DOI: https://doi.org/10.15276/ict.03.2026.14
УДК 004.8:336.76

Порівняння рівнів складності обробки новинного потоку у задачі прогнозування волатильності акцій

Ницька Катерина Олександрівна
ORCID: https://orcid.org/0009-0005-8027-0100; katerynanytska@gmail.com
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»,
Берестейський просп., 37. Київ, 03056, Україна

АНОТАЦІЯ

Предметом дослідження цієї роботи є прогнозування волатильності акцій з використанням новин як потенційного джерела додаткової інформації для покращення якості прогнозування. Способи використання новинного потоку відрізняються як складністю реалізації, так і обчислювальною вартістю, тому метою роботи є визначення рівня складності обробки новин, після якого приріст точності прогнозування припиняється. Для дослідження використовувались дані про акції компаній з однієї галузі за період чотирнадцяти років і близько сорока дев'яти тисяч новинних повідомлень, при цьому глибина новинного покриття за активами різна. У ході роботи здійснювалось прогнозування логарифму денної дисперсії, оціненої на основі чотирьох значень цін за торговий день, а саме ціни відкриття, закриття, мінімуму і максимуму акцій. Базовою моделлю прогнозування виступила гетерогенна авторегресія з трьома основними компонентами, які відповідають денному, тижневому й місячному горизонтам усереднення змінної. Для порівняння результатів використовувались також моделі умовної гетероскедастичності, модель ковзного середнього і наївний прогноз як нижня межа точності. Новинний потік подавався на вхід моделі як додатковий регресор у трьох рівнях складності обробки: перший рівень – кількість новин за день і відхилення цієї кількості від очікуваного рівня; другий – тональність новин, визначена за словниковим методом; третій – тональність, визначена за фінансовою мовною моделлю. Тональність в обох методах подавалась однаковим набором змінних, тому різниця між цими рівнями відображає якість оцінювання тональності, а не кількість оцінюваних коефіцієнтів. Параметри моделей оцінювались на ковзному вікні, а прогноз будувався на день вперед для днів, які в оцінювання не входили. Оскільки цільова змінна сама вимірюється з похибкою, якість прогнозування оцінювалась метрикою, стійкою до цього, а різниця точності між моделями перевірялась формальним тестом порівняння прогностичної точності. Інформативним для прогнозу виявилось саме відхилення обсягу новин від очікуваного рівня, а не їхня тональність. Різниця між словниковою моделлю виявлення тональності новин і фінансовою мовною моделлю є статистично незначущою на розглянутих акціях, попри те що остання потребує на чотири порядки більше обчислювальних ресурсів. Після поправки на множинність лише найпростіша специфікація значуще покращує прогноз, проте її внесок виявився приблизно у шістдесят разів меншим за внесок власної історії волатильності. Перспективи подальших досліджень – перевірка отриманих результатів на оцінці ринкового ризику акцій, застосування великої мовної моделі як наступного рівня обробки новинного потоку для прогнозування.

Ключові слова: фінансові часові ряди; прогнозування волатильності; гетерогенна авторегресія; новинний потік; аналіз тональності новин; словниковий аналіз тональності; фінансова мовна модель; порівняння прогностичної точності

Для цитування: **Ницька К. О.“Порівняння рівнів складності обробки новинного потоку у задачі прогнозування волатильності акцій”. *Інформатика. Культура. Техніка*. 2026; Том 3 №1(3): 166–178. DOI: https://doi.org/10.15276/ict.03.2026.14**

Актуальність. На сьогодні прогнозування волатильності залишається важливою дослідницькою задачею, оскільки волатильність є невід'ємною складовою оцінювання ринкового ризику, визначення розміру позиції, а також ціноутворення похідних інструментів. На відміну від прогнозування ціни, яке здебільшого має на меті отримання прибутку, без прогнозу волатильності ризик неможливо виміряти взагалі. Волатильність також має стійкі емпіричні властивості – кластеризацію, довгу пам'ять, ефект левериджу – саме тому вона піддається прогнозуванню краще, ніж напрямок ціни.

Обсяг доступної текстової інформації різко зріс, що робить новини перспективним джерелом інформації, якої не містить сам ціновий ряд. Раніше використання новин було ускладнене відсутністю відкритих джерел, звідки їх можна було б отримати у придатному для обробки вигляді. Нині ситуація змінилась: з'явились великі машинозчитувані архіви,

одним із яких є набір даних FNSPID [1], використаний і в цій роботі. Він містить 15,7 млн новинних повідомлень для 4775 тікерів за період з 1999 по 2023 рік.

Зростання обсягу доступної текстової інформації стимулювало створення дедалі складніших методів і моделей її обробки. Складніші способи обробки тексту коштують дорожче як у реалізації, так і обчислювально. Крім того, кожна додаткова змінна вимагає оцінювання додаткового коефіцієнта на скінченній вибірці, а це підвищує дисперсію оцінок. Тому постає питання, чи є перехід до складніших методів обробки завжди виправданим. Виникає також сумнів, чи придатна для прогнозування волатильності знакова величина, якою є тональність, якщо сама волатильність знаку не має. Коли за день виходять і новини з різко негативною тональністю, і новини з різко позитивною, середня тональність опиниться біля нуля, хоча саме такий день характеризується підвищеною невизначеністю, що може призвести до збільшення волатильності.

Дослідження зв'язку новин з волатильністю показали, що тональність виявляється менш інформативною, ніж міри уваги – обсяг повідомлень та кількість пошукових запитів [2], [3], [4], [5]. Водночас більшість робіт порівнюють один конкретний спосіб обробки текстової інформації з моделлю без новин, а не кілька способів між собою на єдиній схемі оцінювання [6]. Крім того, порівняння моделей часто проводиться без формального тесту рівної прогностичної точності, тому різницю між моделями може бути складно відрізнити від випадкової. Отже, питання про рівень складності обробки новин, після якого приріст точності припиняється, залишається відкритим.

Мета дослідження полягає у встановленні межі складності обробки новинного потоку, при якій приріст точності прогнозування волатильності припиняється. Це досягається шляхом порівняння способів урахування новинного потоку, упорядкованих за зростанням складності обробки, за однакової для всіх специфікацій процедури оцінювання на денних даних акцій компаній однієї галузі. Об'єктом дослідження виступає процес прогнозування волатильності акцій, а предметом – методи врахування новинної інформації різного рівня складності обробки в моделях прогнозування волатильності.

Емпіричне дослідження проведено на даних восьми компаній галузі напівпровідників, список яких визначено властивостями набору даних новин. Джерелом новинного потоку слугував набір FNSPID [1]. Набір складається з окремих колекцій, зібраних з різних джерел і таких, що містять різні записи новин, але в роботі, з огляду на наведені нижче обмеження, використано лише одну, а саме nasdaq_exteral_data. У колекції All_external лише близько 5,5% записів мають прив'язку до тікера компанії. Крім того, тільки колекція nasdaq_exteral_data містить повні тексти статей, і хоча в цій роботі використовувались лише їхні заголовки, наявність повних текстів може бути корисною для подальших досліджень. Об'єднання двох колекцій розглянуто як недоцільне через структурний розрив у періодах та їхню нетотожність: колекція All_external охоплює період до 2020 року, а nasdaq_exteral_data триває до 2023. Об'єднання призвело б до штучного зменшення обсягу новин для активів після 2020 року, що перевірено на окремих активах. Наприклад, QCOM має приблизно однакову кількість новин за день на спільних періодах в обох колекціях – 1,36 і 1,42 повідомлення на день, що в об'єднаному ряді дорівнювало б 2,78, проте після 10.06.2020 колекція All_external обривається, тому кількість новин за день різко зменшилась би приблизно вдвічі. Для цього дослідження це критично важливо, бо розрив між періодами колекцій припадає на середину тестового періоду і міг би спотворити обсяг повідомлень, який є однією з пояснювальних змінних. Обмеження періоду до 2020 року призвело б до суттєвої втрати текстової інформації; масштабування даних також відкинуто через відмінність коефіцієнтів для різних активів, а для деяких – через неможливість їх обчислення. Тому прийнято рішення відмовитись від колекції All_external задля уникнення спотворення результатів, хоча при цьому втрачається від 825 до 3137 повідомлень на актив за досліджуваний період. У ході фільтрації набору за тікерами отримано вибірку з 49326 повідомлень, а після дедуплікації новин і відкидання повідомлень поза ціновим рядом – 48666. У дослідженні використовувались лише заголовки новин, середня довжина яких становить 59 символів.

Додатковим обмеженням є те, що в 99,9 % повідомлень вказано лише дату публікації без часу доби, тому розділення новин на ті, що вийшли до закриття біржі та після нього, для цього набору даних неможливе. Лише для 59 повідомлень з наявним часом публікації застосовано правило, за яким повідомлення після закриття торгів відноситься до наступного дня. Для решти записів береться дата публікації, зсунута вперед до найближчого доступного дня торгівлі. Усі новинні змінні подаються на вхід моделям із затримкою на один день, що унеможливлює заглядання моделей наперед.

Усі відібрані активи належать до галузі «Напівпровідники та обладнання для їх виробництва»: з одного боку, спільні джерела ризику роблять їх порівнянними, а з іншого – всередині галузі вони суттєво відрізняються як річною волатильністю, від 20,5 % у TSM до 44,2 % в AMD, так і бізнес-моделями – проєктування мікросхем (NVDA, AMD, QCOM, TXN), виробництво пам'яті (MU), контрактне виробництво (TSM) і випуск обладнання (AMAT), тоді як INTC поєднує проєктування й виробництво.

Активи обирались насамперед з огляду на достатнє покриття новинами. У використаному наборі даних діє обмеження приблизно 9000 записів на один тікер, тому для компаній, які дуже активно висвітлювались у новинах, записів за старі роки немає: MSFT має покриття лише з квітня 2022, AAPL – з червня 2022, а LRCX, KLAC і AVGO – взагалі по сім записів кожна. Тобто довга історія і щільний потік новин для цього набору даних є несумісними.

Тому активи поділено на дві групи: компанії з довгою історією і нижчою щільністю покриття (QCOM, MU, TSM, AMAT, TXN) та компанії зі щільним потоком (AMD, INTC, NVDA). Найбільшу щільність повідомлень має NVDA – 14,54 повідомлення на день і 0,1 % днів без новин, а найнижчу TXN – 0,90 на день і 52,8 % порожніх днів, наведено у Таблиці 1:

Таблиця 1. **Склад вибірки**

Актив	Кількість новин	Початок діапазону покриття	Кількість новин за день	Медіана	Максимальне значення	Частка днів без новин	Річна волатильність
QCOM	6001	01.02.2010	1,69	1	27	38,6%	25,2%
MU	6451	07.02.2013	2,34	1	30	20,3%	38,0%
TSM	3895	15.04.2010	1,13	1	15	47,3%	20,5%
AMAT	3477	15.04.2010	1,00	0	17	53,5%	28,9%
TXN	3166	09.12.2009	0,90	0	19	52,8%	22,0%
AMD	8952	12.12.2016	4,96	4	49	2,4%	44,2%
INTC	8675	26.12.2017	5,68	5	49	1,1%	23,6%
NVDA	8709	17.08.2021	14,54	10	95	0,1%	35,3%

***Джерело*: розроблено автором**

Ціни активів отримано з відкритого ресурсу Yahoo Finance [7] за період 04.01.2010 – 29.12.2023, що становить 3522 торгові дні на кожен актив, а разом утворює 28176 спостережень. Проведено перевірки даних на цілісність, відсутність пропусків та аномальних значень. Для уникнення фіктивного обвалу цін було проведено їх коригування на дроблення акцій.

У відкритому доступі немає внутрішньоденних котирувань, потрібних для обчислення реалізованої волатильності, тому було використано проксі-оцінку з денних значень цін відкриття, закриття, мінімуму й максимуму, а саме застосовувався оцінювач Гармана-Класа, який обчислюється за формулою:

$$\sigma^2{}_t = 0{,}5 \cdot \left(ln\left(\frac{H_t}{L_t}\right)\right)^2 - (2 \cdot ln2 - 1) \cdot \left(ln\left(\frac{C_t}{O_t}\right)\right)^2, \quad (1)$$

де H_t – це максимальна ціна за день t; L_t – мінімальна ціна за день t; C_t – ціна на момент закриття торгів за день t; O_t – ціна на момент відкриття торгів за день t.

Цей оцінювач є у 7,4 раза ефективнішим за оцінювач на основі квадрата денної дохідності [8].

Цільова змінна обчислюється за формулою:

$$logRV_t = ln(\sigma^2{}_t). \quad (2)$$

Логарифмування дозволило наблизити розподіл до нормального: асиметрія становить від 0,08 до 0,31, а ексцес – від -0,01 до 0,64, що обґрунтовує застосування методу найменших квадратів для оцінювання. Для порівняння, ексцес самих дохідностей становить від 3,99 до 11,57.

Обмеженням оцінювача Гармана-Класа є те, що він не враховує нічний розрив у даних, на який припадає від 32,6 % до 49,9 % денної дисперсії на розглянутих активах. Оцінювачі, які враховують ці розриви, використано для перевірки результатів на стійкість.

Перевірку проведено на чотирьох оцінювачах волатильності з денних цін: Гармана-Класа, Паркінсона, Роджерса-Сатчелла та модифікації Гармана-Класа, доповненої квадратом нічного розриву. Перші три враховують лише внутрішньоденний рух, четвертий охоплює всю добу. За результатами перевірки модель без новин скрізь передостання, а відхилення обсягу посідає перше місце для трьох оцінювачів із чотирьох.

Автокореляція цільової змінної спадає повільно і навіть на лагу 132 торгові дні лишається на рівні від 0,05 до 0,2 залежно від активу, тоді як автокореляція квадратів дохідностей практично зникає на лагу 22 дні, що зображено на Рис. 1.

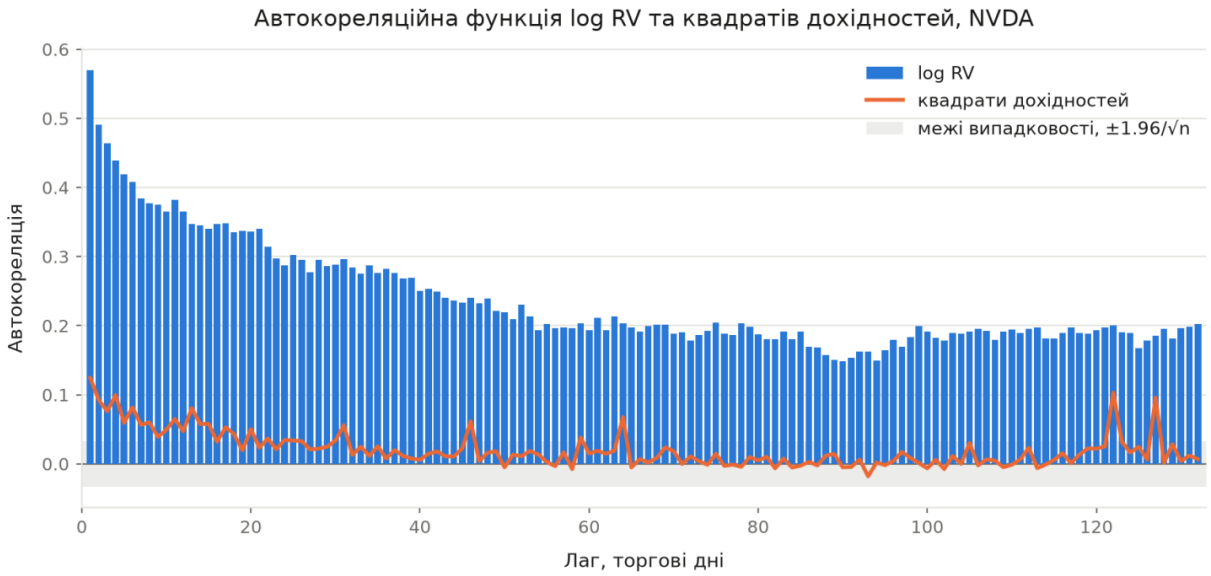

Рис. 1. **Автокореляційна функція логарифма реалізованої волатильності та квадратів дохідностей на прикладі NVDA**

Джерело: **розроблено автором**

Проведений аналіз підтверджує доцільність використання оцінювача на основі цін OHLC: він містить менше шуму, тому довга пам'ять проявляється на ньому виразніше, ніж на квадратах денних дохідностей. Збереження залежності на піврічному горизонті свідчить, що прогноз не повинен спиратися лише на дані попереднього дня. Тому в дослідженні використано модель HAR із трьома горизонтами усереднення: її тижнева й місячна компоненти дають змогу врахувати інформацію віддаленіших лагів [9].

Формально модель HAR можна записати так:

$$logRV_t = \beta_0 + \beta_d \cdot log\ RV_{t-1} + \beta_w \cdot m_1 + \beta_m \cdot m_2 + \varepsilon_t, \quad (3)$$

де m_1 – середнє значення цільової змінної за попередній тиждень; m_2 – середнє значення цільової змінної за попередній місяць.

Параметри моделі оцінюються методом найменших квадратів. Новинні змінні на вхід моделі подаються як додаткові регресори рівняння.

Першою базовою моделлю порівняння обрано наївний прогноз за попереднім днем: модель, яка його не перевершує, не має практичної цінності. Модель ковзного середнього використано для перевірки доцільності врахування динаміки. Моделі GARCH(1,1), EGARCH і GJR застосовано для порівняння з іншим класом моделей [10], [11], [12]; вони оцінювались не за log RV_t, а за дохідностями з розподілом Стьюдента для врахування важких хвостів розподілу. Моделі EGARCH і GJR розглянуто додатково, щоб урахувати ефект левериджу.

Оскільки прогнози цих моделей вимірюють іншу величину – умовну дисперсію дохідностей від закриття до закриття, тоді як цільова змінна побудована з внутрішньоденного розмаху, – для коректного порівняння здійснено їх вирівнювання за рівнем. Без цього моделі сімейства GARCH програвали б через одиниці вимірювання, а не через якість прогнозу. Зсув оцінювався лише на навчальному вікні, причому кожна модель отримувала власний: спільний зсув приховав би те, що моделі зміщені по-різному, і порівняння між ними стало б некоректним.

Як було сказано раніше, новинні змінні подаються на вхід як додаткові регресори, а питання дослідження звужується до порівняння різних рівнів складності їх обробки. Перший рівень не вимагає обробки тексту взагалі.

Використовувався простий підрахунок обсягу новин, а також відхилення цього обсягу від очікуваного рівня:

$$n_t = ln(1 + N_t), \tag{4}$$

$$s_t = n_t - \frac{\sum_{i=1}^{22} n_{t-i}}{22}, \tag{5}$$

де N_t – кількість повідомлень за день t. Логарифмування застосовано через сильну скошеність розподілу.

Варто зазначити, що на першому рівні розглядалося три варіанти: тільки обсяг новин, тільки відхилення від очікуваного рівня і обидві змінні разом.

Наступним рівнем обробки є VADER – словниково-правиловий метод, який не навчався на даних і не враховує контексту, а тому працює майже миттєво. Використовується заготовлений словник слів з визначеною тональністю, а також правила для підсилювачів, заперечень і пунктуації. Як результат, метод повертає оцінки заголовків новин у діапазоні від -1 до 1 [13]. Обмеженням цього рівня є те, що інструмент створювався для оцінювання текстів із соціальних мереж, а тому фінансова лексика йому незнайома.

На противагу попередньому методу, наступний рівень обробки FinBERT є нейронною мережею з архітектурою трансформера, донавченою на фінансових текстах. Це дає їй змогу врахувати специфіку фінансової лексики. Метод оцінює слово в контексті всього речення, тому потребує значних обчислювальних ресурсів порівняно з попередніми рівнями [14]. Мережа повертає три ймовірності, але для коректності порівняння з попереднім рівнем, який також оцінює тональність і повертає одне число, їх було зведено до різниці ймовірностей позитивного й негативного класів.

Обчислювальну вартість рівнів обробки виміряно на повному наборі даних новин з 48666 повідомлень (наведено у Таблиці 2). Обчислення виконувались на центральному процесорі Intel Core i7-10750H. Для реалізації методу обробки FinBERT було використано модель ProsusAI/finbert з розміром батчу 64 послідовності і максимальною довжиною повідомлення у 64 токени.

Рівні відрізняються як часом, потрібним на їх виконання, так і кількістю пам’яті, необхідної для їх роботи, причому рівні не рівновіддалені. Так простий підрахунок обсягу новин у 17 разів швидший за словниковий метод визначення тональності, але вже третій

рівень у 990 разів повільніший за другий. Відмінність також присутня між кількістю параметрів необхідних для опрацювання даними методами.

Таблиця 2. **Порівняння обчислювальної вартості обробки новин**

Рівень обробки	Час, с	Заголовків/с	Пам'ять, МБ	Розмір моделі, МБ	К-сть параметрів
1	0,09	521020	0,63	0	0
2	1,58	30830	0,75	0,4	7506
3	1563,51	31	93,26	437,94	109484547

Джерело: **розроблено автором**

Для другого та третього рівнів на основі оцінок тональності окремих заголовків було обчислено чотири денні агрегати, а саме середню тональність за день, розкид тональності за день, середній модуль тональності і частку повідомлень з оцінкою нижче -0,1. Кожна специфікація цих рівнів містить відхилення обсягу новин з першого рівня обробки, а також один із трьох наборів змінних: середню тональність із розкидом, лише розкид або середній модуль із часткою негативних. Для кожного набору агрегатів специфікації цих рівнів мають однакову кількість змінних і різняться лише способом оцінювання тональності кожного заголовка, тому різниця між ними відображає саме якість цього оцінювання, а не кількість оцінюваних коефіцієнтів. При цьому день, у який не виходило повідомлень, має всі нульові змінні; день з одним повідомленням має нульовий розкид; а якщо день перебуває поза межами періоду покриття, на місце його змінних ставився пропуск, бо заповнення нулями прирівняло б такий день до дня всередині періоду без новин.

Схема оцінювання моделей була визначена ще до їх побудови. Ковзне вікно навчання становить 1000 днів, його кінець завжди передує першому дню прогнозу, прогноз дається на 1 день вперед. Тестовий період обрано починаючи з початку 2019 року. Переоцінка параметрів моделі здійснюється кожні 22 дні, що є компромісом на користь швидкості навчання. Для моделей порівняння сімейства GARCH умовна дисперсія оновлюється на основі щоденних даних, а коефіцієнти перераховуються раз на місяць.

Для дослідження було створено декілька конфігурацій моделі, наведених у Таблиці 3. Для моделей без новин використовувались усі вісім активів, проте порівняння рівнів обробки новин потребує, щоб навчальне вікно було повністю покрите новинними змінними, а межі цього покриття різняться залежно від активу, що розглянуто вище. Тому для основної конфігурації з ковзним вікном навчання 1000 днів було використано лише п'ять активів. Однак активи з пізнім покриттям не виключено з дослідження – їх використано в інших конфігураціях.

Основною обрано саме конфігурацію з ковзним вікном навчання 1000 днів, оскільки на коротшому вікні, яке покривало б усі активи, зростає похибка оцінювання коефіцієнтів при новинних змінних через недостатню кількість спостережень. Тому вибір коротшого вікна послабив би результат дослідження через свідомо гіршу схему оцінювання.

Слабким місцем основної конфігурації є мала щільність новинного потоку – від 0,9 до 2,3 повідомлення за день, через що достовірно виміряти тональність складно. Тому дослідження потребує перевірки в сприятливіших для тональності умовах.

Серед активів, відібраних для дослідження, щільний потік новин мають три: AMD (4,96 повідомлення на день), INTC (5,68) і NVDA (14,54). Останній, проте, через обмеження набору даних має короткий період покриття новинами, що при ковзному вікні 500 днів залишило б лише 78 днів для тестування, тому його до цієї конфігурації не включено. Таку довжину вікна обрано тому, що 500 днів – найбільше вікно, яке AMD встигає заповнити до стандартного початку тестового періоду, завдяки чому її тестовий період збігається з періодом основної конфігурації. Для INTC такої кількості днів не набирається, тому її тестовий період починається з 08.01.2020.

Конфігурацію з ковзним вікном 250 днів застосовано, щоб перевірити результати моделювання на всіх восьми активах дослідження.

Варто зазначити, що довжина вікна не виводиться з даних однозначно, тому проведено перевірки на різних довжинах – від 300 до 1000 днів. Упорядкування специфікацій при цьому зберігається: відхилення обсягу новин лишається найкращою ознакою, а модель без новин – у кінці рейтингу.

Таблиця 3. **Специфікації моделей у дослідженні**

Конфігурація	Вікно, днів	Активи	Початок тесту	Прогнозів на модель
Моделі без новинних змінних	1000	усі вісім	02.01.2019	10064
Основна: рівні обробки новин	1000	QCOM, MU, TSM, AMAT, TXN	02.01.2019	6255
Активи зі щільним потоком	500	AMD, INTC	02.01.2019 і 08.01.2020	2253
Розширена вибірка	250	усі вісім	різний	9087

Джерело: **розроблено автором**

Оскільки цільова змінна не є спостережуваною величиною і виміряна з похибкою, для порівняння якості прогнозування потрібна метрика, стійка до шуму в цільовій змінній [15]. Тому було використано метрику QLIKE, яка обчислюється за формулою:

$$QLIKE = \frac{\sum_1^n (r_t - \ln r_t - 1)}{n}, \tag{6}$$

$$r_t = \frac{\sigma_t^2}{\hat{\sigma}_t^2} \tag{7}$$

де σ_t^2 – оцінка дисперсії, отримана за формулою (1) за день t; $\hat{\sigma}_t^2$ – прогноз дисперсії за день t; n – кількість спостережень у прогнозі.

Вираз під сумою дорівнює нулю лише тоді, коли прогноз точно збігається з фактичним значенням, а в решті випадків є додатним, тому менше значення QLIKE відповідає кращому прогнозу. QLIKE є асиметричною, тобто заниження волатильності за цією метрикою карається сильніше за завищення.

Варто зазначити, що QLIKE обчислюється від дисперсії, а експоненціювання прогнозу повертає медіану замість середнього, через що значення дисперсії систематично занижується. Оскільки QLIKE карає заниження сильніше за завищення, без поправки всі моделі отримали б штраф за це зміщення. Тому при зворотному перетворенні до прогнозу додавалася половина дисперсії залишків моделі, оцінена на навчальному вікні. На відміну від вирівнювання за рівнем, ця поправка застосовується однаково до всіх моделей.

Як додаткові метрики застосовано середньоквадратичну похибку MSE, середню абсолютну похибку MAE і коефіцієнт детермінації, обчислений поза навчальною вибіркою, R^2_oos. Усереднення метрик проведено за спостереженнями, а не за активами.

Як тест значущості використано тест Дібольда-Маріано (далі – тест DM) [16] з двома поправками: Ньюї-Веста на автокореляцію ряду різниць втрат і Гарві-Лейборна-Ньюболда на малу вибірку з розподілом Стьюдента. Нульова гіпотеза полягає в тому, що моделі однаково точні. Розглядається ряд різниць втрат за кожен день, і перевіряється, чи його середнє значуще відрізняється від нуля; від'ємне значення статистики означає, що точнішою є перша модель. Різницевий ряд розглядається для всіх пар активів і днів прогнозування, що підвищує потужність тесту, проте є коректним лише за умови слабкої кореляції між активами.

Перевірено, що середня кореляція між ними становить 0,039, тобто є незначною. Також для дев'яти новинних специфікацій було здійснено дев'ять перевірок гіпотез, при яких частина результатів може виявитись значущою випадково, тому було використано поправку на множинність за низхідною процедурою Холма, яка контролює ймовірність хоча б одного хибного відкидання в усьому сімействі, та за процедурою Бенджаміні-Хохберга, яка контролює очікувану частку хибних відкриттів серед відхилених гіпотез. Сімейство утворюють дев'ять порівнянь новинних специфікацій зі специфікацією без новин.

Також здійснено перевірку коректності реалізації моделювання. Спершу перевірено відсутність заглядання в майбутнє: дані майбутнього помножено на 3, при цьому ознаки минулого не змінились. На штучно створених цінах з відомою волатильністю перевірено константи у формулах оцінювачів – усі вони коректно відновили волатильність. Окремо підтверджено, що тест DM не виявляє різниці між ідентичними моделями, а частка хибних відкидань становить близько 5% при номінальному рівні значущості 5 %. Варто зазначити, що всі оцінювачі на основі розмаху систематично занижують показник, і чим рідші угоди за активом, тим сильніше це заниження. Причина в тому, що максимум і мінімум спостерігаються лише в моменти здійснення операцій, тому оцінка волатильності майже завжди буде меншою за справжню.

Результати показали, що модель HAR за всіма показниками перевершує базові моделі порівняння, що наведено у Таблиці 4.

Таблиця 4. **Порівняння моделей без новинних змінних**

Модель	QLIKE	RMSE (log RV)	R^2_oos	Кореляція	MAE, в.п. річної волатильності	Медіанна відносна похибка
HAR	**0,2773**	**0,7126**	**0,452**	**0,673**	**8,81**	**26,5%**
EGARCH	0,3039	0,7475	0,397	0,635	9,28	27,8%
GARCH	0,3129	0,7680	0,363	0,614	9,61	28,4%
GJR	0,3160	0,7687	0,362	0,617	9,52	28,2%
Наївний прогноз	0,4092	0,8748	0,174	0,587	12,31	34,3%
Ковзне середнє	0,4936	0,8956	0,134	0,435	10,54	31,3%

Джерело: **розроблено автором**

Усі пари моделей мають значущі відмінності, окрім GARCH і GJR ($p = 0,115$), причому EGARCH виявилась кращою за обидві. Наївний прогноз і модель ковзного середнього помітно програють решті моделей, що обґрунтовує доцільність використання складніших моделей. Теоретичною верхньою межею R^2_oos є одиниця, і чим ближче до неї значення, тим краще модель описує дані; значення 0,452 є типовим для прогнозування денної волатильності і не свідчить про низьку якість прогнозу. Варто також додати, що оскільки цільова змінна виміряна з похибкою, верхня межа коефіцієнта детермінації стає меншою за одиницю.

З Таблиці 5 видно, що найвищу кореляцію з цільовою змінною має обсяг повідомлень за день – 0,298, тоді як середня тональність, виміряна за допомогою FinBERT, не корелює з нею взагалі. Розкид тональності зі значеннями 0,220 і 0,218 корелює істотно сильніше за середню тональність зі знаком, для якої кореляція становить 0,087 і -0,004.

Як було сказано раніше, для першого рівня обробки досліджено три варіанти включення змінних. Специфікація, що містить обидві змінні, при короткому вікні поступається специфікаціям з однією змінною, оскільки ці змінні сильно корельовані між собою (0,62-0,85 за активами). Обсяг новин несе менше інформації, ніж відхилення обсягу, що і наведено у Таблиці 6.

Таблиця 5. **Кореляція новинних змінних з логарифмом волатильності**

Змінна	Кореляція з log RV
Обсяг повідомлень	0,298
Розкид тональності, FinBERT	0,220
Розкид тональності, VADER	0,218
Середній модуль тональності, VADER	0,140
Середній модуль тональності, FinBERT	0,140
Частка негативних, FinBERT	0,115
Середня тональність, VADER	0,087
Відхилення обсягу	0,074
Частка негативних, VADER	0,050
Середня тональність, FinBERT	-0,004

Джерело: **розроблено автором**

Таблиця 6. **Варіанти першого рівня за різної довжини навчального вікна**

Варіант	QLIKE, вікно 1000	p	QLIKE, вікно 250	p
Тільки відхилення обсягу	0,2809	0,002	0,2887	0,095
Обидві змінні	0,2812	0,039	0,2942	0,007 (гірше)
Тільки обсяг	0,2819	0,122	0,2902	0,927
Без новин	0,2830	-	0,2901	-

Джерело: **розроблено автором**

Результати порівняння для основної конфігурації моделі наведено у Таблиці 7:

Таблиця 7. **Порівняння рівнів обробки новин**

Специфікація	QLIKE	MSE(log)	MAE(log)	p з L0
L1 відхилення обсягу	**0,2809**	**0,5134**	**0,5658**	**0,002**
L2 VADER-розкид	0,2812	0,5140	0,5662	0,016
L1 обидві	0,2812	0,5145	0,5664	0,039
L3 FinBERT-розкид	0,2813	0,5144	0,5664	0,019
L3 FinBERT-інтенсивність	0,2814	0,5141	0,5661	0,040
L2 VADER	0,2817	0,5147	0,5667	0,112
L3 FinBERT	0,2818	0,5151	0,5669	0,113
L1 обсяг	0,2819	0,5149	0,5666	0,122
L2 VADER-інтенсивність	0,2820	0,5151	0,5668	0,196
L0 без новин	0,2830	0,5153	0,5664	-
EGARCH	0,3087	0,5639	0,5925	< 0,001

Джерело**: розроблено автором**

Найкращою виявилась специфікація першого рівня обробки з включенням лише відхилення обсягу новин (Таблиця 7). Усі специфікації з новинами перевершили модель EGARCH, яка була найкращою серед базових моделей після HAR. VADER і FinBERT статистично нерозрізненні: p = 0,861 для основної агрегації та p = 0,694 для агрегації за розкидом. Жодна специфікація з тональністю новин не перевершує модель з відхиленням обсягу від очікуваного рівня, тобто додавання тональності до відхилення обсягу не покращує прогноз. Навіть у межах першого рівня відхилення обсягу спрацьовує значно краще за сам обсяг новин.

Варто зазначити, що тест DM, об'єднаний за парами «торговий день – актив», дає для найкращої специфікації значення 0,002, проте при перевірці за окремими активами різниця є

значущою лише для QCOM, а для решти значення коливаються від 0,088 до 0,907. При цьому статистика тесту від'ємна для всіх активів, тому можна стверджувати про однаковий напрямок ефекту новин, а окремо взятий актив просто не дає достатньої кількості спостережень для підтвердження на рівні 5 %. Коректність такого об'єднання підтверджує також низька крос-кореляція між активами, розглянута раніше.

Після проведення поправки на множинність було отримано наступні результати, наведені у Таблиці 8:

Таблиця 8. **Значущість специфікацій після поправки на множинні порівняння**

Специфікація	p без поправки	Поріг Холма	p за Холмом	p за Бенджаміні-Хохберга
L1 відхилення обсягу	0,0018	0,0056	0,0158	0,0158
L2 VADER-розкид	0,0158	0,0062	0,1261	0,0576
L3 FinBERT-розкид	0,0192	0,0071	0,1344	0,0576
L1 обидві	0,0394	0,0083	0,2365	0,0711
L3 FinBERT-інтенсивність	0,0395	0,0100	0,2365	0,0711
L2 VADER	0,1124	0,0125	0,4498	0,1366
L3 FinBERT	0,1125	0,0167	0,4498	0,1366
L1 обсяг	0,1215	0,0250	0,4498	0,1366
L2 VADER-інтенсивність	0,1959	0,0500	0,4498	0,1959

Джерело: **розроблено автором**

До проведення поправки на множинність значущими виглядали п'ять новинних специфікацій, після поправки статистично значущою лишається лише специфікація першого рівня обробки з відхиленням обсягу новин. Хоча ці два методи контролюють різні величини, обидва дають однаковий результат.

Щодо абсолютних метрик, ефект від новин на них майже непомітний: R^2_oos зростає з 0,4012 до 0,4035, а MAE у пунктах волатильності практично не змінюється. Тобто новинний ефект є незначним за величиною, але статистично значущим для найпростішої специфікації після поправки на множинність.

Порівняння специфікацій усередині рівнів обробки 2 і 3 показало (наведено у таблиці 7), що більший вплив на результат має не обрана модель оцінювання тональності, а спосіб її агрегації. Заміна середньої тональності на розкид або інтенсивність підвищила якість прогнозу і до поправки на множинність специфікації були значущими, проте після здійснення поправки значущості не досягається. Отриманий результат узгоджується з наведеними вище міркуваннями про невідповідність знакової величини незнаковій цілі.

На інших конфігураціях, з ковзним вікном 500 і 250 днів, ситуація майже ідентична (Таблиця 9).

Результати свідчать, що невисока щільність новин в основній конфігурації не є причиною отриманих результатів, оскільки для конфігурації з ковзним вікном 500 днів порядок специфікацій зберігся. Підтвердилось також припущення про те, що збільшення кількості оцінюваних коефіцієнтів негативно впливає на результати: у конфігурації з вікном 250 днів складніші специфікації обробки новин стають значуще гіршими за модель без новин узагалі.

Для оцінювання масштабу новинного ефекту новинні змінні додавались до моделі HAR по одній, з вимірюванням приросту коефіцієнта детермінації в межах усієї вибірки. Коефіцієнт детермінації для моделі HAR при цьому дорівнює 0,5113 (наведено у Таблиці 10). Власна історія волатильності дає приблизно у 60 разів більший приріст, ніж усі новинні змінні разом. Розкид тональності, обчислений за FinBERT, дає майже вдвічі більший приріст, ніж за методом VADER, що вказує на здатність складнішої моделі краще видобувати сигнал з новин, проте це не привело до суттєвого покращення прогнозу.

Таблиця 9. **Порівняння рівнів обробки новин для конфігурації з вікном 500 і 250**

Специфікація	Вікно 500 (AMD, INTC) QLIKE	p	Вікно 250 (усі вісім) QLIKE	p
L1 відхилення обсягу	**0,2584**	**0,037**	**0,2887**	**0,095**
L2 VADER-розкид	0,2589	0,093	0,2897	0,669
L3 FinBERT-розкид	0,2590	0,087	0,2899	0,846
L3 FinBERT-інтенсивність	0,2593	0,147	0,2929	0,061
L1 обсяг	0,2594	0,169	0,2902	0,927
L3 FinBERT	0,2598	0,282	0,2920	0,140
L1 обидві	0,2600	0,410	0,2942	0,007 (гірше)
L2 VADER	0,2607	0,664	0,2927	0,038 (гірше)
L0 без новин	0,2616	-	0,2901	-
L2 VADER-інтенсивність	0,2624	0,712	0,2942	0,003 (гірше)
EGARCH	0,3905	< 0,001	0,3215	< 0,001

Джерело: **розроблено автором**

Таблиця 10. **Приріст R^2 від додавання новинних змінних до HAR**

Змінна	Коефіцієнт	p	Приріст R^2
Відхилення обсягу	+0,0865	< 0,0001	+0,00175
Розкид тональності, FinBERT	+0,1346	< 0,0001	+0,00087
Розкид тональності, VADER	+0,1586	0,00003	+0,00050
Середній модуль, FinBERT	+0,0745	0,0008	+0,00029
Середній модуль, VADER	+0,0552	0,091	+0,00007
Середня тональність, VADER	+0,0372	0,230	+0,00004
Середня тональність, FinBERT	-0,0194	0,357	+0,00002
Усі новинні разом			+0,00204
Тижнева і місячна компоненти HAR			+0,12102

Джерело: **розроблено автором**

Приріст від усіх новинних змінних разом становить +0,00204. Для порівняння, власна історія волатильності дає приріст у +0,12102. Тобто інформація з новин до поправки на множинність дає статистично значущий, але кількісно малий внесок.

Висновки. У ході роботи було порівняно різні рівні складності обробки новинного потоку в задачі прогнозування волатильності за єдиною схемою оцінювання. Встановлено, що межа, після якої точність перестає зростати, проходить по першому, найпростішому рівні обробки новин. Найбільш інформативним виявилось відхилення обсягу новин від очікуваного рівня, тоді як перехід до наступних рівнів обробки, а саме до оцінювання тональності, приросту не дає. Після поправки на множинність жодна специфікація з тональністю не є значуще кращою за модель без новин. Другий і третій рівні виявились статистично нерозрізненними, а перехід між рівнями збільшує обчислювальні витрати в 990 разів і при цьому не збільшує точність. Водночас більший ефект має вибір способу агрегації тональності за день, ніж модель її оцінювання. Загалом ефект від новин виявився малим: власна історія волатильності лишається основним джерелом інформації для прогнозу, а внесок новинних змінних приблизно у 60 разів менший. Встановлено також, що на короткому навчальному вікні складніші специфікації, навпаки, шкодять порівняно з моделями без новин через додаткові коефіцієнти, вартість оцінювання яких перевищує користь від отриманої інформації. До обмежень дослідження можна віднести аналіз лише восьми активів однієї галузі, використання проксі-оцінки цільової змінної, оцінювання тональності лише за заголовками новин, а також відсутність у наборі даних часу публікації повідомлень, через що розділити новини на ті, що виходили до та після закриття торгів, було неможливо. Перспективами подальших досліджень є оцінювання ринкового ризику на основі отриманих

прогнозів, перевірка результатів на інших галузях, а також використання великої мовної моделі для оцінювання тональності новинного потоку.

Конфлікт інтересів: автор декларує, що не має конфлікту інтересів, зокрема фінансового, особистого, авторського чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті.

Фінансування: дослідження проводилося без фінансової підтримки

Доступність даних: усі дані доступні в цифровій або графічній формі в основному тексті рукопису

Використання засобів штучного інтелекту: під час підготовки рукопису використовувався інструмент Claude, Opus 5 штучного інтелекту (ШІ) для граматичної та стилістичної перевірки тексту. Усі фрагменти, опрацьовані за його допомогою, були перевірені й відредаговані автором. ШІ не використовувався для генерації тексту, формул і таблиць, формування наукових положень, отримання результатів або формулювання висновків. Наукові положення, результати та висновки є власним інтелектуальним внеском автора

СПИСОК ЛІТЕРАТУРИ

1. Dong, Z., Fan, X. & Peng, Z. “FNSPID: A comprehensive financial news dataset in time series”. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2024. p. 4918–4927. DOI: https://doi.org/10.1145/3637528.3671629.

2. Audrino, F., Sigrist, F. & Ballinari, D. “The impact of sentiment and attention measures on stock market volatility”. *International Journal of Forecasting.* 2020; 36 (2): 334–357. DOI: https://doi.org/10.1016/j.ijforecast.2019.05.010.

3. Bodilsen, S. T. & Lunde, A. “Exploiting news analytics for volatility forecasting”. *Journal of Applied Econometrics*. 2025; 40 (1): 18–36. DOI: https://doi.org/10.1002/jae.3095.

4. Lei B. & Song Y. “Volatility forecasting for stock market incorporating media reports, investors' sentiment, and attention based on MTGNN model”. *Journal of Forecasting*. 2024; 43 (5): 1706–1730, https://www.scopus.com/pages/publications/85186609626.
DOI: https://doi.org/10.1002/for.3101

5. Yang T., Zhuo S. & Yang Y. “Investor attention fluctuation and stock market volatility: Evidence from China”. *PLOS ONE*. 2023; 18 (11): e0293825, https://www.scopus.com/pages/publications/85178500612.
DOI: https://doi.org/10.1371/journal.pone.0293825

6. Léber, D. & Egyed, B. “The sentiment augmented GARCH-LSTM hybrid model for value-at-risk forecasting”. *Computational Economics.* 2025; 67 (1): 313–353. DOI: https://doi.org/10.1007/s10614-025-11042-8.

7. Yahoo Finance. “Historical market data”. – Available at: https://finance.yahoo.com – [Accessed: 20.08.2026].

8. Garman, M. B. & Klass, M. J. “On the estimation of security price volatilities from historical data”. *Journal of Business*. 1980; 53 (1): 67–78. DOI: https://doi.org/10.1086/296072.

9. Corsi F. “A simple approximate long-memory model of realized volatility”. *Journal of Financial Econometrics.* 2009; 7 (2): 174–196. DOI: https://doi.org/10.1093/jjfinec/nbp001.

10. Bollerslev, T. “Generalized autoregressive conditional heteroskedasticity”. *Journal of Econometrics.* 1986; 31 (3): 307–327. DOI: https://doi.org/10.1016/0304-4076(86)90063-1.

11. Nelson D. B. “Conditional heteroskedasticity in asset returns: A new approach”. *Econometrica.* 1991; 59 (2): 347–370. DOI: https://doi.org/10.2307/2938260.

12. Glosten L. R., Jagannathan R. & Runkle D. E. “On the relation between the expected value and the volatility of the nominal excess return on stocks”. *The Journal of Finance.* 1993; 48 (5): 1779–1801. DOI: https://doi.org/10.1111/j.1540-6261.1993.tb05128.x.

13. Hutto, C. & Gilbert, E. “VADER: A parsimonious rule-based model for sentiment analysis of social media text”. *Proceedings of the International AAAI Conference on Web and Social Media*. 2014; 8 (1): 216–225. DOI: https://doi.org/10.1609/icwsm.v8i1.14550.

14. Araci, D. “FinBERT: Financial sentiment analysis with pre-trained language models”. *arXiv*. 2019. DOI: https://doi.org/10.48550/arXiv.1908.10063.

15. Patton, A. J. “Volatility forecast comparison using imperfect volatility proxies”. *Journal of Econometrics.* 2011; 160 (1): 246–256. DOI: https://doi.org/10.1016/j.jeconom.2010.03.034.

16. Diebold, F. X. & Mariano, R. S. “Comparing predictive accuracy”. *Journal of Business & Economic Statistics.* 1995; 13 (3): 253–263. DOI: https://doi.org/10.1080/07350015.1995.10524599.

Отримано 30.07.2026
Отримано після перегляду 28.08.2026
Прийнято 04.09.2026

DOI: https://doi.org/10.15276/ict.03.2026.14
UDC 004.8:336.76

Comparison of news flow processing complexity levels in the task of stock volatility forecasting

Kateryna O. Nytska
ORCID: https://orcid.org/0009-0005-8027-0100; katerynanytska@gmail.com
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37, Beresteiskyi Ave.
Kyiv, 03056, Ukraine

ABSTRACT

The subject of this study is the forecasting of stock volatility using news as a potential source of additional information to improve the quality of forecasting. Methods for utilizing news flow vary in terms of both implementation complexity and computational cost; therefore, the objective of this study is to determine the level of complexity in news processing beyond which the increase in forecasting accuracy ceases. The study utilized data on stocks from companies in a single industry over a fourteen-year period and approximately forty-nine thousand news reports, with varying depths of news coverage across assets. The logarithm of the daily variance was forecast, estimated based on four price values per trading day: the opening, high, low and closing prices of the stocks. The baseline forecasting model was a heterogeneous autoregression with three main components corresponding to the daily, weekly, and monthly averaging horizons of the variable. To compare the results, conditional heteroskedasticity models, a moving average model, and a naive forecast serving as a lower bound for accuracy were also used. The news flow was fed into the model as an additional regressor at three levels of processing complexity: the first level – the number of news items per day and the deviation of this number from the expected level; the second – the sentiment of the news, determined using a lexicon-based method; the third – the sentiment determined using a financial language model. Sentiment was represented by the same set of variables in both methods; therefore, the difference between these levels reflects the quality of sentiment estimation rather than the number of estimated coefficients. Model parameters were estimated using a rolling window, and forecasts were generated one day in advance for days not included in the estimation. Since the target variable itself is measured with an error, the quality of the forecast was assessed using a metric that is robust to this error, and the difference in accuracy between the models was verified using a formal test for comparing predictive accuracy. It was found that the deviation of news volume from the expected level, rather than the sentiment of the news, was informative for the forecast. The difference between the lexicon-based method for detecting news sentiment and the financial language model is statistically insignificant for the stocks under consideration, despite the latter requiring four orders of magnitude more computational resources. After correction for multiple comparisons, only the simplest specification significantly improves the forecast; however, its contribution was approximately sixty times smaller than that of the stock's own volatility history. Prospects for further research include validating the obtained results in market risk assessment and applying a large language model as the next level of news flow processing for forecasting.

Keywords: Financial time series; volatility forecasting; heterogeneous autoregression; news flow; news sentiment analysis; lexicon-based sentiment analysis; financial language model; predictive accuracy comparison

ІНФОРМАЦІЯ ПРО АВТОРА

Ницька Катерина Олександрівна - студентка магістратури кафедри Математичних методів системного аналізу. Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», пр. Берестейський, 37. Київ, 03056, Україна
Галузь досліджень: системний аналіз фінансових ринків

Kateryna O. Nytska - Master's Student, Department of Mathematical Methods of System Analysis. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37 Beresteiskyi Ave. Kyiv, 03056, Ukraine

ORCID: https://orcid.org/0009-0005-8027-0100; katerynanytska@gmail.com
Research field: *systems analysis of financial markets*