рішень; однак, демонструючи упереджене ставлення до об'єкту інвестицій, коли вони проводять дослідження самостійно, зазвичай не коригують свої прогнози, оскільки не сприймають альтернативних суджень із інших джерел. Іноді консультанти виявляють, що клієнт-незалежний індивідуаліст зробив інвестиції, не порадившись ні з ким. У свою чергу, така поведінка може бути проблематичною, оскільки завдяки своєму незалежному мисленню ці клієнти зберігають власну точку зору незмінною, навіть коли ринкові умови змінюються. Їм часто подобається інвестувати і легко ризикувати, але часто протистоять фінансовому плану.

Деякі незалежні індивідуалісти розглядають інвестування як спосіб заробити гроші, заради отримання безумовного пасивного доходу. Вони можуть бути хорошими клієнтами, оскільки вони зазвичай зайняті люди, хоча деякі не приймають фінансових порад. Інші ж фокусуються на спробах миттєвого збагачення і можуть мати сконцентровані недиверсифіковані портфелі. Упередження незалежних індивідуалістів є

когнітивними: conservatism, availability, confirmation, representativeness і self-attribution.

*Conservatism bias*. Дана упередженість зазвичай проявляється у тому, що люди зосереджуються на застарілих поглядах або прогнозах, не визнаючи нової актуальної інформації. Наприклад, припустимо, що інвестор купує цінний папір на основі очікуваного оголошення про новий продукт. Потім компанія оголошує, що має проблеми з виведенням продукту на ринок. Інвестор може концентруватися на акціях, при цьому звертаючи увагу лише на позитівні аспекти компанії, і не продавати їх через негативні новини.

*Availability bias*. Дана упередженість - це людська схильність покладатися на інформацію, яка легко спадає на думку під час оцінки ситуації або прийняття рішень. Через цю упередженість люди вважають, що доступна інформація більш репрезентативна, ніж насправді. Упередження доступності, також відоме як евристика доступності, є лише одним із ряду когнітивних упереджень, які перешкоджають критичному мисленню та, як наслідок, обґрунтованості наших рішень. Інформація може бути отримана з останніх або особливо яскравих спогадів. Це також може ґрунтуватися на особистому досвіді або підживлюватися зовнішніми джерелами, такими як новинні видання чи Інтернет. Покладання на цю інформацію допомагає людям уникнути копіткої перевірки фактів і аналізу, але збільшує ймовірність того, що їхні рішення будуть помилковими. Інвестори, які схильні до цього упередження, змушені вибирати компанії, що є медійно активними, але не є надійними, ігноруючи той факт, що деякі з найбільш ефективних фондів рекламуються дуже мало.

*Representativeness bias.* Дана упередженість виникає, коли подібність об’єктів або подій вводить людей в оману щодо ймовірності результату. Люди часто роблять помилку, вважаючи, що дві схожі речі чи події

тісніше пов’язані між собою, ніж є насправді. Це відбувається в результаті неправильної системи сприйняття під час обробки нової інформації. Незалежний індивідуаліст може розглядати певну акцію, наприклад, як цінну акцію, оскільки вона нагадує акцію попередньої вартості, яка була успішною інвестицією, але нові інвестиції насправді не є цінними акціями. Наприклад, активні біотехнологічні акції з мізерними прибутками або активами впадають на 25 відсотків після негативного оголошення продукту. Деякі незалежні індивідуалісти можуть вважати це репрезентативним для цінних акцій, оскільки вони дешеві; але біотехнологічні акції, як правило, не мають прибутку, тоді як акції традиційних галузей економіки мали прибутки в минулому, але вони тимчасово низькі.

*Self-attribution bias* Це відноситься до тенденції приписувати успіхи вродженим талантам і звинувачувати у невдачах за зовнішній вплив. Наприклад, припустимо, що Незалежний Індивідуаліст вкладається в інвестиції, які зростають. Причина його зростання пов’язана не з випадковими факторами, такими як економічні умови чи невдачі конкурентів, а з інвестиційною підкованістю інвестора. Це класична упередженість самоатрибуції.

*Confirmation bias.* Дана упередженість виникає, коли люди спостерігають, переоцінюють або активно шукають інформацію, яка підтверджує їхні твердження, ігноруючи чи знецінюючи докази, які можуть спростувати їхні твердження. Упередженість підтвердження може змусити інвесторів шукати лише інформацію, яка підтверджує їхні переконання щодо інвестиції, а не інформацію, яка може суперечити їхнім переконанням. Така поведінка може залишити інвесторів у збитках.

Незважаючи на можливу надмірну самооцінку, незалежні індивідуалісти є в основному підкованими гравцями на фондових ринках, а отже, є цінними клієнтами для фінансових консультантів.

### 1.4.4. Активний накопичувач

Даний тип інвестора характеризується активністю, високою толерантністю до ризику і характеризується емоційним та когнітивним упередженням.

Активний накопичувач є найбільш проактивним поведінковим типом інвестора. При високих рівнях достатку активні накопичувачі часто раціонально розпоряджаються результатами неінвестиційної діяльності та вважають, що вони здатні до аналогічної інвестиційної діяльності. Така поведінка може призвести до надмірної впевненості. Не послуговуючись необхідними консультаціями, активні накопичувачі часто мають високу оборотність портфеля, що гальмує ефективність інвестицій. Активні накопичувачі схильні асоціювати високі ризикі і волатильність з прибутковостями.

Деяким активним накопичувачам часто не прислуховуються до порад фінансових консультантів, оскільки вони зазвичай не вірять у основні принципи інвестування, такі як диверсифікація та розподіл активів. Вони прагнуть брати активну участь у процесі прийняття інвестиційних рішень, при цьому визнаючи, що бракує компетенції інвестиційних знань. Упередженнями активних накопичувачів є overcofidence, self-control, optimism, і illusion of control.

*Overconfidence bias*. Дана упередження характеризується необґрунтованою вірою у власні компетенції та здібності та містить як когнітивні, так і емоційні елементи. Надмірна самовпевненість проявляється в переоцінці інвесторами якості їхніх суджень. Багато активних накопичувачів стверджують, що здатні проаналізувати ринок ефективно, обравши лише найприбутковіші акції, однак численні дослідження показали, що дане твердження є помилковим. Наприклад,

дослідження, проведене дослідниками Одеаном і Барбером [53], показало, що після вирахування витрат на торгівлю (але до сплати податків) пересічний інвестор даного типу поступався медіанному гравцю на ринку приблизно на 2 відсотки на рік через необґрунтовану віру в свою здатність правильно оцінити вартість інвестиційних цінних паперів.

*Illusion of control bias.* Це когнітивне упередження виникає, коли люди вірять, що вони можуть контролювати результати інвестицій або принаймні впливати на них, хоча насправді дане твердження є хибним. Активні накопичувачі, які схильні до ілюзії упередженого контролю, вважають, що найкращий спосіб керувати інвестиційним портфелем — це це постійно його коригувати. Наприклад, орієнтовані на торгівлю інвестори, які толерують високий рівень ризику, вважають, що вони мають більше контролю над результатом своїх інвестицій, ніж є насправді, оскільки вони керують кожним власним рішенням.

*Self-control bias.* Це емоційне упередження, яке виникає, коли людям бракує самодисципліни та вони віддають перевагу короткостроковому задоволенню над довгостроковими цілями. Такі люди можуть віддавати перевагу малим виграшам зараз за рахунок більших виграшів у майбутньому (гіперболічне дисконтування). На фінансових ринках упередження самоконтролю може призвести до надмірної уваги до активів, що приносять дохід, для задоволення короткострокових потреб. У довгостроковій перспективі упередження самоконтролю може призвести до недостатніх заощаджень для фінансування пенсійних потреб, що, у свою чергу, може змусити інвестора піти на надмірний ризик пізніше у своєму житті, щоб спробувати компенсувати недостатнє накопичення заощаджень. Зазвичай, активні накопичувачі надають перевагу агресивним інвестиціям і мають високі поточні потреби у витратах, і у разі виникнення серйозної турбулентності на ринку, такий інвестор може бути

змушений продати солідні довгострокові інвестиції, вартість яких знизилася через поточні ринкові умови просто для покриття витрат.

*Optimism bias.* Інвестори з таким типом упередженості схильні переоцінювати ймовірність переживання позитивних подій і недооцінювати ймовірність переживання негативних подій. Така ілюзія може стати причиною конструювання збиткового портфелю, оскільки люди не визнають потенційних несприятливих наслідків інвестиційних рішень, які вони приймають. Яскравим прикладом такого емоційного упередження можна назвати випадок, коли працівники виділяють значну частку своїх доходів на акції компанії, в якій вони працюють. Невиправданий оптимізм спонукає цих працівників сприймати свою фірму як таку, що навряд чи постраждає від економічних негараздів.

## 1.5 Метрики технічного аналізу ринків у задачі дослідження фінансових рядів

Задача дослідження фінансових часових рядів є досить складною оскільки на даний ряднакладається умова на стохастичність та неповну проінформованість особи, що приймає рішення. Тобто, дані ряди складно прогнозувати, озираючись лише на базові його характеристики, такі як ціни відкриття позиції, ціни закриття позиції та об’єми торгів, тощо.

Дисципліна технічного аналізу фінансових ринків ставить перед собою мету синтезу нових характеристик досліджуваних рядів задля спрощення задачі аналізу та прогнозування цих рядів. В рамках еволюції даної дисципліни було створено низку функцій, які називаються індикаторами. Роль індикаторів – агрегувати певні особливості досліджуваних рядів за певні проміжки часу. Індикатори широко використовуються професійними трейдерами, а також ботами

високочастотної торгівлі (алгоритмічної торгівлі). Саме тому в рамках даного дослідження пропонується використати наступні індикатори:

1. $\mathrm{BOP} = \frac{\mathrm{Close_i} - \mathrm{Open_i}}{\mathrm{High_i} - \mathrm{Low_i}}$
2. $\mathrm{CCI} = \frac{\text{Typical Price} - \mathrm{MA}}{0.15 \times \mathrm{MeanDeviation}}$
3. $\mathrm{MFI} = 100 - \frac{100}{1 + \mathrm{MoneyFlowRatio}}$
4. $\mathrm{StochRSI} = \frac{\mathrm{RSI_i} - \mathrm{MIN[RSI]_i}}{\mathrm{MAX[RSI]_i} - \mathrm{MIN[RSI]_i}}$

Вибір даних індикаторів зумовлений їх широко розповсюдженістю, а тобто практичністю.

Balance of Power (BOP) – індикатор, основна особливість якого – коливання навколо 0, що є центральною віссю графіка BOP. Індикатор BOP вимірює здатність покупців підштовхувати ціни до вищих показників проти здатності продавців штовхаати ціни до нових мінімумів. Це відбувається шляхом коливання графіку (хвилеподібного руху вгору і вниз) між -1 і +1, при цьому позитивні значення інтерпретуються як сильніший тиск з боку покупців, а негативні значення вказують на сильніший тиск з боку продавців. Це означає, що коли індикатор BOP має позитивне значення, ринок віддає перевагу висхідному тренду, а коли негативне – нисхідному, що означає домінування покупців/продавців відповідно. Індикатори BOP часто використовуються в поєднанні з іншими індикаторами технічного аналізу, щоб дати трейдерам краще уявлення про те, в якому напрямку може рухатися ринок, і як довго він буде продовжувати рухатися в тому ж напрямку[35].

Індекс товарних каналів (CCI) - це технічний індикатор, який вимірює різницю між поточною ціною та середньою історичною ціною. Коли CCI вище нуля, це означає, що ціна вище середнього історичного значення. І

навпаки, коли CCI нижче нуля, ціна нижче історичного середнього значення.

CCI в основному використовується для виявлення нових трендів, спостереження за рівнями перекупленості і перепроданості, а також для виявлення слабкості трендів, коли індикатор розходиться з ціною. В даному контексті, перекуплений ринок – це перегрітий покупками цінних паперів ринок, простими словами, «бульбашка», що скоро лусне, а перепроданий – навпаки ринок з домінуючим нисхідним трендом, але з потенціалом «вистрілити».

Коли CCI переходить з від'ємного або близького до нуля значення до рівня вище 100, це може означати, що ціна починає новий висхідний тренд. Як тільки це відбувається, трейдери можуть спостерігати за відкатом ціни, за яким слідує узгодженість напрямків як ціни, так і CCI, що сигналізує про можливість купівлі.Ця ж концепція застосовується і до спадного тренду, що зароджується. Коли індикатор переходить від позитивних або близьких до нуля значень до рівня нижче -100, це може означати, що починається спадний тренд. Це сигнал, щоб вийти з довгих позицій або почати стежити за можливостями для відкриття позицій на продаж [36].

Індекс грошового потоку (MFI) - це функція осцилятор, який використовує дані про ціну та обсяг для виявлення сигналів перекупленості або перепроданості активу. Він також може бути використаний для виявлення розбіжностей, які попереджають про зміну тенденції в ціні.

Показник MFI вище 80 вважається перекупленим, а показник MFI нижче 20 вважається перепроданим, хоча рівні 90 і 10 також використовуються як порогові значення.

Варто звернути увагу на розбіжність між індикатором і ціною. Наприклад, якщо індикатор зростає, а ціна падає або є сталою, ціна може почати зростати [37].

Стохастичний RSI або StochRSI — це індикатор, головна функція якого є визначення того, чи даний актив є перекупленим або перепроданим. Варто зазначити, що даний індикатор є стохастичностою версією індексу відносної сили (RSI), а отже є осцилятором, що означає наявність коливань навколо центральної лінії [38].

Для розрахунку стохастичного RSI враховується ціна закриття актива, а також найвища та найнижча ціна протягом певного періоду (зазвичай, 14 днів). Однак, коли формула використовується для розрахунку StochRSI, вона безпосередньо застосовується до значень RSI (сама по собі ціна не враховується) [38].

# РОЗДІЛ 2 УРАХУВАННЯ УПЕРЕДЖЕНОСТІ В СИСТЕМАХ ГЛИБИННОГО НАВЧАННЯ ПРИ ВИРІШЕННІ ЗАДАЧІ ОПТИМІЗАЦІЇ ІП

## 2.1 Проблема упередження при інвестиційних рішеннях

Поведінкові упередженості – це одна із складових людської природи. При прийнятті фінансових рішень, інвестори несвідомо або свідомо приймають рішення з емоційним забарвленням, що безпосередньо впливає на дохідність портфелю цінних паперів. Коли інвестори припускаються упереджених дій, вони не визнають доказів, які суперечать їхнім припущенням. Інвесторам слід усвідомити необхідність уникання двох основних типів упередженості: емоційного та когнітивного. Урахування цих факторів дозволяє інвестору на основі наявних даних прийняти зважене рішення.

Тому важливо розуміти та уникати поведінкових упереджень у прийнятті інвестиційних рішень.

До найбільш розповсюджених поведінкових упереджень відносяться наступні:

1) Ефект надмірної самовпевненості (Overconfidence bias) - це схильність до хибної та оманливої оцінки власної експертизи.
2) Ефект почуття жалю (Regret aversion bias) – це схильність до страху втраченої можливості. Інвестори можуть зрештою витратити більше грошей і часу, ніж потрібно з точки зору альтернативної вартості, щоб уникнути відчуття «втраченої можливості» про те, що вони прийняли неправильне рішення. Таким чином, це може перешкодити інвесторам прийняти правильне рішення в потрібний час.

3) Ефект сліпого слідування тренду (Trend-chasing bias) – це схильність інвестора приймати рішення на основі історичних даних. Незважаючи на те, що минулі показники не обов'язково відображають майбутні результати, багато інвесторів приймають інвестиційні рішення переважно на основі минулих доходів. Ця тенденція може призвести до неправильного інвестиційного рішення, якщо є будь-який фактор, який інвестори не врахували.
4) Ефект самопідтвердження (Confirmation bias) - це схильність інвестора вірити та шукати інформацію, яка підтверджує власні переконання, ігноруючи об'єктивну дійсність. Через упередженість підтвердження інвестор може знехтувати інформацією, яка не відповідає його переконанням.
5) Ефект уникання втрат (Loss aversion bias) – це схильність інвестора віддавати перевагу уникненню втрат, а не пошуку прибутку. Уникнення цієї упередженості може окупитися в майбутньому, якщо інвестиції зроблені усвідомлено.
6) Ефект настрою (Mood bias) оптимізм (або песимізм) і надмірна впевненість додають ірраціональності та емоційності процесу прийняття рішень.
7) Ефект упередженості ЗМІ (Media bias and Internet information bias) – це схильність до некритичного прийняття інвесторських рішень на основі широко поширених думок і припущень.

Пости в соціальних мережах важливі як для розуміння людського сприйняття, так і для того, щоб вони могли впливати на людське сприйняття. З цієї точки зору соціальні мережі можуть мати певний вплив на поведінку людей у процесі прийняття рішень, і ця ситуація також може проявлятися у фінансових рішеннях. Динаміка, створена взаємним впливом

соціальних мереж і фінансових ринків, стала важливим предметом дослідження.

Shen, Urquhart, & Wang (2019) [13] провели дослідження, в якому досліджували зв'язок між інтересом інвесторів і доходністю біткойнів, обсягом торгів і реалізованою волатильністю. У цьому дослідженні, яке проводилося за допомогою лінійних і нелінійних тестів причинності Грейнджера, було зроблено висновок, що кількість твітів була важливою рушійною силою обсягу торгів наступного дня та фактичної волатильності, що підтверджено лінійними та нелінійними тестами причинності Грейнджера. Останнім часом було проведено багато досліджень щодо кореляції між висловлюваннями в Twitter і ринком криптовалют. Kraaijeveld, & De Smedt (2020) [14] дійшли висновку у своїй статті, Media bias, що має місце серед користувачів Twitter може вплинути на прибутковість Bitcoin, Bitcoin Cash і Litecoin за допомогою аналізу настроїв на основі семантичного аналізу твітів. Занг (2020) [15] виявив, що в той час як вартість криптомонет зросла у відповідь на настрої Twitter, обсяги торгів зросли у відповідь на абсолютну вартість. Згідно з результатами роботи Аарона, Деміра, Лау та Заремби (2020) [16], існує чітка причинно-наслідкова кореляція між неоднозначністю публікацій у соціальних мережах і дохідністю криптовалюти Bitcoin, Ethereum, Bitcoin Cash і Ripple.

У своїй статті Öztürk і Bilgin (2021) [17] досліджували, чи твіти впливають на прибуток біткойнів або зміни обсягу торгів з точки зору важливих облікових записів Twitter. Результати дослідження показують, що твіти можна використовувати для прогнозування доходів біткойнів, і, зокрема, найвпливовіші облікові записи, а не всі твіти, є драйверами цих доходів. У своєму дослідженні Анте (2021) ділиться результатом того, що після кожного допису Ілона Маска в Твіттері про криптовалюти виявляється дуже важливий аномальний обсяг транзакцій і прибутковість біткойнів і доджкойнів (Анте, 2021) [18].

Вивчення процесів прийняття рішень людьми в рамках різних факторів, таких як психологічні, соціологічні та когнітивні, сприяє розумінню того, як ірраціональність людської природи може вплинути на індивідуальні інвестиційні рішення та відображення цих ефектів на фінансових ринках.

Вважається, що чутливість людей до соціальних медіа також зростає зі збільшенням їх використання, і з цієї точки зору поширення соціальних мереж може ірраціонально впливати на рішення інвесторів. Загалом, можна вважати, що в цьому процесі є вплив сукупності різних когнітивних упереджень, що може виникати в періоди, коли ринок має як спадний, так і висхідний тренд [19,20].

Але варто, зазначити, що у зазначених вище наукових роботах потенціал використання текстових даних при побудові інвестиційних портфелів дуже обмежено та частково, а саме:

- у роботах [13-16] досліджувались причинно-наслідкові зв’язки між заявами в соціальних мережах та ціною на криптовалюту, проте не враховувались предиктивні здібності текстових даних щодо цін на дані активи

- у дослідженнях [17,18] головна роль приділялась дослідженню впливу постів конкретних осіб в Twitter на прогноз ціни біткоїну.Більш того, у роботі [18] досліджувався вплив лише вплив твітів Ілона Маска.

- у роботах [19,20] основний акцент був спрямований на дослідження впливу когнітивних упереджень на загальний стан ринку, що є достатньо поверхневим дослідженням, оскільки висвітлює патерни та причинно-наслідкові зв’язки між колективними упередженнями та індивідуальними інвестиційними рішеннями.

У свою чергу, у даній роботі пропонується розширити використання текстових даних від дописів окремих інфлуенсерів до висвітлення

суб'єктивних думок широкого спектру індивідуальних інвесторів, а також відомих журналістів. Також, було зміщено акцент дослідження з крипторинку на фондовий ринок.

Наукова новизна поточного дослідження полягає у винайденні алгоритму прогнозування сукупності доходностей цінних паперів – кандидатів на позиції інвестиційного портфелю, опираючись на інструментарій передових методів штучного інтелекту - Алгоритми обробки текстової інформації Word-to-Vec та рекурентних нейронних мереж LSTM.

## 2.2 Алгоритми обробки текстової інформації Word-to-Vec

Вивчення патернів нейронною мережею накладає певні обмеження на вибірку навчальних данних. Одне з ключових – представлення данних у вигляді числових послідовностей. Оскільки інформація, що передає поведінкове упередження базується на текстових данних, то необхідно інтерпретувати текстові дані як числові послідовності. З цією метою було вирішено використовувати Word2vec для навчання представлених слів. Представленням слова тут і надалі називатимемо числову послідовність невеликого розміру, що асоціюється з даним словом. На представлення слів також введемо умову збереження семантичного сенсу. Тобто логічні операції над словами в одному контексті мають відображатися на арифметичних операціях над представленнями цих слів.

Word2vec засвоює значення слів за рахунок обробки великого корпусу немаркованого тексту. Немає потреби маркувати слова у словнику Word2vec. Не треба розповідати алгоритму забезпечення, що Марія Кюрі – вчений, а «Тімберс» – футбольна команда, Сіетл – місто чи Портленд – два міста у штатах Орегон та Мен. І не треба розповідати, що футбол – вид

спорту, команда – група людей чи міста – як географічні, так і демографічні об'єкти. Word2vec може сам навчитися всієї цієї інформації, а також багато іншого! Все, що йому потрібно, — досить великий корпус, у якому Марія Кюрі, «Тімберс» та Портленд згадувалися б поряд з іншими словами, пов'язаними з наукою, футболом та містами відповідно.

Модель Word2vec містить інформацію про взаємозв'язки між словами, у тому числі їх подібність. Модель Word2vec знає, що термін SpaceX знаходиться приблизно на такій же відстані від Elon Musk, що і Tesla від Elon Musk. І ці відстані (різниці між парами векторів) лежать приблизно одному напрямку. Тому можна отримати вектор Tesla, віднявши від вектору SpaceX вектор rocket і додавши вектор automobile.

Існує два шляхи представлення слів що генеруються Word2Vec - CBOW (Continuous Bag of Words) та Continuous Skip-Gram Model.

Почнемо з CBOW. Дана мережа намагається передбачити ключове слово в контексті (оточення даного слова) за допомогою контексту. Таким чином, модель Word2Vec являтиме собою одношарову нейронну мережу з n входами і 1 виходом.

У свою чергу Continuous Skip-Gram Model підхід є інтуітивно оберненим до CBOW. Тобто, маючи ключове слово Word2Vec намагається передбачити слова, що є в контексті з даним ключовим словом.

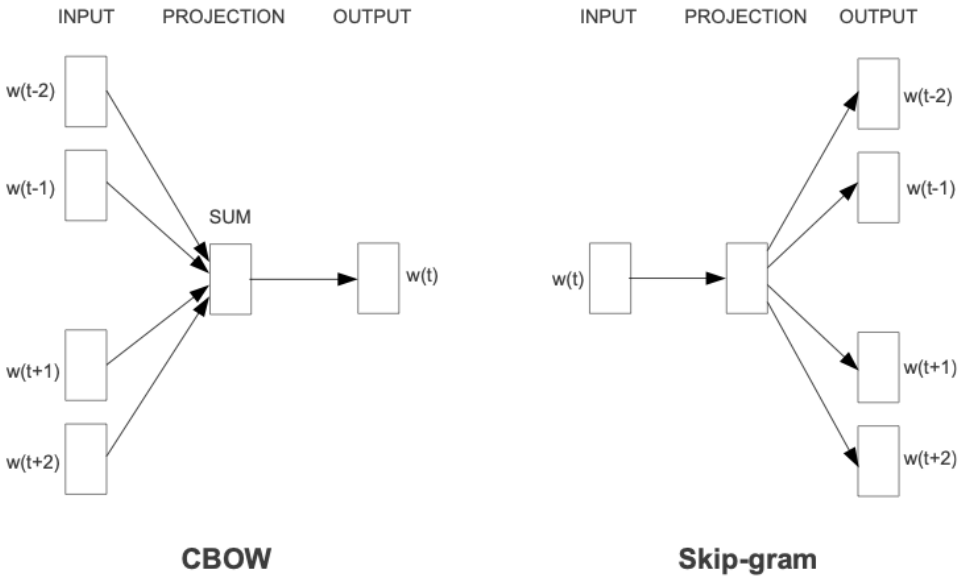


Рисунок 2.1. CBOW та Skip-Gram підходи

*Джерело: Побудовано на основі* [33]

Модель word2vec знайшла широке застовування при вирішенні широкого спектру задач, при чому обидві варіації - Continuous Bag of Words (CBOW) і Continuous Skip-Gram Model (SG) є ефективними архітектурами, які дозволяють нейронним мережам вивчати слова та їхній контекст [44]. Метод CBOW використовує контекст для передбачення наступного слова, а модель SG використовує слово для передбачення контексту [44]. Word2vec використовується в поєднанні з іншими алгоритмами для точної класифікації настроїв.

У роботах Міколова було запропоновано дві архітектури для безперервних векторних представлень великих наборів даних і проаналізував ефективність різних алгоритмів і архітектур [33]. Серед них можна виокремити модель нейронної мережі з упередженим зв’язком (NNLM), модель рекурентної нейронної мережі (RNNLM), паралельне навчання нейронних мереж і нові логарифмічно-лінійні моделі, такі як

CBOW і SG. Їхня робота виявила, що можна навчити високоякісні вектори на простих модельних архітектурах, таких як CBOW або SG, і що завдяки меншій обчислювальній складності цих архітектур можна досягти вищої точності більших наборів даних [33].

У [45] було проаналізувано та було порівняно кілька різних методів навчання та алгоритмів, які були протестовані для класифікації настроїв у даних Twitter. Дані дослідження порівнювали продуктивність ряду алгоритмів машинного навчання, таких як Naive Bayes, Max Entropy, Support Vector Machines, а також інших підходів на основі аналізу лексикографічних данних [45].

Найвищої точності досягається метод опорних векторів (support vector machine method, SVM), який був імплементований в [46]. У цьому експерименті було отримано результат класифікаційного прогнозування 86%, а низка інших алгоритмів – у діапазоні точності 80% [46]. Основний недолік даного дослідження полягає в тому, що два різні дослідження використовували різні набори даних [46]. Рівень точності був би більш показовим, якби все було зроблено з використанням одного набору даних.

У [47] була представлена модель, яка змішує як техніки навчання з учителем, так і навчання без учителя для визначення емоційного забарвлення текстових документів. В рамках данного дослідження було застосувано цю модель до широко перевірених текстових корпусів і змогли добитись більшої точності, ніж методи, представлені раніше [47]. Традиційні стоп-слова, які зазвичай видаляються з мовної обробки, були залишені через те, як вони допомагають передати емоційне забарвлення[47]. Це також справедливо для знаків пунктуації (таких як «!» і «:-»). [47]. Автори взяли свій набір даних із попередньої роботи, наведеної в [46], і використали отримані результати в якості орієнтиру. У [47] було правильно класифікувано 88% документів. Важливим внеском цього дослідження є класифікація на рівні речень, яка має відношення до

задачі класифікації повідомлень у Twitter, вирішення якої є пріоритетним в рамках потоного дослідження.

У дослідженні [48], був продемонстрований інший підхід, оскільки було видалено всі смайлики та знаки пунктуації  під час навчання своїх алгоритмів. У даній роботі показано, що видалення несловесних токенів дозволило класифікаторам зосередитися на емоційному забарвленні кожного слова, тоді як їхнє включення мало негативний вплив на метрики двох  алгоритмів, які були досліджені, і незначний вплив на третій [48]. Після навчання моделей, описаних вище, були застосовані алгоритми Naive Bayes, Maximum Entropy і Support Vector Machine, щоб класифікувати настрої Twitter з точністю в діапазоні 80% [48], що відповідає результатам роботи, проведеної у [47]. Найкращий результат, 83%, було досягнуто з класифікатором максимальної ентропії при використанні як уніграм, так і біграм. Проте всі результати були в межах трьох відсоткових пунктів, тому жодна комбінація класифікатора та ознаки не показала себе набагато гірше [48].

Хештеги — ще один вид токенів, схожих на смайли, який є особливою рисою для текстів соціальних мереж. У  [49] було створено набір даних, який включав хештеги (наприклад, #bestfeeling, #epicfail, #news), щоб підвищити точність аналізу емоційного забарвлення. У їх тренувальному наборі данних було використано різні види хештегів, що дослідники вважали позитивними, негативними чи нейтральними, наприклад #iloveitwhen, #thingsilove, #success, #worst, #itsnotok, #ihate, з метою покращення способу збору та класифікації вхідних даних [49]. У рамках даного дослідження було протестовано включення смайлів у порівнянні з набором даних тільки хештегів і виявили, що включення смайлів не показало значного покращення порівняно з окремими хештегами [49].

У [50] була реалізована модель на основі word2vec для класифікації тексту. Була перевірена модель на основі word2vec з моделлю, що використовувала алгоритм tf-idf, що базується на інформації про частоту слова в тексті порівняно до інверсної частоти документа, і змогли показати, що комбінація word2vec з tf-idf покращує базовий алгоритм tf-idf сам по собі [50]. Також, у даному дослідженні порівнювали продуктивність з урахуванням стоп-слів і без них, а також із різними зваженими комбінаціями цих двох підходів. Слід зазначити, що окрім порівняння алгоритмів між собою, у рамках дослідження було вивчено залежність між метриками якості моделі та кількості категорій класифікації даних. Було встановлено, що для 2 категорій метрика точності складала 85%, а зі збільшенням категорій до 4 точність падала до 65% [50].

У дослідженні [51] вивчались задачі аналізу емоційного забарвлення для інших соціальних мереж з метою порівняння ефективності алгоритму TF-IDF і алгоритму TF для класифікації тексту з векторною моделлю з шістьма кортежами на основі ознак у поєднанні з High Adverb of Degree Count (HADC). Було встановлено, що модель із шістьма кортежами та алгоритм зважування HADC у поєднанні змогли точно класифікувати від 88% до 92% текстових даних [51]. Необхідно провести додаткові дослідження, щоб зробити висновок, чи є це найефективнішим алгоритмом для аналізу настрою, але очевидно, що ця модель забезпечує високий рівень точності. Ця модель ще не перевірялася на англійському тексті.

У [52] освітлені нові перспективи для використання алгоритму наївного Байєса. Незважаючи на те, що метод Наївного Байєса не такий ефективний, як багато інших алгоритмів, проте він може бути надзвичайно точним, якщо його використовувати разом у поєднанні з іншими методами [52]. Було запропоновано двоетапну ієрархію вибору ознак, яка базується

на пошуку найбільш використовуваних слів у наборах даних з подальшим групуванням для зменшення розмірності простору ознак [52]. Дослідження вивчало вибір «важливих слів» на основі значення статистики хі-квадрат і визначило ці слова як основу за допомогою матриці входжень слів в документ [52]. Цей запропонований метод покращив продуктивність алгоритму наївного Байєса, зменшив набір ознак ефективніше, ніж одновимірний хі-квадрат, і є більш ефективним ніж традиційні методи на основі кореляції[52].

## 2.3 Глибинне навчання у задачах оптимізації ІП

Теорія оптимізації портфеля, представлена Марковіцем, відіграє значну роль як у дослідженнях, так і в практиці. Базова модель MV (Mean-Variance) розглядає лише довгі позиції (long positions) та вводить обмеження на параметри моделі; наприклад, сума ваг дорівнює 1. Її можна сформулювати як задачу квадратичного програмування [21], і рішення залежить лише від очікуваного середнього та коваріаційної матриці прибутковості активів. Найпростішим способом оцінки очікуваного середнього та коваріаційної матриці є використання вибіркового середнього та коваріаційної матриці. Однак невідомо, чи вибіркові оцінки є репрезентативними з точки зору генеральної сукупності. Простими словами, невідомо, чи оцінки прибутковостей та волатильностей, отримані з наявних даних, відображають реальний потенціал активів, що цікавлять інвестора. Дана проблема відома як «загадка оптимізації Марковіца», описана в [22, 23, 24].

Щоб розв’язати «загадку оптимізації Марковіца», основним способом вирішення є пошук кращих оцінок очікуваних середніх, наприклад,

застосовуючи аппарат Байєсівської статистики [25, 23, 26]. Однак динаміка фінансових прибутків є досить стохастичною, як і прогноз прибутковості акцій. Також існує велика кількість літератури, в якій використовуються моделі машинного навчання для прогнозування прибутковості активів, у тому числі мережі довготривалої короткочасної пам'яті (LSTM) [29], та ін. Крім того, оцінка коваріаційної матриці є нестабільною, коли наявна велика кількість змінних [27, 28].

Також слід зазначити роботу Zohren et al. (2020) [30], де різні нейронні мережі навчаються оптимізувати портфельний Sharpe Ratio. З-поміж різних мережевих архітектур вони визнали модель довгострокової пам'яті (LSTM) найефективнішою. Ціна активу та прибутковість із поточними та останніми 50-денними значеннями подаються на вхід в модель, і вони повідомляють про перенавчання(overfitting) з мережею прямого розповсюдження через велику кількість параметрів. Батлер і Квон (2021)[31] використовують нейронні мережі з диференційованими рівнями оптимізації, щоб знайти рішення там, де аналітичне рішення не існує. Реальні емпіричні тести показують переваги в продуктивності в порівнянні з двоетапними моделями, де традиційний підхід полягав у виконанні лінійної регресії, а потім оптимізації портфеля.

Рекурентні нейронні мережі (Simple RNN) є сімейством нейронних мереж для обробки послідовних даних. Рекурентні мережі можуть масштабуватися до набагато більших послідовностей, ніж це було б практично для мереж без спеціалізації на основі послідовностей. Більшість рекурентних мереж також можуть обробляти послідовності змінної довжини.

Нехай матимемо часовий ряд

$$Y = \{y(t), t \in \mathbb{Z}\}$$

Тоді можна скласти такий навчальний набір таким чином

$$D = \left\{d_i \mid d_i = \langle \overrightarrow{X_i}, y_{t-i} \rangle\right\}$$
$$\overrightarrow{X_i} = \left\{y_{t-i-1}, y_{t-i-2}, \cdots, y_{t-i-p}\right\}$$

Отже, була поставлена задача навчання з учителем, яка полягає в передбаченні значень наступного кроку даної послідовності при наявності p попередніх значень. Цю задачу можна вирішити за допомогою методів машинного навчання. Один з найефективніших фреймворків для прогнозування та обробки послідовностей є рекурентні нейронні мережі (RNN).

Уведемо позначення: W - матриця ваг, яка відповідає за зв'язки між прихованими станами мережі; U - матриця ваг для вхідного вектора; V - матриця ваг для виходів. Тоді, отримуємо таке формальне визначення рекурентної мережі:

$$a_\tau = b + Ws_{\tau-1} + Ux_\tau, \quad (2.1)$$

$$o_\tau = c + Vs_\tau, \quad (2.2)$$

$$s_\tau = f(a_\tau), \quad (2.3)$$

$$y_\tau = h(o_\tau), \quad (2.4)$$

де $b, c$ - вектори зміщення;

$f, h$ - функції активації вхідного та вихідного шарів відповідно [32].

З системи (2.1)-(2.4) видно, що мережі мають цікаву особливість, відмінну від мереж прямого розповсюдження, наприклад, багатошарового

перцептрона. У рекурентних мережах зв'язки між нейронами можуть йти не тільки послідовно від попереднього шару до наступного, але також можуть повертатися "до самого себе", конкретніше, до свого стану в попередній момент часу.

У зв'язку з необхідністю моделювання послідовностей, для даної мережі можна використовувати функцію помилок у вигляді від'ємної логарифмічної правдоподібності (neg-log-likelihood), яку необхідно мінімізувати:

$$L = \sum_{i} L_{t=i} = -\sum_{i=0}^{t-p} \log\left(p\left(y_{t-i} | y_{t-i-1}, y_{t-i-2}, \dots, y_{t-i-p}, \vec{\theta}\right)\right), \tag{2.5}$$

$$\hat{y}_{t-i} = \arg\min_{\vec{\theta}} L_i(\vec{\theta}) \tag{2.6}$$

де $\vec{\theta} = \langle W, V, U, b, c, o, s \rangle$.[32]

Усі рекурентні мережі, які ми розглядали досі, мають «причинну» структуру, тобто стан у момент часу фіксує лише інформацію з минулого, x(1), . . . , x(t−1) і поточний вхід x(t). Деякі моделі рекурентних мереж також дозволяють інформації з минулих значень ряду впливати на поточний стан, коли вони доступні.

Однак у деяких випадках ми хочемо вивести прогноз y(t), який залежить від усієї вхідної послідовності. Наприклад, під час розпізнавання мовлення правильна інтерпретація поточного звуку як фонеми може залежати від кількох наступних фонем через співартикуляцію та може навіть залежати від кількох наступних слів через лінгвістичні залежності між сусідніми словами: якщо існує дві інтерпретації поточні слова, які є акустично вірогідними, нам, можливо, доведеться зазирнути далеко в майбутнє (і минуле), щоб усунути їх двозначність. Це також стосується

розпізнавання рукописного тексту та багатьох інших задач обробки послідовностей.

Двонаправлені рекурентні нейронні мережі були винайдені для задоволення цієї потреби (Schuster and Paliwal, 1997). Вони були надзвичайно успішними (Graves, 2012) у програмах, де виникає така потреба, наприклад, розпізнавання рукописного тексту (Graves та ін., 2008; Graves та Schmidhuber, 2009), розпізнавання мовлення (Graves та Schmidhuber, 2005; Graves та al., 2013) та біоінформатики (Baldiet al., 1999) [32].

## 2.4 Метрики оцінки моделей глибинного навчання

Для використання побудованих моделей глибинного навчання треба впевнитись в якості прогнозів в подальшій перспективі. Оскільки навчання нейронних мереж базується на принципі вдосконалення зворотного зв’язку, для оцінки якості моделі використовують різноманітні метрики. Вибір метрики залежить передусім від специфіки поставленої задачі. Зазвичай, при аналізі фінансових часових рядів використовують наступні метрики оцінки моделей регресії:

1. Root Mean Squared Error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\mathrm{Predicted} - \mathrm{Actual_i})^2}{N}}$$

RMSE є найпопулярнішим показником оцінки, який використовується в задачах регресії. Це випливає з припущення, що помилки є незміщеними та мають нормальний розподіл.

2. R-Squared/Adjusted R-Squared

$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}}$$

$$\frac{\text{MSE(model)}}{\text{MSE(baseline)}} = \frac{\sum_i^N (y_i - \hat{y}_i)^2}{\sum_i^N (\bar{y}_i - \hat{y}_i)^2}$$

Ця метрика, опираючись на дослідження дисперсії ряду, що вивчається, показує, наскільки добре модель пояснює вхідні дані.

3. Mean absolute percentage error (MAPE)

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{\text{Actual}_t - \text{Predicted}_t}{\text{Actual}_t} \right|$$

MAPE показує середню абсолютну похибку в процентах, що допомагає дізнатися, наскільки відносно модель добре прогнозує ряд, що вивчається.

4. Mean Absolute Error (MAE)

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

MAE вимірює середню абсолютну величину помилок у наборі прогнозів. Це медіанне значення за тестовою вибіркою абсолютних відмінностей між прогнозом і фактичним спостереженням, де всі індивідуальні відмінності мають однакову вагу.

Оскільки в контексті поставленої задачі існує проблема оцінки емоційного забарвлення інформації, що виражає поведінкові упередження. Зрозуміло, що дана задача відноситься до задач класифікації, а отже, для оцінки моделі доцільно також використати наступні метрики оцінки класифікацій:

1. F1

$$F_1 = \left(\frac{\text{recall}^{-1} + \text{precision}^{-1}}{2}\right) = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

2. AUC-ROC

3. Recall

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

4. Precision

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

## 2.5 Архітектура обраної нейронної мережі

З огляду на специфіку поточної задачі, було вирішено збудувати модель, що могла б сприймати часові ряди різної довжини, а також вміти аналізувати текстову інформацію. Тому було вирішено створити 2 окремі моделі, які згодом будуть інтегровані в одну нейронну мережу.

Перша модель має аналізувати емоційне забарвлення текстової інформації, а друга у свою чергу має прогнозувати задані часові ряди. Фінальна модель має корегувати прогнози другої моделі за допомогою оцінок емоційного забарвлення упереджень щодо досліджуваного ряду.

Було вирішено, що перша модель буде імплементована як класифікатор в основі якого лежить Word2Vec модель для навчання числовим представленням слів та LSTM(Довга короткострокова пам'ять) для аналізу сукупностей цих представлень (репрезентації тексту). Великою перевагою такої мережі є здатність не лише розширювати та перезаписувати "пам'ять" (як у звичайних рекурентних мережах), але й використовувати будь-яку лінійну комбінацію значень "пам'яті" з різних часових проміжків. Ця особливість архітектури дозволяє мережам дуже

гнучко запам'ятовувати закономірності. Справді, ускладнена архітектура LSTM дозволяє вловлювати тренди довільної довжини, що було неоднократно підтверджено на практиці.

Для другої моделі було вирішено використовувати каркас з LSTM нейромереж, кожна з яких оброблює ознаку часового ряду котирувань (ціна відкриття на поточний момент часу,закриття на поточний момент часу,об’єм торгів, найвища ціна позиції, найнижча ціна позиції)

Для прогнозування майбутньої прибутковості прогнозні значення, отримані з каркасу моделей описаних вище зважуються за допомогою нейронної мережі прямого розповсюдження. Далі, до зваженного прогнозу додається індикатор семантичного забарвлення і виконується операція корекції прогнозу. Згодом, отримані значення прогнозу для кожного з активів, що треба включити в інвестиційний портфель, використовуються для отримання ваг портфелю за допомогою максимізації Sharpe Ratio. Результативність портфелю вимірюється шляхом підстановки істиних значень прибутковостей за аналогічний проміжок часу.

# РОЗДІЛ 3 СТВОРЕННЯ ТА АНАЛІЗ ІНВЕСТИЦІЙНОГО ПОРТФЕЛЮ НА ОСНОВІ ДАНИХ БІРЖІ NASDAQ ТА СОЦІАЛЬНОЇ МЕРЕЖІ TWITTER

## 3.1. Обгрунтування вибору наборів даних

Фондова біржа - це місце, де можна здійснювати операції з різними фінансовими інструментами, такі як акції, облігації, ф'ючерси тощо. Торгові майданчики цього типу є важливими індикаторами стану національної та світової економіки. Фінансовий ринок є важливим механізмом, що забезпечує збільшення зайнятості та доходів населення. Крім того, успіх фінансового ринку є показником загального рівня довіри до економічної ситуації в країні та світі.

Раніше торгові платформи були розміщені у фізичних приміщеннях, але з розвитком інтернет-технологій вони перейшли до онлайн-режиму. Проте це не мало впливу на їх капіталізацію, і багато з них активно працюють дотепер. Найбільші з таких платформ мають ринкову вартість в трильйонах доларів.

За показниками капіталізації на даний момент однією з найбільших та найвпливовіших є фондова біржа NASDAQ. Ця платформа є однією з трьох головних фінансових бірж у США, разом з NYSE та AMEX, і була створена в 1971 році [2]. Вона є першою у світі електронною біржею і ніколи не мала фізичного представництва. На цій торговій платформі акціями торгують більше 3 тис. компаній, головним чином з технологічної сфери, таких як Apple (APPL), Microsoft (MSFT), Facebook (FB) та Tesla (TSLA). Серед зведених торгових індексів, що розраховуються на біржі, варто відзначити NASDAQ Composite та NASDAQ National Market Composite index.

Варто зазначити, що біржа NASDAQ користується великим попитом у трейдерів, що займаються алгоритмічною торгівлею. Таким чином, об’єм

транзакцій, що проходять через дану біржу, є більш ніж достатнім для ретельного аналізу обраних акцій та синтезу прибуткового потрфелю в перспективі. В еру домінування високочастотного трейдингу, стохастична природа фондового ринку є важкопрогнозованою, тобто маючи інформацію лише про фінансові показники певного лістингу неможливо спрогнозувати його рух в довгостроковій перспективі. На стан ринку також впливають настрої трейдерів, що часто залежать від інформаційної повістки. Це також відомо, як media bias, або упередження ЗМІ. У зв'язку з бурхливим розвитком інтернету та його доступності для широкого кола осіб у 21 столітті соціальні мережі витісняють нішу, що раніше була зайнята традиційними ЗМІ, але на відміну від останніх, фільтрування інформації в соціальних мережах є незрівняно меншим. Тому змішення об'єктивної інформації та суб'єктивних тверджень може мати непередбачуваний вплив на ринок цінних паперів. З урахуванням вищесказаного, необхідно доповнювати наявні дані показників з фондових бірж інформацією з найвпливовіх цифрових медіа.

Враховуючи вищесказане, було прийняте рішення в рамках даного експерименту взяти дані на тематику фондового ринку з соціальної мережі Twitter [39] та поєднати їх з показниками відповідних цінних паперів з біржі NASDAQ за аналогічний період.

## 3.2. Постановка експерименту

В якості досліджуваних даних фондової біржі nasdaq з метою конструювання інвестиційного портфелю були обрані наступні лістинги:

- 'AAPL' – Apple Inc.
- 'TSLA' - Tesla Inc.
- 'MSFT' – Microsoft

- 'AMZN' - Amazon
- 'GOOG' – Alphabet Inc.
- 'V' – VISA
- 'NVDA' – Nvidia

Було вирішено використовувати дані тікери, зважаючи на високу капіталізацію, значні показники прибутковостей та стійкість компаній. Останній фактор полягає у швидкій відновлювальності ключових фінансових показників технологічних компаній після криз та потрясінь [40], таких як криза dot-com, пандемія COVID-19 [41, 42] та глобальної економічної рецесії 2022-2023 років, спричинених вторгнення рф на Україну [43].

У якості періоду стостережень було обрано період останнього циклу економічного зростання у світі 2015-2019 та початок рецесії, спричинений пандемією COVID-19 – початок-середина 2020 року.

У якості текстових даних з метою покращення результатів конструювання інвестиційних портфелів були обрані твіти як світових видань, серед яких Financial Times(FT), Bloomberg, NY Times, лідерів думок(інфлуенсерів) (наприклад, твіти Ілона Маска), так і твіти блогерів меншого масштабу за аналогічний період 2015-2020. Подібне різномаїття вибору містить в собі ціль дослідити як емоційні забарвлені фрази інфлуенсерів впливають на флуктуацію ринку.

В якості показників, що беруться до уваги використовуємо наступні :

1) Ціна відкриття позиції (Open)
2) Ціна закриття позиції (Close)
3) Об’єм торгів (Volume)
4) Ціна найвищої позиції (High)
5) Ціна найнижчої позиції (Low)

Наведемо приклад візуалізації одного з тікерів, а саме MSFT на рисунку 3.1.

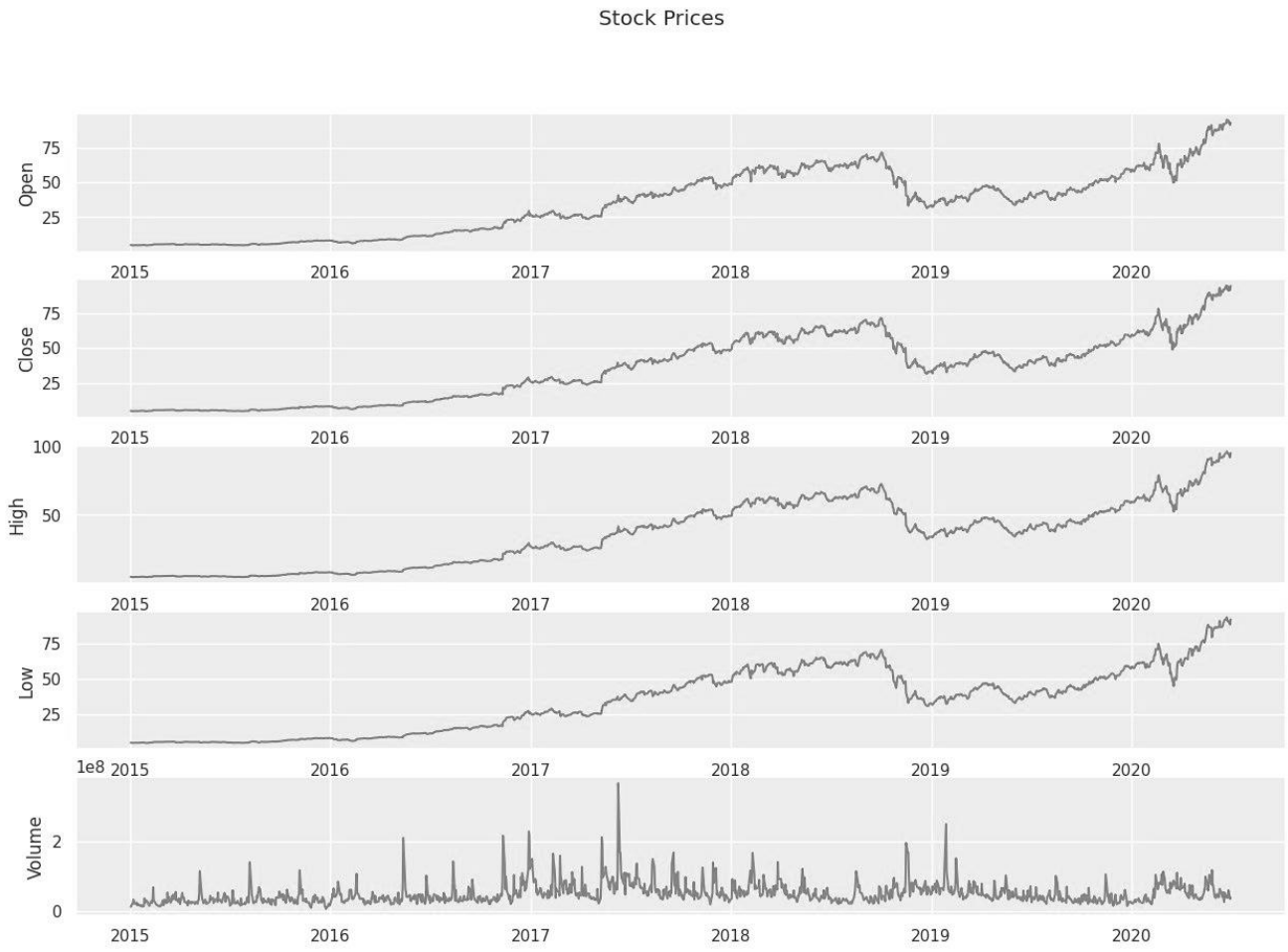


Рисунок 3.1. Показники акцій Microsoft у період 2015–2020 роки

*Джерело: Побудовано автором на основі джерел [54]*

Наявні дані з Twitter мають наступні властивості:

- tweet_id – ідентифікатор твіту
- writer – автор твіту
- post_date – час створення твіту
- body – зміст твіту
- comment_num – кількість коментарів під твітом
- retweet_num – кількість ретвітів
- like_num – кількість лайків
- ticker_symbol – символ акції

Приклад текстового набору даних зображений на рисунку 3.2.

| | tweet_id | writer | post_date | body | comment_num | retweet_num | like_num |
|---|---|---|---|---|---|---|---|
| 0 | 550441509175443456 | VisualStockRSRC | 1420070457 | lx21 made $10,008 on $AAPL -Check it out! htt... | 0 | 0 | 1 |
| 1 | 550441672312512512 | KeralaGuy77 | 1420070496 | Insanity of today weirdo massive selling. $aap... | 0 | 0 | 0 |
| 2 | 550441732014223360 | DozenStocks | 1420070510 | S&P100 #Stocks Performance $HD $LOW $SBUX $TGT... | 0 | 0 | 0 |
| 3 | 550442977802207232 | ShowDreamCar | 1420070807 | $GM $TSLA: Volkswagen Pushes 2014 Record Recal... | 0 | 0 | 1 |
| 4 | 550443807834402816 | i_Know_First | 1420071005 | Swing Trading: Up To 8.91% Return In 14 Days h... | 0 | 0 | 1 |
| 5 | 550443808606126081 | aaplstocknews | 1420071005 | Swing Trading: Up To 8.91% Return In 14 Days h... | 0 | 0 | 1 |
| 6 | 550443809700851716 | iknowfirst | 1420071005 | Swing Trading: Up To 8.91% Return In 14 Days h... | 0 | 0 | 1 |
| 7 | 550443857142611968 | Cprediction | 1420071016 | Swing Trading: Up To 8.91% Return In 14 Days h... | 0 | 0 | 1 |
| 8 | 550443857595600896 | iknowfirst_br | 1420071017 | Swing Trading: Up To 8.91% Return In 14 Days h... | 0 | 0 | 1 |

Рисунок 3.2. Набір текстових даних із Twitter у період 2015–2020 роки

*Джерело: Побудовано на основі [39]*

У рамках даної роботи було прийнято рішення розробити три моделі:

- **перша**: для оцінки емоційного забарвлення твітів (див. рисунок 3.3)
- **друга**: для оцінки прибутковостей акцій в інвестиційному портфелі без урахування упередженостей (bias) (див. рисунок 3.4)
- **третя**: для оцінки прибутковостей акцій в інвесиційному портфелі з урахуванням упередженостей (bias) (див. рисунок 3.5)

Наведемо архітектури моделей нижче.

Тут і надалі Embedding позначає шар генерацій вкладень текстової інформації, Conv 1D – згортковий шар, Avg Pooling – шар пулінгу (усереднення вагів моделі по певному вікну), Dense - повнозв’язний шар, LSTM – шар довгої короткотрокові пам’яті.

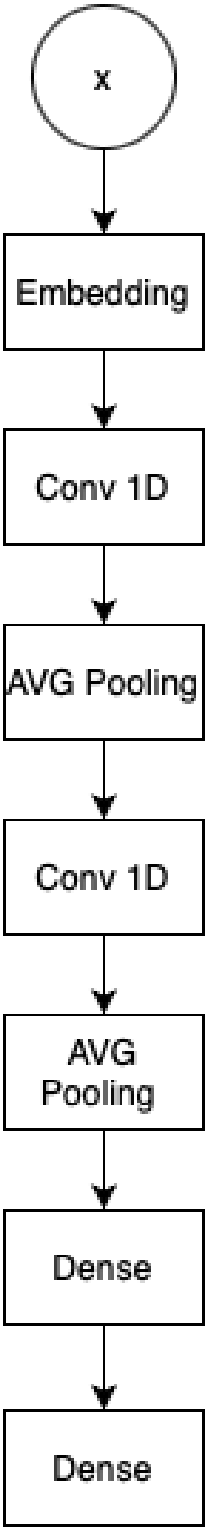


Рисунок 3.3. Архітектура першої моделі

*Джерело: Побудовано автором*

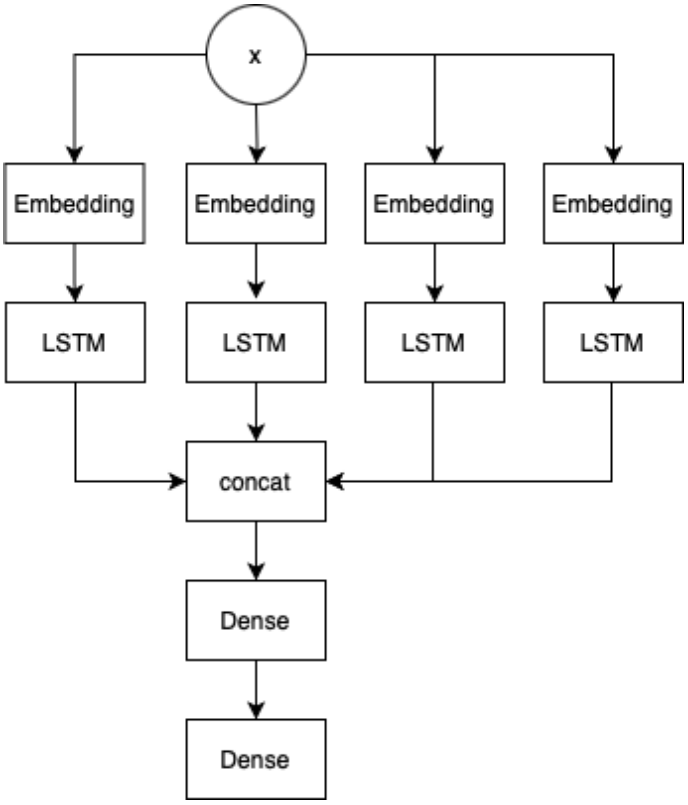


Рисунок 3.4. Архітектура другої моделі

*Джерело: Побудовано автором*

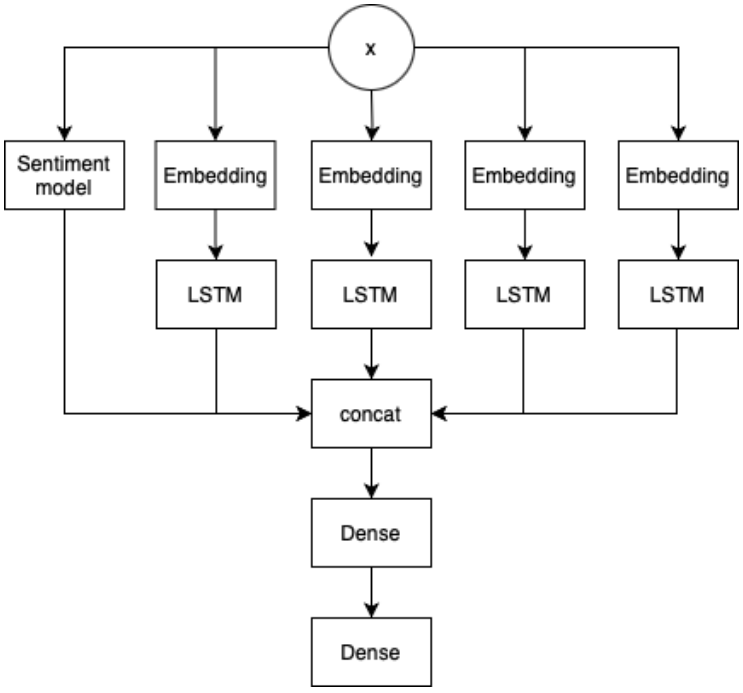


Рисунок 3.5. Архітектура третьої моделі

*Джерело: Побудовано автором*

Далі необхідною є задача конфігурації гіперпараметрів мереж, описаних вище. Гіперпараметри в даному контексті означають розміри шарів запропонованих нейронних мереж, крок навчання мереж і кількість епох навчання. В рамках есперименту були обрані наступні показники:

**Модель 1**:

1) кількість шарів (фільтрів) згортки – 64
2) розмір ядра згортки – 3
3) розмір першого повнозв’язного шару – 64
4) розмір останнього повнозв’язного шару – 3 (кількість класів)
5) крок навчання - 0.001
6) кількість епох – 50

Оцінка моделі проводиться за метриками precision, recall та f1.

**Модель 2**:

1) кількість комірок кожного з шарів довгої короткострокової пам’яті (LSTM) – 64
2) розмір першого повнозв’язного шару – 128
3) розмір останнього повнозв’язного шару – 7 (кількість тікерів)
4) крок навчання - 0.001
5) кількість епох – 100

Оцінка моделі проводиться за метриками середньоквадратичного відхилення, абсолютного відхилення та кореня середньоквадратичного відхилення.

**Модель 3**:

1) кількість комірок кожного з шарів довгої короткострокової пам’яті (LSTM) – 64
2) розмір першого повнозв’язного шару – 128
3) розмір останнього повнозв’язного шару – 7 (кількість тікерів)
4) крок навчання - 0.001
5) кількість епох – 100

6) моделі аналізу тексту – *Модель 1*

Оцінка моделі проводиться за метриками, аналогічними до моделі 2.

Після наведення опису моделей та набору даних необхідно виконати низку трансформацій над данними для приведення їх до прийнятного виду для навчання нейронних мереж.

Для першої моделі текстові дані були оброблені методом регулярних виразів з метою видалення пунктуацій, приведення тексту до нижнього реєстру, а також очистка текстку від «емодзі» символів та хештегів. Також за допомогою бібліотеки nltk (natural language tool kit) були відфільтровані стоп-слова (сполучники, частки), що не несуть ніякого змістового навантаження. Також приведення іменників та дієслів до нейтрального стану, а саме іменники до називного відмінку однини, а дієслів – до інфінитиву, за допомогою операції лематизації.

У свою чергу обробка фінансових даних включала в себе низки індикаторів, зокрема:

- Balance of Power (BOP)
- Commodity Channel Index (CCI)
- Money Flow Index (MFI)
- Stochastic RSI (SRSI)
- Ковзне середнє (moving average) від кожної з ознак, а саме Open, Close, Volume, High, Low

Окрім обчислення індикаторів було виконано логарифмування всіх базових ознак. У випадку цін дана операція була виконана заради нівелювання експоненційного тренду, а в випадку об’єму торгів для приведення розподілів цього показника з лог-нормального до нормального.

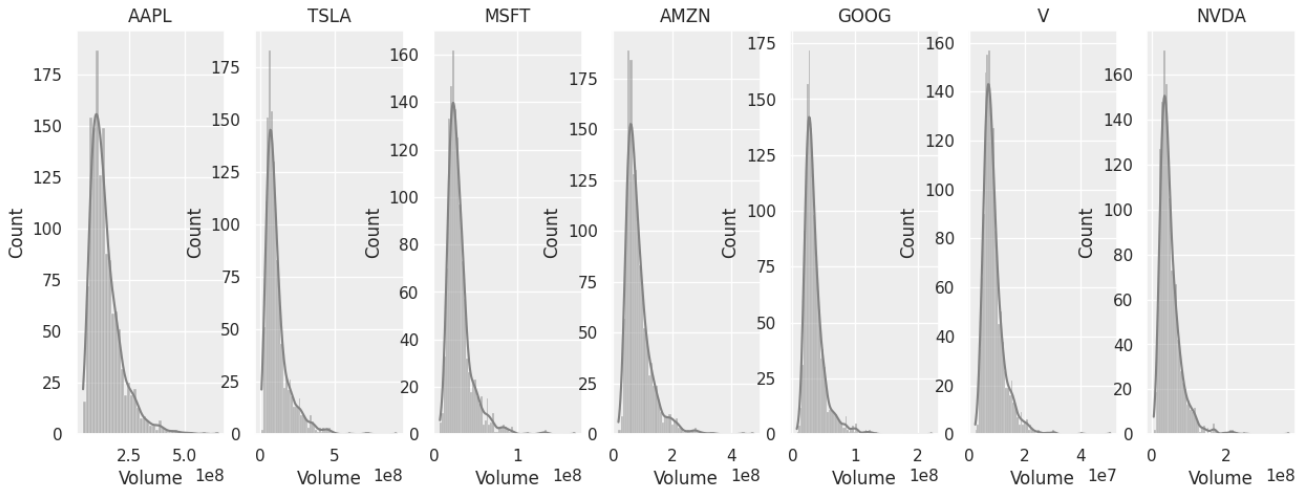


Рисунок 3.6. Лог -нормальний розподіл об’єму торгів

*Джерело: Побудовано автором на основі джерел [54]*

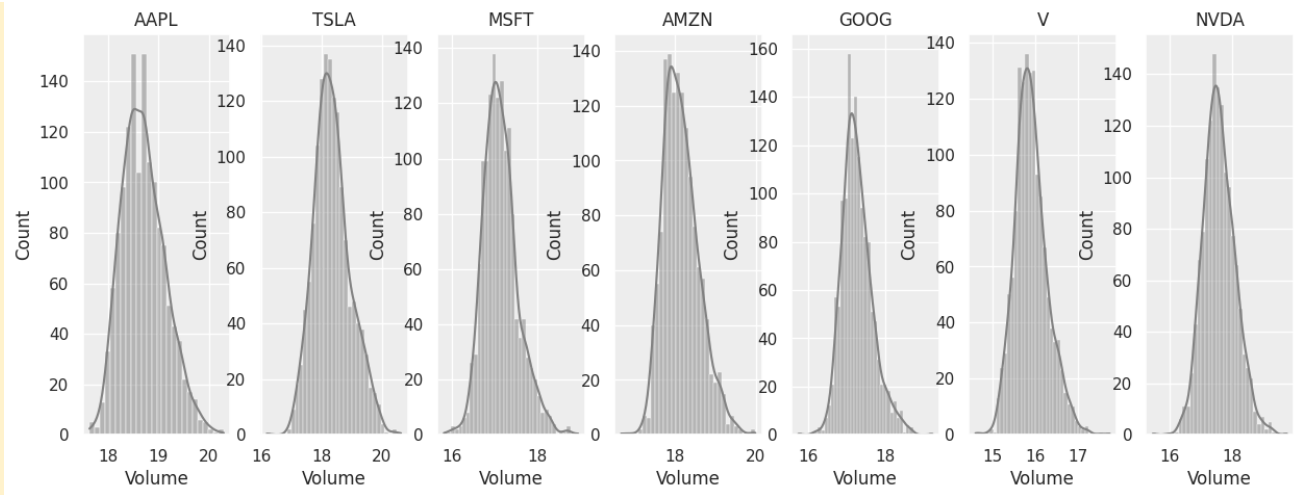


Рисунок 3.7. Нормальний розподіл логарифтму об’єму торгів

*Джерело: Побудовано автором на основі джерел [54]*

У якості одних з результуючих змінних було обрано прибутковість акцій. Варто зазначити, що в якості прибутковостей були взяті прибутковості відносно минулих періодів, так і прибутковості відносно майбутніх періодів. Під минулими періодами розуміємо відношення поточної ціни закриття до ціни відкриття дослідженого періоду , під майбутніми періодами - відношення ціни закриття дослідженого періоду

до поточної ціни відкриття. Для обох показників період становить 2 тижні.

На рисунку 3.8 наведені розподіли прибутковостей минулих періодів за вибраний інтервал дослідження (14 днів). Відповідно на рисунку 3.9 - розподіли прибутковостей майбутніх періодів за вибраний інтервал дослідження (14 днів).

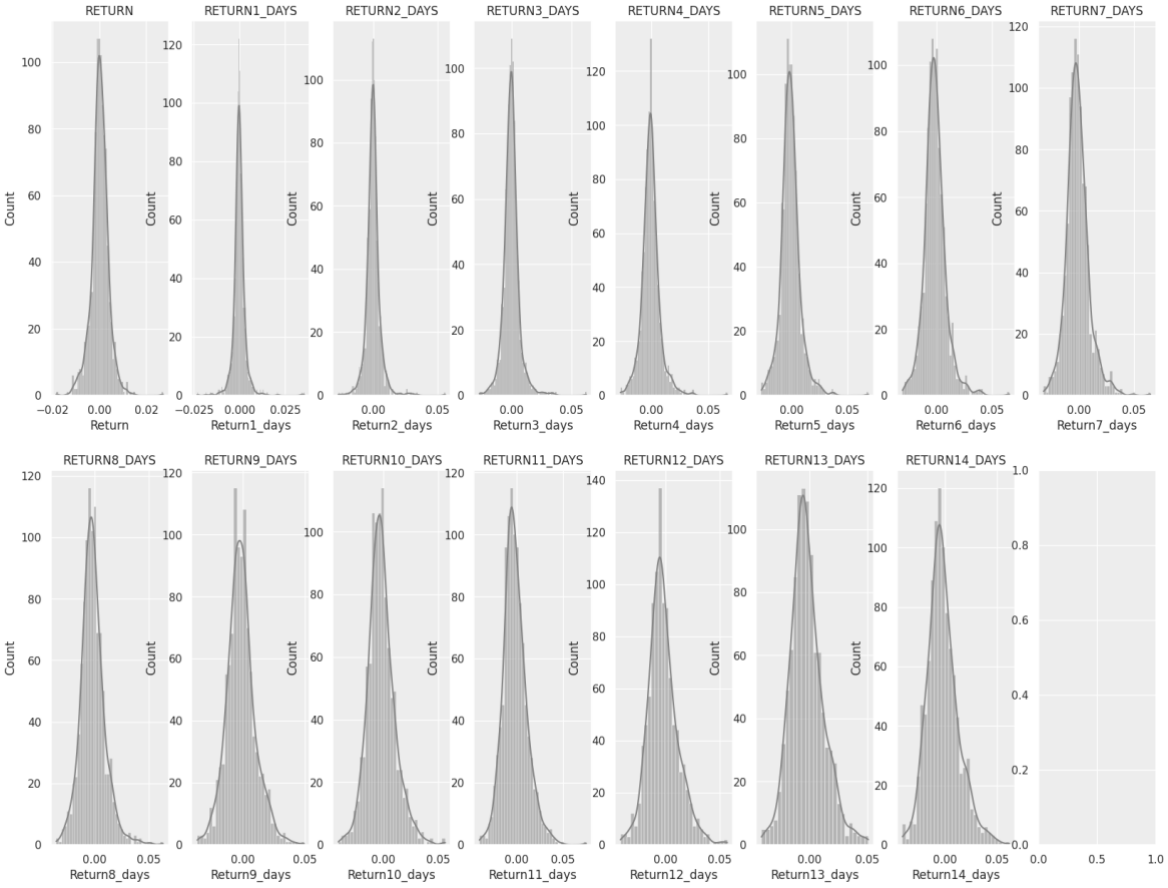


Рисунок 3.8. Розподіл прибутковостей минулих періодів за 14 днів

*Джерело: Побудовано автором на основі джерел [54]*

З огляду на аналіз розподілів вище були обрані прибутковості за 1, 3, 5 та 14 день. Цінним показником також є прибутковість в моменті, оскільки відображає ефект новизни у випадку наявності новин щодо досліджуваної акції.

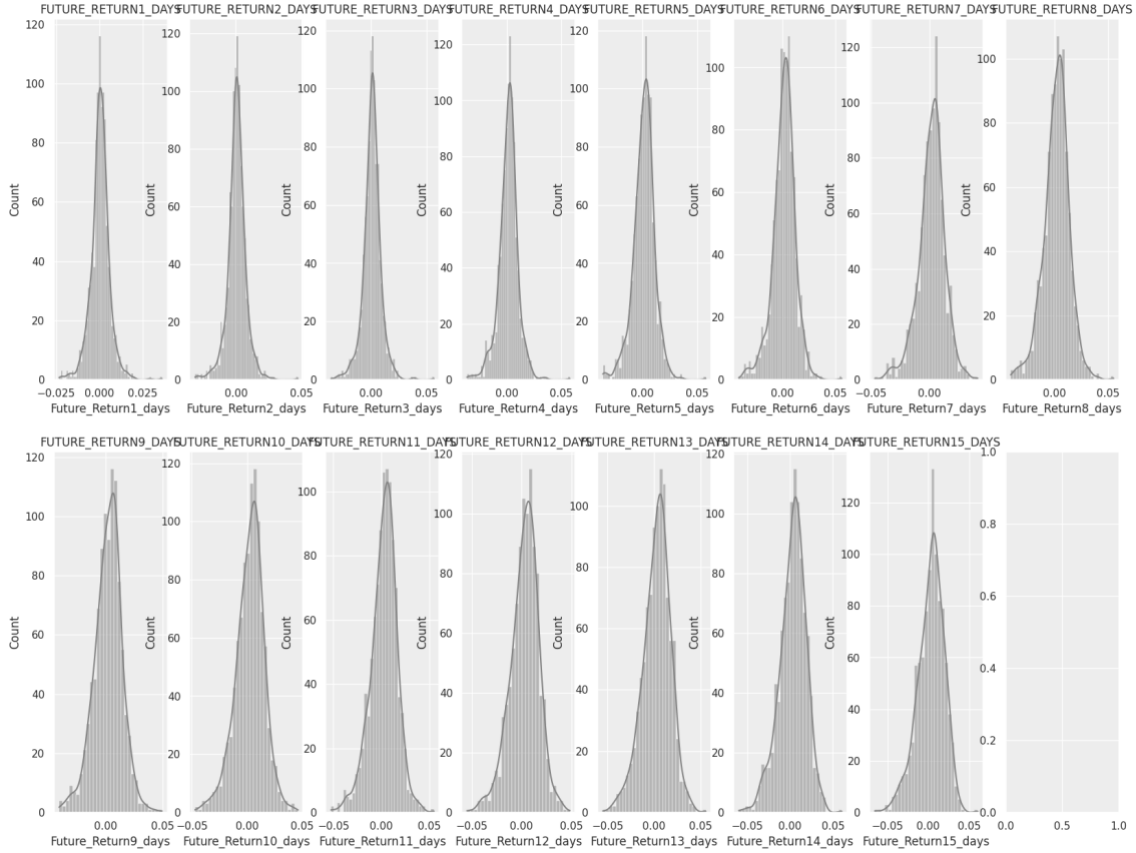


Рисунок 3.9. Розподіл прибутковостей майбутніх періодів за 14 днів

*Джерело: Побудовано автором на основі джерел [54]*

Дані майбутні прибутковості розглядаються як цільові цілі для прогнозування, тому в рамках експерименту кожна з представлених моделей 2 і 3 прогнозує по одній з даних прибутковостей.

Оскільки задача, що поставлена полягає у прогнозуванні майбутніх прибутковостей дані необхідно представити у вигляді набору послідовностей, кожна з яких прогнозується суб-моделлю моделей 2 і 3. Тобто, кожну характеристику необхідно представити у вигляді певних k членів послідовності, де k- максимальний порядок лагу кожного ряду, що представлений характеристикою вхідних данних (наприклад, ціна відкриття чи закриття позиції).

Для знаходження порядку лагу використана емпірика, що базується на значенні функцій часткової автокореляції. В данному експерименті

порядок лагу визначався як останнє значення функції ЧАК, що більша або дорівнює 0.1.

Оскільки модель 3 використовує тектові дані щодо акцій необхідно інтегрувати наявні твіти до набору данних, синтезованому вище. Для цієї цілі, спираючись на кількість лайків, ретвітів та коментарів по критерію суми були відібрані топ k новин. Також враховуючи, що твіти можуть бути написані користувачами в будь-який час, а торги на біржі ведуться лише в будні дні, то було прийнято рішення привести дату публікацій твіту до найближчого робочого дня. Для експерименту k набуває значень або 3 або 5. Це обумовлено дослідженням впливу кількості текстової інформації на якість прогнозів отриманих моделей, оскільки ціль включає в себе побудову прибуткового портфелю за розумний час.

Далі постає необхідність розділення отриманого набору даних на дві вибірки : тренувальну та тестову(валідаційну). Користуючись кращими практики індустрії, співвідношення тестової до тренувальної вибірок становить 30 до 70%.

**Отже**, для задачі побудови інвестиційного портфелю необхідно вирішити дві задачі:

- Прогнозування показників акцій, що входять в даний портфель
- Аналіз емоційного забарвлення тексту (твітів) з метою виявлення упереджених ставлень та впливу їх на характер ринку

Для вирішення даних задач були сконструйовані 3 моделі – модель для оцінки емоційного забарвлення твітів, модель для оцінки прибутковостей акцій в інвесиційному портфелі без урахування упередженостей (bias), модель для оцінки прибутковостей акцій в інвесиційному портфелі з урахуванням упередженостей (bias).

Для цілей експерименту, окрім регулювання гіперпараметрів моделей, а саме, кількості епох, кількості параметрів в кожному шарі, кроку навчання, тощо, також регулюються гіпермаратери навчального набору даних, зокрема – кількість текстових даних(топ k новин), порядок минулих прибутковостей, порядок прогнозування майбутніх прибутковостей.

## 3.3. Аналіз результатів експерименту

Наведемо в таблиці 3.1 результати навчання моделі аналізу для емоційного забарвлення твітів та заголовків новин.

Таблиця 3.1 Результати навчання моделі аналізу для емоційного забарвлення твітів та заголовків новин.

| **Метрика** | **Precision** | **Recall** | **Loss** |
|---|---|---|---|
| **Результат (train)** | 0.9414 | 0.9999 | 0.0013 |
| **Результат (test)** | 0.7227 | 0.8638 | 2.6236 |

*Джерело: Розраховано автором*

З огляду на результати, можна прийти до висновку , що модель натренована добре розпізнавати твіти по емоційному забарвленню. Невелика різниця в значенні метрик на тестувальній та навчальній вибірках свідчить про уникнення «ефекту» перенавчання мережі, а високий показник метрик Precision та Recall свідчить про адекватність моделі в цілому. В якості параметру Loss було обрано функцію категоріальної перехресної ентропії:

$$H(X) = \begin{cases} -\int\limits_{x \in X} p(x) \log p(x)\, dx, & x - \text{неперервний} \\ -\sum\limits_{x \in X} p(x) \log p(x), & x - \text{дискретний} \end{cases} \quad (3.1)$$

Де x – кількість класів. У данному випадку, множина X, себто відтінків емоційного забарвлення твітів є зліченною, а отже, формула перехресної ентропії використовується за другим сценарієм.

Маючи набір фінансових данних, була побудована та навчена низка варіацій моделі 2. У першу чергу варіація моделей є не стільки архітектурною, а скільки полягає у вигляді цільової метрики прогнозування. Тобто, моделі мають однаковий вигляд з точки зору відповідності архітектурі, зображеної на рисунку 3.4., а головна відмінність між ними полягає у виборі показника прогнозування, а саме майбутніх прибутковостей акцій.

Результати навчання моделі аналізу інвестиційного портфелю без урахування текстових данних наведемо в таблиці 3.2.

Таблиця 3.2 Результати навчання другої моделі аналізу

| **Крок прогнозування** | **Метрика** | **RMSE** | **MAE** | **MSE** |
|---|---|---|---|---|
| **1 день** | **Результат (train)** | 0.6696 | 0.5087 | 0.4484 |
| | **Результат (test)** | 1.0088 | 0.7297 | 1.0177 |
| **7 днів** | **Результат (train)** | 0.4939 | 0.3768 | 0.2439 |
| | **Результат (test)** | 0.7500 | 0.5573 | 0.5624 |
| **14 днів** | **Результат (train)** | 0.4774 | 0.3641 | 0.2279 |
| | **Результат (test)** | 0.7812 | 0.6140 | 0.6102 |

*Джерело: Розраховано автором*

Згідно з таблицею 3.2 можна дійти висновку, що модель, що прогнозує прибутковості на тиждень вперед (7 днів) є найкращою за всіма критеріями відбору, оскільки модель показує найкращі результати на тестовій вибірці.

Результати прогнозування даної моделі зображені на рисунку 3.10.

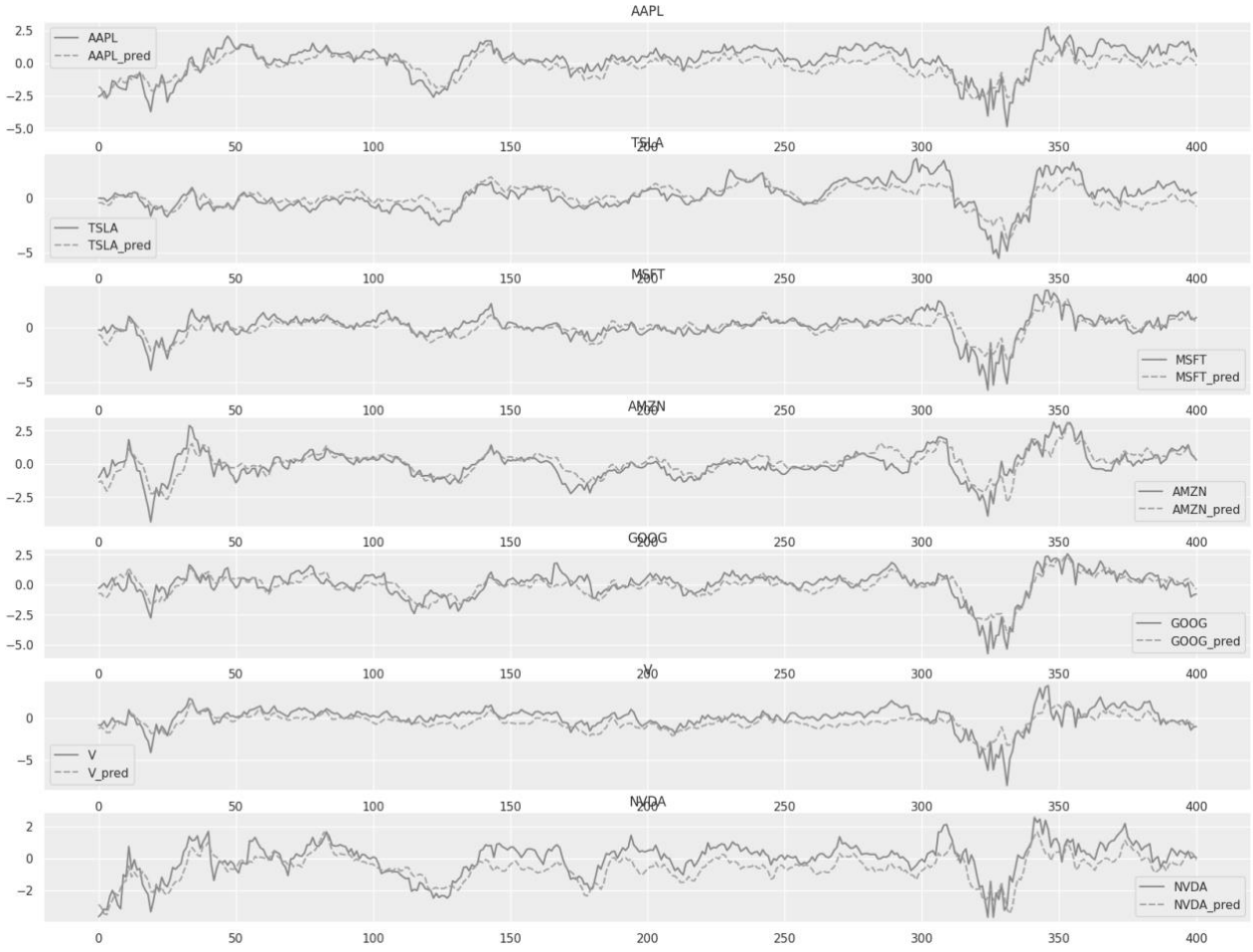


Рисунок 3.10. Результати прогнозування прибутковостей моделлю 2

*Джерело: Розраховано автором*

Таблиця 3.3 Результати навчання третьої моделі аналізу

| **Крок прогнозування** | **Метрика** | **RMSE** | **MAE** | **MSE** |
|---|---|---|---|---|
| **1 день** | **Результат (train)** | 0.5440 | 0.4063 | 0.2959 |
| | | | | |
| | **Результат (test)** | 0.8399 | 0.5769 | 0.7054 |
| **7 днів** | **Результат (train)** | 0.5005 | 0.3826 | 0.2505 |
| | **Результат (test)** | 0.7130 | 0.5316 | 0.5084 |
| **14 днів** | **Результат (train)** | 0.4653 | 0.3624 | 0.2165 |
| | **Результат (test)** | 0.7283 | 0.5411 | 0.5304 |

*Джерело: Розраховано автором*

Згідно з таблицею 3.3 можна дійти висновку, що модель, що прогнозує прибутковості на тиждень вперед (7 днів) є найкращою за всіма критеріями відбору, оскільки модель показує найкращі результати на тестовій вибірці. Також посилаючись на таблицю 3.2, що наявність текстових даних дещо покращує прогнозувальну здатність моделі і дає більш якісні результати при моделюванні портфелю.

Результати прогнозування моделі з урахуванням текстових даних можна побачити на рисунку 3.11.

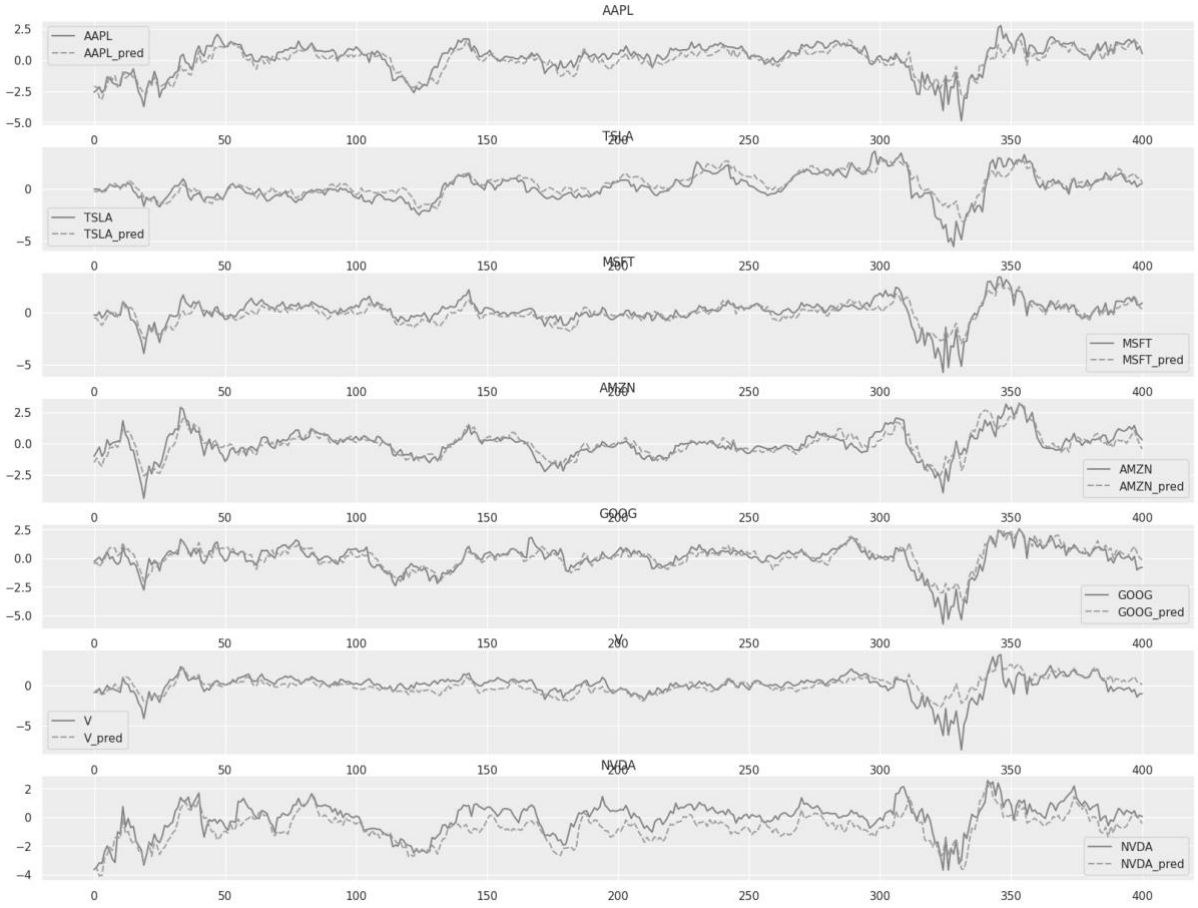


Рисунок 3.11. Результати прогнозування прибутковостей моделлю 3

*Джерело: Розраховано автором*

Далі, отримавши прогнози, перейдемо до формування інвестиційного портфелю. Дана процедура полягає у отриманні ваг кожного з цільових активів методом максимізації sharpe ratio. Після отримання ваг продуктивність портфелю вимірюємо на тестових даних підставляючи ваги, отримані з прогнозованих даних моделями 2 та 3 відповідно. Також для порівняння побудуємо інвестиційний портфель на наявних тренувальних данних методом максимізації sharpe ratio, тобто повторимо процедуру, яка є стандартною при формуванні інвестиційного

портфелю. Наведено розраховані ваги описаних вище портфелів в таблиці 3.4.

Таблиця 3.4 Ваги моделей

| | **AAPL** | **TSLA** | **MSFT** | **AMZN** | **GOOG** | **V** | **NVDA** |
|---|---|---|---|---|---|---|---|
| Ваги 2 моделі | 0.0 | 0.0 | 0.37473 | 0.62527 | 0.0 | 0.0 | 0.0 |
| Ваги 3 моделі | 0.0 | 0.89822 | 0.0 | 0.0 | 0.0 | 0.10178 | 0.0 |
| Ваги класичної моделі | 0.0 | 0.0 | 0.53166 | 0.0 | 0.0 | 0.0 | 0.46834 |

*Джерело: Розраховано автором*

Застосовуючи ваги отриманих портфелів до наявних тестових данних, виміряємо прибутковість та волатильність таких портфелів. Результати наведені в таблиці 3.5.

Таблиця 3.5 Порівняльна характеристика продуктивності отриманих портфелів

| | Модель 2 | Модель 3 | Класична модель |
|---|---|---|---|
| Expected annual return, % | 13.7 | 31.9 | 21.0 |
| Annual volatility, % | 23.2 | 43.1 | 28.7 |
| Sharpe Ratio | 0.51 | 0.69 | 0.66 |

*Джерело: Розраховано автором*

## 3.4. Висновки та перспективи

У даному розділі була сфомована задача на створення інвестиційного портфелю цінних паперів з урахування упередженості. У якості фінансових показників були обрані історичні дані з біржі NASDAQ. Вибір біржі обгрунтовується її популярністю і наявністю великої кількості різноманітних активів, що можуть бути включені до ІП. У свою чергу, у якості показників даних були обрані наступні:

- Ціна відкриття позиції
- Ціна закриття позиції
- Об’єм торгів
- Найменша ціна позиції
- Найвища ціна позиції

Також, над данними були виконані операції логарифмування заради приведення трендів в цінових показниках до лінійного виду, а об’ємів торгів до вигляду нормально розподіленої випадкової величини. Отримана інформація була доповнена розрахованими індексами Balance of Power (BOP), Commodity Channel Index (CCI), Money Flow Index (MFI), Stochastic RSI (SRSI) та ковзним середнім.

У якості текстових даних були обрані пости користувачів соціальної мережі Twitter. Серед користувачів були обрані як персональні акаунти, так і сторінки впливових фінансових видавництв, таких як FT, NYT, тощо. Рішення включити особисті акаунти пов’язана з дослідженням впливу упередженостей на ціну вказано активу, та відповідно формування інвестиційного портфелю, оскільки на відміну від заголовків світових видань персональні твіти виражають лише думку автора і не модерується редакційною колегією конкретного засобу масової інформації.

З результатів моделювання можна дійти висновку, що модель без урахування упередженості формує портфель, що є менш волатильним, ніж інші, проте і менш прибутковим (див. Таблицю 3.5), що підійде особам, що намагаються скласти більш консервативний портфель заради заощадження власних коштів і збагачення у довгостроковій перспективі. На відміну від останньої моделі, додавання текстових данних змінює структуру портфелю на таку, що має найбільшу прибутковість, проте у високу волатильність, що пов’язано з фокусом на активи, що є предметом найбільших дискусій у поточний час. Особливо цікавим є те, що дана модель рекомендує інвестувати в акції Tesla, що належать одіозному Ілону Маску, що відомий своїми контроверсійними заявами та твітами, а також займав позицію CEO Twitter.

У якості перспектив даного дослідження можна вказати:

1) Розширення вибору активів не тільки на технологічний сектор
2) Використання інших популярних соціальних мереж
3) Вивчення взаємозв’язку між компаніями, що знаходяться в одному індексі. Наприклад, наскільки взаємозалежними є два активи, що входять в один індекс, проте не є частиною одного сектору (наскільки технологічні компанії  залежні від фармацевтичних)
4) Дослідити інші підходи до створення ІП, що базується на альтернативній теорії до Sharpe Ratio

# ВИСНОВКИ

Створення оптимального інвестиційного портфелю завжди було складною задачею. Складнощі даної проблеми продиктовані стохастичною природою ринку, а також масовою діджиталізацією фондових бірж, що започаткувало появу алгоритмічного трейдингу. Задачу не спрощує також і те, що особи, що приймають рішення схильні до несвідомих або свідомих упереджень до відношенню до окремих активів, зокрема, на їх рішення впливають масові медіа.

Класичні методи пошуку оптимального портфелю зазвичай ігнорують дані про упередження фокусуючись лише на цінах активів. Однією з найпопулярніших теорій портфелю є теорія Марковіца. Вона грунтується на статистиці, а саме шлях отримання вагів портфелю полягає в пошуку оптимуму задачі квадратичної оптимізації, де в якості функціоналу прибутків є оцінка середньої прибутковісті акцій, а в якості сурогатної метрики ризику використовується волатильність даної акції.

Проте, припущення, що оцінка середньої акції є хороше наближення до середньої генеральної сукупності (потенціальний середній прибуток), а також, аналогічно, дисперсії не завжди є правильним, оскільки фінансові дані схильні до випадкових флуктуацій, в такій мірі як і до невипадкових коливань, що викликані поведінковими упередженнями. Тому, необхідно враховувати вплив останніх на майбутню ціну активу, а отже і на його прибутковість.

З розвитком обчислювальних технологій, з’явилась можливість використовувати апарат статистичного навчання на великих обсягах данних, в тому числі і фінансових. З даного твердження випливає мотивація використовувати моделі глибинного навчання для прогнозування майбутніх прибутковостей. Проаналізувавши відповідну літературу, було

ухвалено у якості основного підходу до даної задачі, зважаючи на їх специфіку, а саме прогнозування кількох часових рядів паралельно, використовувати апарат рекурентних нейронних мереж. Дані мережі неодноразово доводили свою ефективність при вирішенні задач аналізу, обробки та прогнозування данних, представлених послідовностями довільної довжини та з наявністю складних закономірностей.

Слід також зауважити, що машинне навчання дедалі частіше використовується в задачах обробки природньої мови, зокрема обробки текстів. У рамках поточного дослідження для реалізації семантичного забарвлення «твітів» було вирішено використати модель вивчення представлених слів - Word2Vec.

У даній роботі з метою встановлення впливу когнітивних та емоційних упереджень, виражених текстовою інформацією, а саме, постами в соціальній мережі Twitter, на структуру та прибутковість портфелів цінних паперів були збудовані дві архітектури нейронних мереж по прогнозуванню прибутковостей кожного активу з заданого портфелю – з урахуванням текстової інформації та без неї. Шляхом проведення експерименту емпірично було встановлено, що наявність текстової інформації зміщує центр уваги інвестора (збільшує ваги) на активи, що мають популярність в обговореннях в поточний час, тим самим збільшуючи прибутковість та волатильність синтезованого портфелю. Отримана модель дозволяє здійснювати інвестору змішану стратегію між дружнім послідовником та активним накопичувачем, нівелюючи значну кількість упереджень обох та підкреслюючи схильність до середньо та високоризикових угод.

У результаті проведення даних досліджень була встановлена перспективність подальшої роботи з метою урахування та аналізу упереджених тверджень у задачі диверсифікації портфелів цінних паперів.

# Список використаних джерел


1. Фондовий ринок: підручник : у 2 кн. — Кн. 1 / Базилевич В.Д. та ін. ; за ред. В.Д. Базилевича; Київ. нац. ун-т ім. Т. Шевченка. — Київ. : Знання, 2015. — 621 с.
2. The History: How Nasdaq Was Born. *Traders Madazine*: веб-сайт. URL: https://www.tradersmagazine.com/news/the-history-how-nasdaq-was-born/ (Дата звернення 23.04.2023)
3. Hayes A. Conservative Investing: Definition, Strategy Goals, Pros and Cons. *Investopedia*: веб-сайт. URL: https://www.investopedia.com/terms/c/conservativeinvesting.asp (Дата звернення 23.04.2023)
4. Scott G. Aggressive Investment Strategy: Definition, Benefits, and Risks. *Investopedia*: веб-сайт. URL: https://www.investopedia.com/terms/a/aggressiveinvestmentstrategy.asp (Дата звернення 23.04.2023)
5. Fernando J. Balanced Investment Strategy: Definition and Examples. *Investopedia*: веб-сайт. URL: https://www.investopedia.com/terms/b/balancedinvestmentstrategy.asp (Дата звернення 23.04.2023)
6. Зубковський Є. Що таке інвестиційний портфель? *Блог Євгена Зубковського*: веб-сайт. URL: https://e-zubkovskiy.com/blog/shcho-take-investitsiyniy-portfel#5 (Дата звернення 23.04.2023)
7. Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance*. 1952. Vol. 7. №1. P. 77–91.
8. Elton, E. J., Gruber, M. J. Modern portfolio theory, 1950 to date. *Journal of Banking & Finance*. 1997. Vol. 21. №11-12. P. 1743-1759.
9. Goetzmann, W., Ingersoll, J., Spiegel, M. I., Welch, I. Sharpening sharpe ratios. *National bureau of economic research*. 2002.
10. Jason Fernando. Sharpe Ratio Formula and Definition With Examples. *Investopedia*: веб-сайт. URL: ttps://www.investopedia.com/terms/s/sharperatio.asp (Дата звернення 23.04.2023)
11. Hayes A. Understanding Common Types of Bias in Investing. *Investopedia*: веб-сайт. URL: https://www.investopedia.com/terms/b/bias.asp (Дата звернення 23.04.2023)
12. 7 Behavioural Biases Affecting Investment Decisions. *WealthDesk*: веб-сайт. URL: https://wealthdesk.in/blog/behavioural-biases/ (Дата звернення 23.04.2023)
13. Shen, D., Urquhart, A., Wang, P. Does Twitter Predict Bitcoin? *Economics Letters*. 2019. №174. P. 118-122.

14. Kraaijeveld, O., De Smedt, J. The Predictive Power of Public Twitter Sentiment for Forecasting Cryptocurrency Prices. *Journal of International Financial Markets, Institutions, and Money*. 2020. № 65. P. 101-188.
15. Zhang, J. (2020). Do Cryptocurrency Markets React to Issuer Sentiments? Evidence from Twitter. Evidence from Twitter. URL: http://dx.doi.org/10.2139/ssrn.3675196 (Дата звернення 24.04.2023)
16. Aharon, D. Y., Demir, E., Lau, C. K. M., Zaremba, A. Twitter-Based Uncertainty and Cryptocurrency Returns. 2020.
17. Öztürk, S. S., Bilgiç, M. E. Twitter & Bitcoin: Are the Most Influential Accounts Really Influential?. *Applied Economics Letters.* 2021. P. 1-4.
18. Ante, L. How Elon Musk's Twitter Activity Moves Cryptocurrency Markets. 2021. URL: http://dx.doi.org/10.2139/ssrn.3778844 (Дата звернення 24.04.2023)
19. Tversky, A., & Kahneman, D. Judgement Under Uncertainty: Heuristics and Biases. *Science*. 1974. Vol. 185. № 4157. P. 1124-1131.
20. Kahneman, D. & Tversky, A. Prospect Theory: An Analysis of Decision Under Risk. *Econometrica*. 1979. Vol. 47, P. 263–291.
21. Marguerite Frank, Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*. 1956. Vol. 3. №1-2. P. 95–110, 1956.
22. Richard O Michaud. The markowitz optimization enigma: Is ‘optimized’ optimal? *Financial analysts journal*. 1989. Vol. 45. №1. P. 31–42. URL: https://deliverypdf.ssrn.com/delivery.php?ID=310013021027067030080091031002119123059041038044021064118028086030123004002098017104053039042027114097000125122064082096101076019061023049001007070104083124009125054050041087077123025108090002112018013122011114079073006125101012105021110117087114074&EXT=pdf&INDEX=TRUE (Дата звернення 05.03.2023)
23. Lai T. L., Xing H., Chen Z. Mean–variance portfolio optimization when means and covariances are unknown. *The Annals of Applied Statistics*, 2011. Vol. 5. № 2A. P. 798–823.
24. Kan R, Zhou G. Optimal portfolio choice with parameter uncertainty. *Journal of Financial and Quantitative Analysis*. 2007. Vol. 42. № 3. P. 621–656.
25. Jorion P. Bayes-stein estimation for portfolio analysis. *Journal of Financial and Quantitative analysis*. 1986. Vol. 21. № 3. P. 279–292.
26. Samo Y-L. K., Vervuurt A.. Stochastic Portfolio Theory: A machine learning perspective. *In Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*. 2016.
27. Ledoit O., Wolf M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*. 2003. Vol. 10. № 5. P. 603–621.

28. Ledoit O., Wolf M. Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management*. 1994. Vol. 30. № 4. P. 110-119.
29. Fischer T., Krauss C. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*. 2018. Vol. 270. № 2. P. 654–669.
30. Zohren, S., Zhang Z., Roberts S. Deep learning for portfolio optimization. *The Journal of Financial Data Science*, 2020. URL: https://jfds.pm-research.com. (Дата звернення 15.03.2023)
31. Butler A., Kwon R. H. Integrating prediction in mean-variance portfolio optimization. 2021. URL: http://arxiv.org/abs/2102.09287.
32. Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press, 2016. URL: https://www.deeplearningbook.org/contents/rnn.html (Дата звернення 04.04.2023)
33. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*. 2013. URL: https://arxiv.org/pdf/1301.3781.pdf (Дата звернення 01.04.2023)
34. Pástor L., Stambaugh R. F. Comparing asset pricing models: An investment perspective. *Journal of Financial Economics*. 2000. Vol. 56. № 3. P. 335–381.
35. What is Balance of Power (BOP) Indicator: Measure the Strength of Buying and Selling Pressure. *Phemex*: веб-сайт. URL: https://phemex.com/academy/what-is-balance-of-power-indicator-bop (Дата звернення 28.03.2023)
36. Cory Mitchell. What Is the Commodity Channel Index (CCI)? How To Calculate. *Investopedia*: веб-сайт. URL: https://www.investopedia.com/terms/c/commoditychannelindex.asp(Дата звернення 28.03.2023)
37. Cory Mitchell. Money Flow Index - MFI Definition and Uses. *Investopedia*: веб-сайт. URL: https://www.investopedia.com/terms/m/mfi.asp (Дата звернення 28.03.2023)
38. Про стохастичний RSI. *Binance Academy*: веб-сайт. URL: https://academy.binance.com/uk/articles/stochastic-rsi-explained (Дата звернення 28.03.2023)
39. Tweets about the Top Companies from 2015 to 2020. *Kaggle*: веб-сайт URL: https://www.kaggle.com/datasets/omermetinn/tweets-about-the-top-companies-from-2015-to-2020 (Дата звернення 26.02.2023)
40. Yasukochi C. Tech Among Most Resilient US Employment Sectors. *CBRE*: веб-сайт. URL: https://www.cbre.us/real-estate-services/real-estate-industries/technology-and-media/tech-insights/articles/tech-among-most-resilient-us-employment-sectors (Дата звернення 15.04.2023)

41. Innovation in a crisis: Why it is more critical than ever. *McKinsey & Company*: веб-сайт. URL: https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/innovation-in-a-crisis-why-it-is-more-critical-than-ever (Дата звернення 15.04.2023)
42. Which companies will survive the next crisis? *IMD*: веб-сайт. URL: https://www.imd.org/news/updates/which-automotive-and-financial-companies-will-survive-the-next-crisis/ (Дата звернення 15.04.2023)
43. Beyond Silicon Valley, Spending on Technology Is Resilient. *The New York Times*: веб-сайт. URL: https://www.nytimes.com/2023/02/13/technology/technology-spending-resilient.html (Дата звернення 15.04.2023)
44. X, Rong, “word2vec Parameter Learning Explained”, arXiv:1411.2738v4 (2016)
45. Kharde, Vishal, and Sheetal Sonawane. "Sentiment Analysis of Twitter Data: A Survey of Techniques." arXivpreprint arXiv:1601.06971 (2016).
46. B. Pang and L. Lee, "Opinion mining and sentiment analysis," 2008. [Online]. Available: http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf. Accessed: Oct. 23, 2016.
47. Maas, R. Daly, P. Pham, & D. Huang, (n.d.). Learning Word Vectors for Sentiment Analysis. Proceeding HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 142-150.
48. Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N. Project Report, Stanford 1 (2009): 12.
49. Kouloumpis, Efthymios, Theresa Wilson, and Johanna Moore. Twitter Sentiment Analysis: The Good the Bad and the OMG! in Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-2011). 2011.
50. Lilleberg, Joseph, and Yun Zhu. "Support Vector Machines and Word2vec for Text Classification with Semantic Features." Cognitive Informatics & Cognitive Computing (2015): n. pag. Print.
51. H. Wang, L. Liu, W. Song, and J. Lu, “Feature-based sentiment analysis approach for product reviews,” Journal of Software, vol. 9, no. 2, p. 274, Feb. 2014.
52. S. Sarkar, S. Goswami, A. Agrwal, J. Aktar. “A Novel Feature Selection Technique for Text Classification Using Naïve Bayes” International Scholarly Research Notices, vol. 2014, no. 717092, p. 1-10, Apr. 2014
53. Terrence Odean and Brad Barber, “Boys Will Be Boys: Gender, Overconfidence, and Common Stock Investment,” Quar- terly Journal of Economics 116, 1 (Febru- ary 2001): 261–292.

54. Polygon Finance API. *Polygon:* веб-сайт. URL: https://polygon.io/stocks (Дата звернення: 07.05.2023)

# ДОДАТОК А. ЛІСТИНГИ КОДУ

Модуль обробки даних

```
import gensim.downloader as apiimport nltk
import re
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('wordnet')
nltk.download('omw-1.4')
from nltk.corpus import wordnet, stopwords
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize
import talib
import yfinance
import gzip
import typing as tp
import pandas as pd
from tqdm.notebook import tqdm
from collections import defaultdict
import os

twitter_vectors = api.load('glove-twitter-50')
wnl = WordNetLemmatizer()

def filter_out_non_text(text: str) -> str:
  filtered_text = re.sub(r'[^\w\s]', '', text)
  filtered_text = repr(filtered_text.encode('utf-
8'))[2:-1]
  filtered_text = filtered_text.replace('\\n',' ')
```

```
  filtered_text = re.sub(r"won\'t", "will not",
filtered_text)
  filtered_text = re.sub(r"can\'t", "can not",
filtered_text)
  filtered_text = re.sub(r"n\'t", " not",
filtered_text)
  filtered_text = re.sub(r"\'re", " are",
filtered_text)
  filtered_text = re.sub(r"(he|He)\'s", "he is",
filtered_text)
  filtered_text = re.sub(r"(she|She)\'s", "she is",
filtered_text)
  filtered_text = re.sub(r"(it|It)\'s", "it is",
filtered_text)
  filtered_text = re.sub(r"\'d", " would",
filtered_text)
  filtered_text = re.sub(r"\'ll", " will",
filtered_text)
  filtered_text = re.sub(r"\'t", " not",
filtered_text)
  filtered_text = re.sub(r"[#@$]", "", filtered_text)
  filtered_text = re.sub(r"(\'ve|has)", " have",
filtered_text)
  filtered_text = re.sub(r"\'m", " am",
filtered_text)
  filtered_text = re.sub(r"\\x[a-f0-9]*", "",
filtered_text) # прибрати емодзі
  filtered_text = filtered_text.split(' ')
  filtered_text = ' '.join([
```

```
      wnl.lemmatize(word.lower()) for word in
filtered_text if word.isalpha()
      ])
  return filtered_text

text =
pd.read_csv('train/stock_market_crash_2022.csv')[['te
xt', 'text_sentiment']]
filtered_text = text['text'].map(filter_out_non_text)
filtered_list = list(map(lambda sentence:
sentence.split(' '), filtered_text.values.tolist()))
all_vocab = set(sum(filtered_list, [])) &
set(twitter_vectors.index_to_key[:])

dataset = text[['text_sentiment']]
dataset['texts'] = filtered_text

def construct_vocabulary():
  res = []
  for word in all_vocab:
    if word != '':
        res.append(twitter_vectors[word])
  return np.vstack(res)

vocabulary = construct_vocabulary()
all_vocab = [x for x in all_vocab if x!='']

columns = ['Open', 'Close', 'High', 'Low', 'Volume']
```

```
tickers = ['AAPL', 'TSLA', 'MSFT', 'AMZN', 'GOOG',
'V', 'NVDA']

def read_tickers(tickers):
  """
  Read and process stock prices fro given list of
tickers
  """
  df = dict()
  for ticker_name in tickers:
    ticker = yfinance.Ticker(ticker_name)
    history_prices = ticker.history('1d',
start='2015-01-01', end='2020-07-01')[columns]
    df[ticker_name] = history_prices
  return df

def plot_tickers(tickers_dataset: pd.DataFrame, name:
str = 'Stock Prices'):
  """
  Plot tickers from tickers dataset
  """
  number_of_plots = len(tickers_dataset.columns)
  fig, ax = plt.subplots(number_of_plots,
figsize=(15, 10))
  fig.suptitle(name)
  for idx, ticker in
enumerate(tickers_dataset.columns):
    if number_of_plots > 1:
      axis = ax[idx]
```

```
    else:
      axis = ax
    axis.plot(tickers_dataset.index.values.ravel(),
tickers_dataset[ticker].values.ravel())
    axis.set(ylabel=ticker)
  plt.show()

def shift_data(tickers_dataset, y_col, days_shift):
  """
  Add to data some lags for given column
  """
  new_dataset = tickers_dataset.copy()
  for lag in range(1, days_shift + 1):
    new_dataset[f'{y_col}_-{lag}'] =
new_dataset[y_col].shift(lag)

  return new_dataset

def compute_return_for_k_days(df: pd.DataFrame, k:
int = 0):
  column_name = 'Return' + (f'{k}_days' if k else '')
  df[column_name] = df['Close'] - df['Open'].shift(-
k)

def compute_future_return_for_k_days(df:
pd.DataFrame, k: int = 0):
  column_name = 'Future_Return' + (f'{k}_days' if k
else '')
```

```
  df[column_name] = df['Close'].shift(-k) -
df['Open']

def compute_stoch(x: pd.DataFrame,
                  fastk_period: int = 14,
                  slowk_period: int = 3,
                  slowk_matype: int = 0,
                  slowd_period: int = 3,
                  slowd_matype: int = 0):
  slowk, slowd = talib.STOCH(x['High'].ffill(),
x['Low'].ffill(), x['Close'].ffill(),

fastk_period=fastk_period,

slowk_period=slowk_period,

slowk_matype=slowk_matype,

slowd_period=slowd_period,

slowd_matype=slowd_matype)
  x['slowd'] = slowd
  x['slowk'] = slowk

def compute_bop(x: pd.DataFrame):
  x['BOP'] = talib.BOP(x['Open'],
                       x['High'],
                       x['Low'],
                       x['Close'])
```

```
def compute_cci(x: pd.DataFrame):
  x['CCI'] = talib.CCI(x['High'],
                       x['Low'],
                       x['Close'])

def compute_mfi(x: pd.DataFrame, timeperiod: int =
14):
  x['MFI'] = talib.MFI(x['High'],
                       x['Low'],
                       x['Close'],
                       x['Volume'],
                       timeperiod=timeperiod)

def compute_wma(x: pd.DataFrame, column: str):
  x['WMA_'+column] =
talib.WMA(x[column].fillna(method='pad'))

dict_nasdaq = read_tickers(tickers)

twts = pd.read_csv('tweets/Tweet.csv')
twts['post_date'] = pd.to_datetime(twts.post_date,
unit='s')
comp_tweet = pd.read_csv('tweets/Company_Tweet.csv')
twts = twts.merge(comp_tweet, on='tweet_id',
how='inner')
twts['total_engagement'] = twts['like_num'] +
twts['retweet_num'] + twts['comment_num']
```

```
twts = twts[twts['total_engagement'] >
100][['writer', 'body', 'ticker_symbol',
'total_engagement', 'post_date']]
twts['post_date'] = twts.post_date.dt.ceil("d")
twts = twts.rename(columns={'post_date': 'Date',
'body': 'text'}).set_index('Date')
twts['text'] = twts.text.map(filter_out_non_text)
twts = twts.drop_duplicates(subset='text')[['writer',
'ticker_symbol', 'total_engagement', 'text']]
twts = twts.groupby([twts.index,
'ticker_symbol']).aggregate({'text': list})
twts['text'] = twts['text'].apply(lambda x: '.
'.join(x))
twts = twts.reset_index(level=1)
twts_2 =
pd.read_csv('stock_market_tweets.csv').dropna()
twts_2['post_date'] =
pd.to_datetime(twts_2.post_date)
twts_2['total_engagement'] =
twts_2['like_num'].astype(int) +
twts_2['retweet_num'].astype(int) +
twts_2['comment_num'].astype(int)
twts_2 = twts_2[twts_2['total_engagement'] >
100][['writer', 'body', 'ticker_symbol',
'total_engagement', 'post_date']]
twts_2['post_date'] = twts_2.post_date.dt.ceil("d")
twts_2 = twts_2.rename(columns={'post_date': 'Date',
'body': 'text'}).set_index('Date')
twts_2['text'] = twts_2.text.map(filter_out_non_text)
```

```
twts_2 =
twts_2.drop_duplicates(subset='text')[['writer',
'ticker_symbol', 'total_engagement', 'text']]
news_datasets = [

'/content/train/raw_analyst_ratings.csv',

'/content/train/raw_partner_headlines.csv'
                ]
datasets = []
for fname in news_datasets:
  dataset = pd.read_csv(fname, usecols=['date',
'stock', 'headline'])
  dataset['date'] =
pd.to_datetime(dataset.date.str[:10])
  dataset = dataset.rename(columns={'date': 'Date',
'headline': 'text', 'stock':
'ticker_symbol'}).set_index('Date')
  dataset['text'] =
dataset.text.map(filter_out_non_text)
  datasets.append(dataset)
datasets = pd.concat(datasets)
twts = pd.concat([datasets, twts[['ticker_symbol',
'text']], twts_2[['ticker_symbol', 'text']]])
twts =
twts.sort_values(by='Date').drop_duplicates('text')

def add_text(stock: str):
```

```
  tweets_about_stock = twts[twts.ticker_symbol ==
stock][['text']].groupby('Date').agg({'text': set})
  tweets_about_stock['text'] =
tweets_about_stock['text'].map(list)
  dict_nasdaq[stock] =
dict_nasdaq[stock].join(tweets_about_stock,
how='left')
  dict_nasdaq[stock]['text'] =
dict_nasdaq[stock]['text'].fillna('')

from pandas.tseries.offsets import BDay, Day
twts.index = twts.index.map(lambda x : x - Day(1) +
BDay(1))

for t in tickers:
  dict_nasdaq[t].index =
dict_nasdaq[t].index.tz_localize(None)

tickers_mean_std = {ticker: [] for ticker in tickers}

for t in tickers:
  dict_nasdaq[t]['Open'] =
np.log(dict_nasdaq[t]['Open']) # звести все до
лінійного тренду
  dict_nasdaq[t]['Close'] =
np.log(dict_nasdaq[t]['Close'])
  dict_nasdaq[t]['Low'] =
np.log(dict_nasdaq[t]['Low'])
```

```
  dict_nasdaq[t]['High'] =
np.log(dict_nasdaq[t]['High'])
  dict_nasdaq[t]['Volume'] =
np.log(dict_nasdaq[t]['Volume'])
  dict_nasdaq[t] =
dict_nasdaq[t].fillna(method='pad')
  compute_stoch(dict_nasdaq[t])
  compute_bop(dict_nasdaq[t])
  compute_cci(dict_nasdaq[t])
  compute_mfi(dict_nasdaq[t])
  compute_wma(dict_nasdaq[t], ('Open'))
  compute_wma(dict_nasdaq[t], ('Close'))
  compute_wma(dict_nasdaq[t], ('Low'))
  compute_wma(dict_nasdaq[t], ('High'))
  for k in range(0, 15):
    compute_return_for_k_days(dict_nasdaq[t], k)
  for k in range(1, 15):
    compute_future_return_for_k_days(dict_nasdaq[t],
k)
  mean = dict_nasdaq[t].mean()
  std = dict_nasdaq[t].std()
  tickers_mean_std[t].extend([mean, std])
  dict_nasdaq[t] = (dict_nasdaq[t] - mean)/std
  add_text(t)
  for text_idx in range(5):
    dict_nasdaq[t][f"text_top{text_idx + 1}"] =
dict_nasdaq[t]['text'].map(lambda x: x[text_idx] if
text_idx<len(x) else '')
```

```
import statsmodels.tsa.api as sm

def construct_dataset(stock: str,
column_lag_order_dict: tp.Dict[str, int],
columns_to_keep: tp.List[str], target_feature: str) -
> tp.Dict[str, tp.List[str]]:
  new_ds = dict_nasdaq[stock][columns_to_keep].copy()
  new_ds = new_ds.rename(columns={column:
f'{stock}_{column}' for column in columns_to_keep})
  column_mapping = {column: [column] for column in
new_ds.columns}
  target_feature_name = f"{stock}_{target_feature}"
  new_ds[target_feature_name] =
dict_nasdaq[stock][target_feature]
  for column, lag_order in
column_lag_order_dict.items():
    if lag_order=='auto':
      lag_order_temp =
sum(sm.stattools.pacf(dict_nasdaq[stock][column].drop
na()) >= 0.1)
    else:
      lag_order_temp = lag_order
    column_mapping.update({ f"{stock}_{column}":
[f"{stock}_{column}_{lag}_days" for lag in
range(lag_order_temp+1)]})
    for lag in range(lag_order_temp+1):
      new_ds[f"{stock}_{column}_{lag}_days"] =
dict_nasdaq[stock][column].shift(lag)
  dict_nasdaq[stock] = new_ds.dropna()
```

```
  return column_mapping, target_feature_name

def convert_to_dataset(pandas_dataset: pd.DataFrame,
                       column_mapping:
tp.Dict[tp.List[str], str],
                       target_feature_names:
tp.List[str]) -> tf.data.Dataset: #ще раз запитати у
Діми
  tf_datasets = []
  for subset in column_mapping.values():
    vals = pandas_dataset[subset].values
    shape = (*vals.shape, 1) # (entries, timesteps,
1)
    vals = vals.reshape(shape)

tf_datasets.append(tf.data.Dataset.from_tensor_slices
(vals))
  tf_datasets =
tf.data.Dataset.zip(tuple(tf_datasets)).map(lambda
*features: {feature_n: (feature if
not(feature_n.split('_')[1].startswith('text')) else
feature[:, 0]) for feature_n, feature in
zip(column_mapping.keys(), features)})
  target_vals = pandas_dataset[target_feature_names]
  target_vals =
tf.data.Dataset.from_tensor_slices(target_vals)
  tf_datasets = tf.data.Dataset.zip((tf_datasets,
target_vals))
  return tf_datasets
```

```
column_lag_order_dict = {
    "Open": "auto",
    "Close": "auto",
    "Low": "auto",
    "High": "auto",
    "Volume": "auto",
    "Return": "auto",
    "Return1_days": "auto",
    "Return3_days": "auto",
    "Return5_days": "auto",
    "Return14_days": "auto",
    'slowd': "auto",
    'slowk': "auto",
    'BOP': "auto",
    'CCI': "auto",
    'MFI': "auto",
    'WMA_Open': "auto",
    'WMA_Close': "auto",
    'WMA_Low': "auto",
    'WMA_High': "auto"
    }
columns_to_keep = ['Open', 'text_top1', 'text_top2',
'text_top3']
#columns_to_keep = ['Open']

target_feature = 'Future_Return7_days'
features_mapping = dict()
target_variables = list()
```

```
for t in tickers:
  dict_features, targets = construct_dataset(t,
column_lag_order_dict, columns_to_keep,
target_feature)
  features_mapping.update(dict_features)
  target_variables.extend([targets])

target_dataset =
pd.concat(list(dict_nasdaq.values()),
axis=1).fillna(method='pad').dropna()

tf_ds = convert_to_dataset(target_dataset,
features_mapping, target_variables)
n_records = target_dataset.shape[0]
train_fraction = 0.7
batch_size = 32
n_train = int(train_fraction * n_records)
train_ds =
tf_ds.take(n_train).batch(batch_size).cache().prefetc
h(2048)
test_ds =
tf_ds.skip(n_train).batch(batch_size).cache().prefetc
h(2048)
```

Модель аналізу тексту

```
class SentimentModel(tf.keras.Model):
  def __init__(self,
               max_token: int,
               kernel_size: int,
```

```
                filters: int):
    super().__init__(self)

self.vectorizer=tf.keras.layers.TextVectorization(out
put_sequence_length=max_token, vocabulary=all_vocab)

self.embeddings=tf.keras.layers.Embedding(len(all_voc
ab) + 2, 50, weights=tf.constant([vocabulary]))

self.conv_1=tf.keras.layers.Conv1D(filters=filters,
kernel_size=kernel_size, activation='relu')
    self.average_1 =
tf.keras.layers.AveragePooling1D(2)

self.conv_2=tf.keras.layers.Conv1D(filters=filters,
kernel_size=kernel_size, activation='relu')
    self.average_2 =
tf.keras.layers.AveragePooling1D(2)
    self.flatten = tf.keras.layers.Flatten()
    self.dense = tf.keras.layers.Dense(64,
activation='relu')
    self.output_layer = tf.keras.layers.Dense(3,
activation='softmax')

  def call(self, input_):
    o=self.vectorizer(input_)
    o=self.embeddings(o)
    o=self.conv_1(o)
    o=self.average_1(o)
```

```
    o=self.conv_2(o)
    o=self.average_2(o)
    o=self.flatten(o)
    o=self.dense(o)
    o=self.output_layer(o)
    return o

semantic_model = SentimentModel(max_token=30,
kernel_size=3, filters=64)
semantic_model.compile(optimizer=tf.keras.optimizers.
Adam(0.001),
loss=tf.keras.losses.CategoricalCrossentropy(from_log
its=False), metrics=[tf.keras.metrics.Precision(),
tf.keras.metrics.Recall()])
semantic_model.fit(x=dataset[['texts']].values,
y=y.values, validation_split=0.4, epochs=50,
batch_size=64, verbose=0)
```

Модель прогнозу прибутковостей

```
class TimeSeriesPredictor(tf.keras.Model):
  def __init__(self, feature_list: tp.List[str],
num_reccurrent_dims: int, output_size: int,
text_model: tf.keras.Model = None):
    super().__init__()
    self.rnn_layers =
defaultdict(tf.keras.layers.Layer)
    self.feature_list = feature_list
    self.text_model=False
    for feature in self.feature_list:
      if feature.split('_')[1] =='text':
```

```
        if text_model:
          self.text_model=True
          text_model_ = text_model
          text_model_.trainable = False
          self.rnn_layers[feature] = text_model_
      else:
        self.rnn_layers[feature] =
tf.keras.Sequential(
            [

tf.keras.layers.LSTM(num_reccurrent_dims)
            ]
      )

    self.dense_output =
tf.keras.layers.Dense(output_size)

  def call(self, x: tf.Tensor):
    rnn_outputs = []
    for feature in self.feature_list:
      if feature.split('_')[1] == 'text':
        if self.text_model:

rnn_outputs.append(self.rnn_layers[feature](x[feature
]))
      else:

rnn_outputs.append(self.rnn_layers[feature](x[feature
]))
```

```
rnn_outputs = tf.concat(rnn_outputs, axis=-1)
out = self.dense_output(rnn_outputs)
return out
```