Serhii Kanyhin, Postgraduate Student
Simon Kuznets Kharkiv National University of Economics
Kharkiv, Ukraine

DOI: https://doi.org/10.30525/978-9934-26-242-5-48

NATURAL LANGUAGE PROCESSING IN FORECASTING FINANCIAL MARKETS

ВИКОРИСТАННЯ ОБРОБКИ ПРИРОДНЬОЇ МОВИ У ПРОГНОЗУВАННІ ФІНАНСОВИХ РИНКІВ

Прогнозування фінансових ринків є важливим і складним завданням. Розвиток веб-технологій призвів до експоненціального зростання обсягів доступних даних. Це може стати ключем до розкриття мінливості фінансового ринку та сприяти точності прогнозування. Однак, такі дані, як правило, є неструктурованими. Хоча людина може легко їх інтерпретувати, проте вручну опрацювати їх у великій кількості досить складно. Важливо розуміти різні підходи до обробки таких даних, щоб оптимізувати ефективність прогнозування.

З точки зору попиту, ціни на фінансових ринках відображають очікування інвесторів щодо майбутньої вартості конкретного активу. Торговельна діяльність інвесторів базується на обробці інформації. Останнім часом розвиток веб-технологій сприяє зростанню обсягів текстових даних, і масово доступні дані можуть бути використані для виявлення непояснених коливань на фінансовому ринку та покращення якості прогнозування. Природна мова в тексті, як правило, неструктурована і може містити різнорідне логічне навантаження. Вона може бути легко інтерпретована людиною, але її обробка машиною є проблематичною. Таку величезну кількість текстових даних важко обробити людині вручну через брак часу, здібностей, енергії тощо. Винахід технології обробки природної мови (англ. natural language processing) дозволив розробити обчислювальні моделі, які дають змогу машині розуміти людську мову та автоматично вирішувати практичні завдання.

У попередніх дослідженнях прогнозування біржових показників на основі тексту здебільшого покладалися на обробку слів за допомогою таких підходів як «торба слів», N-грам, TF-IDF, тощо [4].

Однак ці підходи не враховують порядок слів та обробляють слова окремо.

Семантичний підхід долає вищезазначену проблему шляхом векторизації та є ключем до підвищення продуктивності моделей. Використання передавального навчання (англ. transfer learning) посилило семантичний підхід. При його використанні велика модель спочатку проходить попереднє навчання на великому корпусі тексту, а в подальшому додатково навчається під більш конкретні задачі. Це значно знизило складність застосування та зменшило витрати на обчислення, оскільки модель може бути ініціалізована з попередньо навченої моделі замість навчання з нуля. При використанні передавального навчання важливо, щоб домен моделі попереднього навчання не сильно відрізнявся від фінального. Велика розбіжність може призвести до погіршення ефективності. Наприклад, деякі моделі можуть розпізнавати слово «знецінення» як таке, що має негативний відтінок, хоча це слово може вказувати на сприятливу інвестиційну можливість. В зв'язку з цим, потрібно використовувати найбільш придатні попередньо навчені моделі або збільшувати обчислювальні витрати для вдосконалення фінальної моделі.

Сентиментальний підхід може визначати емоційну тональність тексту через векторизацію. При роботі з абзацами або цілими документами, стає проблематично трансформувати їх у вектор фіксованого розміру без певного зниження якості репрезентації, що, в свою чергу, погіршує ефективність моделей [1; 2; 3; 5]. Крім того, велика кількість інформаційного шуму призводить до зниження якості векторизації та погіршує кінцеву ефективність моделі [3]. До того ж, закодована інформація у векторній формі не може бути інтерпретована людиною. Нейронна мережа – це чорний ящик, що може перешкоджати процесу інтерпретації. З позиції фінансів можливість інтерпретації є важливою, оскільки це дозволяє зменшувати інвестиційні ризики та уникати фінансових втрат.

Подієво-орієнтований підхід дозволяє за допомогою метода графів створювати модель зв'язків всередині речень та абзаців для кращої інтерпретації [6]. Це важливо, та як нехтування дрібнозернистою інформацією є критичним і може призвести до введення в оману. Наприклад, речення «А добре, а Б погано» має одразу два твердження і його важко узагальнити. Як позитивна, так і негативна оцінка такого речення не є точною, і узагальнення до рівня речення призведе до втрати додаткової інформації.

Отримати якісну оцінку деяких видів тексту можна тільки при використанні окремих підходів [7]. Контент соціальних медіа, як правило, містить сильно виражені настрої та лексично неповний та добре аналізується за допомогою сентиментального підходу. Як приклад: коментар у соціальних мережах «Акція А хороша через вчорашнє оголошення». Експертні фінансові новини, як правило, мають нейтральне емоційне навантаження і є доволі чіткими та зрозумілими, що дозволяє використовувати подієво-орієнтований підхід. Виходячи з цього, дослідникам рекомендується враховувати такі властивості вхідного тексту, як суб'єктивність, довжина та деталізація. Підбір оптимального підходу на основі цих параметрів може дозволити отримувати більш точні прогнози.

Результати порівняння різних підходів обробки природньої мови у прогнозуванні фінансових ринків підсумовано у таблиці 1.

Таблиця 1

Порівняння різних підходів обробки природньої мови у прогнозуванні фінансових ринків

	Семантичний підхід	**Сентиментальний підхід**	**Подієво-орієнтований підхід**
Інтерпретація результатів	Ні	Ні	Так
Визначення впливу асиметричного контенту	Так	Так	Так
Визначення взаємопов'язаності контенту	Так	Ні	Так
Опрацювання дрібнозернистої інформації	Так	Так	Так
Оптимальний тип вхідних даних	Будь-який	Суб'єктивний	Об'єктивний
Деталізація вхідних даних	Будь-який	Будь-який	Необхідна деталізація
Довжина тексту	Погано працює з великими текстами	Погано працює з великими текстами	Добре працює з великими текстами
Напрям застосування	Новини, фінансові звіти	Соціальні мережі, форуми	Новини, фінансові звіти

Таким чином, в рамках дослідження було розглянуто підходи до прогнозування біржових показників на основі обробки природної мови, включаючи такі, що базуються на аналізі семантики, настроїв, подій та гібридні варіанти. Описано сильні та слабкі сторони кожного підходу, їх відмінності та напрями застосування. Було здійснено спробу допомогти людям з різним досвідом легко зрозуміти відповідні підходи. Закладено основу для подальших досліджень в напрямку прогнозування біржових показників.

Література:

1. Chen D., Zou Y., Harimoto K. Incorporating fine grained events in stock movement prediction. 2019.

2. Hu Z., Liu W., Bian J., Liu X., Liu T. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. 2018. P. 261–269.

3. Xing, F.Z.; Cambria, E.; Welsch, R.E. Natural language based financial forecasting: A survey. 2018. Vol. 50. Issue 1. Pp 49–73.

4. Verma, R., Verma, P. Noise trading and stock market volatility. 2007. Vol. 17, Issue 3. Pp. 231–243.

5. Chan J., Leow, S., Bea K. Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. 2022. Vol. 10, Pp. 1–17.

6. Chan J., Leow, S., Bea K. A Correlation-Embedded Attention Module to Mitigate Multicollinearity: An Algorithmic Trading Application. 2022. Vol. 10, 1231.

7. Jiang W. Applications of deep learning in stock market prediction: Recent progress. 2021. Vol. 184.