

УДК 004.855:[336.76+004.032.26]

МЕТОД ПРОГНОЗУВАННЯ КРИЗ ФІНАНСОВИХ РИНКІВ НА ОСНОВІ ВЕКТОРНОЇ БАЗИ ДАНИХ ІСТОРИЧНИХ ПОДІЙ ТА СЕМАНТИЧНОГО АНАЛІЗУ З ВИКОРИСТАННЯМ LLM

Недодаєв В.А.

e-mail: vladyslav.nedodaiev@nure.ua

Науковий керівник – к.т.н., проф. Шевченко О.Ю.

Харківський національний університет радіоелектроніки, каф. III
м. Харків, Україна

This work is devoted to the development of a method for forecasting financial market crises based on a vector database of historical events. The proposed method employs a Multi-Factor Financial Foundation Model (MFFM) that fuses quantitative market indicators with semantic news embeddings into a unified vector representation. Unlike prior ML approaches to crisis prediction that rely on fixed feature sets and lack interpretability, the proposed method combines multimodal embeddings with retrieval-based reasoning. Walk-forward backtesting on data spanning 2000–2025 is used to evaluate the method against established baselines including VIX threshold, logistic regression, turbulence index, and absorption ratio.

З кожним роком зростає потреба у надійних системах раннього попередження фінансових криз. Традиційні методи прогнозування, такі як: порогові моделі на основі індексу волатильності VIX, логістичні регресії на кривій дохідності, моделі кредитних спредів - працюють із вузьким набором заздалегідь визначених ознак та не здатні врахувати мультимодальну, контекстно-залежну природу передвісників кризових явищ. Водночас зростаюча доступність ринкових даних у реальному часі, макроекономічних індикаторів та неструктурованих новинних потоків створює можливість для принципово нових методів виявлення криз, що поєднують кількісний та семантичний аналіз.

Ключова гіпотеза роботи полягає в тому, що фінансові кризи демонструють повторювані структурні патерни (каскади зараження, спіралі ліквідності тощо), які проявляються як схожі «відбитки» в різні часові періоди, навіть коли поверхневі тригери відрізняються. Гіпотеза вважається підтвердженою, якщо пошук найближчих сусідів у просторі вкладень забезпечує статистично значуще покращення якості прогнозування порівняно з пороговими моделями на окремих індикаторах.

Запропонований метод базується на трирівневій архітектурі. Перший рівень - мультифакторна фінансова модель (MFFM), яка кодує гетерогенні фінансові дані у єдиний простір вкладень через дві паралельні гілки. Кількісна гілка обчислює набір ознак із ринкових та макроекономічних джерел: волатильність, нахил кривої дохідності, кредитні спреди, міжбанківські ставки, співвідношення золото/акції тощо. Текстова гілка агрегує заголовки фінансових новин за часове вікно та кодує їх за допомогою попередньо навченого трансформера для генерації вкладень

речень [2]. Для історичних періодів, де цифрові новинні архіви відсутні, система працює виключно на кількісних ознаках. Обидві гілки об'єднуються у спільне вкладення за допомогою навченого модуля злиття, оптимізованого методами метричного навчання для забезпечення близькості подібних ринкових станів у векторному просторі.

Другий рівень – векторна база даних, яка зберігає вкладення історичних ринкових станів у векторному індексі разом із метаданими: тип кризи, рівень серйозності, уражені сектори, макроконтекст. Каталог охоплює анотовані кризові події за період 1970–2025 рр. на основі даних Reinhart–Rogoff та Laeven–Valencia [1, 3].

Третій рівень – LLM-агент, який працює за паттерном Retrieval-Augmented Generation (RAG) [4]. Його основна роль – забезпечення інтерпретованості: тоді як другий рівень генерує числову оцінку кризи, LLM синтезує структуроване пояснення на основі поточного кількісного контексту та топ-К історичних збігів, формуючи класифікацію типу кризи, оцінку серйозності та рекомендації з моніторингу.

Оцінювання проводиться за протоколом walk-forward бектестування (2000–2025) із суворою темпоральною дисципліною. Враховуючи суттєвий дисбаланс класів, основний акцент робиться на метриках, стійких до дисбалансу: AUPRC, Brier score, час випередження та частота хибних спрацювань на рік; AUROC наводиться для порівняльності з існуючими роботами. Запропонований метод порівнюється з базовими моделями: порогова модель VIX, логістична регресія, індекс турбулентності Крітцмана [5] та коефіцієнт абсорбції. Абляційні дослідження ізолюють внесок кожного компонента: кількісні ознаки проти текстових, лінійне злиття проти нелінійного, а також порівняння різних LLM-моделей.

Список використаних джерел:

1. Reinhart C. M., Rogoff K. S. This Time Is Different: Eight Centuries of Financial Folly. Princeton : Princeton University Press, 2009. 512 p.
2. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2019. P. 3982–3992. DOI: <https://doi.org/10.18653/v1/D19-1410>.
3. Laeven L., Valencia F. Systemic Banking Crises Database II // IMF Economic Review. 2020. Vol.68. P. 307–361. doi:10.1057/s41308-020-00107-3.
4. Lewis P., Perez E., Piktus A. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems (NeurIPS). 2020. Vol. 33. P. 9459–9474.
5. Kritzman M., Li Y. Skulls, Financial Turbulence, and Risk Management // Financial Analysts Journal. 2010. Vol. 66, No. 5. P. 30–41. DOI: <https://doi.org/10.2469/faj.v66.n5.3>.