УДК 004.05

АНАЛІЗ МЕТОДІВ ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ В ЗАВДАННЯХ OSINT

А. П. Фегер[1], Д. В. Ланде[1]

[1] Навчально-науковий Фізико-технічний інститут

Анотація

Прогнозування часових рядів є важливою нішею в сучасному процесі прийняття рішень та вибору тактик, в перерізі з технологією OSINT такий підхід може допомогти передбачати події та дозволяти ефективно реагувати на них. Для цього було вибрано LSTM, ARIMA, LPPL (JLS), N-gram в якості методів прогнозування часових рядів, реалізовано їх прості форми на основі часового ряду кількісних згадувань про системи старлінк, отриманих та сформованих за допомогою технології OSINT. Базуючись на цьому, було досліджено їх загальну ефективність та можливості по застосуванню в комбінації з технологією OSINT для формування прогнозу майбутнього.

Ключові слова: time-series, prediction, osint

Вступ

В бізнесі, фінансах, логістиці, медицині, біології, хімії прогнозування є одним з найбільш прикладних методів науки який допомагає ефективно вирішувати типові задачі, та сприяє загальному розвитку. Разом з тим в сучасному світі актуальності в сфері також додають і останні нейронно-мережеві розробки в різних галузях кібербезпеки, безпосередньо таких як розвідка загроз, розпізнавання шкідливих програм, захист кінцевих точок, при побудові яких використовуються концепти ймовірнісного прогнозування.

Методи прогнозування часових рядів використовують історичні та поточні дані для прогнозування майбутніх значень протягом певного періоду часу або в певний момент у майбутньому. Аналізуючи наявні дані, які зберігались в минулому прогнозування допомагає зрозуміти майбутні тенденції, та дозволяє найбільш ефективно реагувати на них.

В сучасному світі якісно спроектована система прогнозування розв'язує руки та дає свободу в сфері точкового застосування навіть в рамках національної та кібербезпеки. З точки зору військової та цивільної безпеки така система дозволяє правильно будувати та корегувати тактику та стратегію на різних часових проміжках відповідно до прогнозованих подій.

Завдання дослідження полягає в першу чергу в створенні базису з найбільш ефективних методів прогнозування для ефективного проведення подальших досліджень, та якісному порівнянні між різними по природі методами. Проведення аналізу самих методів та використання їх в перерізі з технологією Open Source Intelligence (OSINT). Розглянутий часовий ряд для дослідження прогнозування складають колективну інформацію отриману за допомогою технологій OSINT.

Методика

Вибраний часовий ряд для дослідження репрезентує комплексну залежність кількості вибраних подій отриманих за допомогою технології OSINT від часу проміжком в рік. Вибраною подією для аналізу було представлено як залежність кількісної характеристики згадувань про системи старлінк на просторах інтернету до відповідного часового проміжку довжиною в один рік.

Для створення порівняльної бази в подальшому для навчання вибраних моделей було використано лише 333 днів з 364 в якості часового проміжку, де прогнозованим підсумком описувався останній місяць року.

Основними же сучасними підходами до прогнозування прийнято вважати такі як: нейронно-мережеві, статистичні, економетричні, лінгвістичні. Кожен з яких активно використовується в своїх галузях, в окремих випадках використовується комбінація з декількох підходів для отримання найбільш релевантних значень.

В якості типових сучасних представників описаних підходів, для дослідження часових рядів та побудови відповідних їм прогнозів було вибрано наступні методи:

- Long Short-Term Memory (LSTM) в якості найпоширенішого нейронно-мережевого методу;
- Autoregressive Integrated Moving Average (ARIMA) як найбільш широко вживаним статистичним методом;
- Log Periodic Power Law (LPPL) або Johansen-Ledoit-Sornette (JLS) як економетричного, який піддається критиці й не є популярним, проте використовується в окремих випадках;
- N-gram як лінгвістичного, який доволі сильно імплементовий в сучасні технології та життя.

По вибраним даним за допомогою моделей прогнозування можна передбачити кількісну характеристику вибраних подій та вирахувати середню похибку між реальними та прогнозованими даними.

Для визначення точності моделей прогнозування було взято середньоквадратичну похибку (MSE) за формулою (1), та корінь з середньоквадратичної похибки (RMSE) за формулою (2), як найбільш точні методи визначення подібних похибок.

$$MSE = \frac{1}{n}\Sigma_{i=1}^{n}\left(x_i - \overline{x_i}\right)^2 \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(x_i - \overline{x_i}\right)^2} \quad (2)$$

LSTM

LSTM – це різновид моделі рекурентної нейронної мережі, з тією відмінністю, що LSTM може обробляти довгі часові ряди даних. Крім того, звичайна рекурентна модель має проблему зникаючого градієнта для довгих послідовностей даних, при цьому LSTM може запобігти цій проблемі під час навчання.

Модель може згадувати попередні довготривалі часові ряди даних [1] і має автоматичний контроль для збереження релевантних ознак або відкидання нерелевантних ознак. Саме через ці фактори LSTM був вибраний серед інших рекурентних в якості методу для дослідження.

Для LSTM було використано його одношарову конфігурацію з 32 блоками (units), та омтимізатором Adam, з значенями batch size та epoch в 512, вихідні дані було (inverse) зворотньо трансформовано нормалізацією для отримання прогнозованого ряду.

ARIMA

ARIMA – це модель авторегресії з інтегрованим ковзним середнім, де AR частина показує, що часовий ряд регресується на власні минулі дані. MA частина показує, що помилка прогнозу є лінійною комбінацією минулих відповідних помилок. I частина показує, що значення даних були замінені різними значеннями порядку d для отримання стаціонарних даних, що є вимогою підходу моделі ARIMA.

Саме завдяки такій комплексності модель ARIMA є ефективною для перебору минулих даних за допомогою такого комбінованого підходу і допомагає ефективно прогнозувати майбутні точки часового ряду [2].

Для ARIMA було використано його одношарову конфігурацію з $p = 33$, $d = 2$, $q = 0$, значення яких було визначено шляхом проб відповідно до більш релевантного вихідного прогнозу.

LPPL

LPPL – або модель Йохансена-Ледойта-Сорнетта (JLS) – намагається діагностувати, визначати час і передбачати кінець фінансових «бульбашок», поширений термін в фінансовій сфері визначаючих кризові моменти при появі недовіри більшості учасників під час спекулятивного зростання.

Не дивлячись на широку критику [3], творці моделі надають мотивацію, побудовану на деяких природних припущеннях, включаючи нейтральні до ризику активи, раціональні очікування, локальну самопідсилюючу імітацію та ймовірнісні критичні моменти для розрахунку алгоритмом безпосередньо стадій розвитку «бульбашки» [4]. Завдяки цьому ми можемо побачити як вибраний алгоритм для прогнозування буде працювати з нетиповим для нього часовим рядом.

Для LPPL (JLS) було використано його модифікацію з використанням стратегії еволюційної адаптації коваріаційної матриці Covariance Matrix Adaptation Evolution Strategy (CMA-ES), що дає більш релевантні прогнозовані ряди.

N-gram

N-gram – представляють собою безперервну послідовність з N елементів із заданого набору текстів.

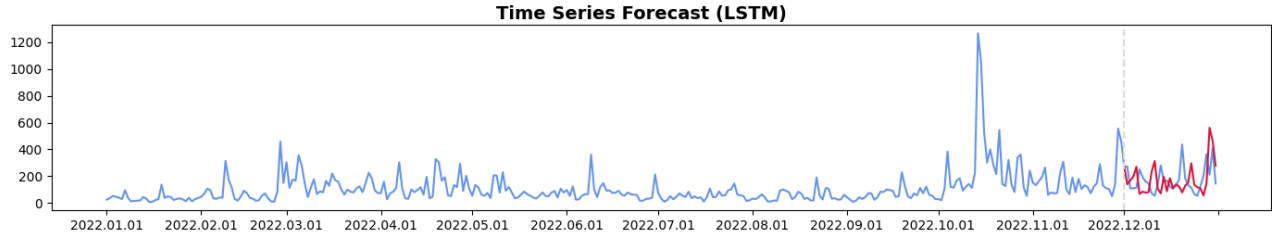
N-грами знайшли своє основне застосування в області імовірнісних мовних моделей. Вони оцінюють ймовірність появи наступного елемента в послідовності слів, це постало в основу теоретичного підходу щодо прогнозування часових рядів. Цей підхід до моделювання мови передбачає тісний взаємозв'язок між позицією кожного елемента в рядку, обчислюючи появу наступного слова по відношенню до попереднього та частоти їх появи. У широкому сенсі, такі елементи не обов'язково означають рядки слів, вони також можуть бути фонемами, складами або буквами [5], залежно від того, що саме потрібно отримати, саме завдяки такій гнучкості робота змогла бути базованою також і на часових рядах.

Існує додаткова варіативність в моделюванні за допомогою створення почергово семантично зв'язаних елементів, в даній роботі таким чином було досліджена юніграма – N-грама з одним зв'язаним рядком всередині, задля забезпечення цілісного прогнозу в 31 день, при інших значеннях N модель не могла видати ланцюг значень довжиною в 31 значень, та було використано простий генеральний тип токенізації всіх елементів.

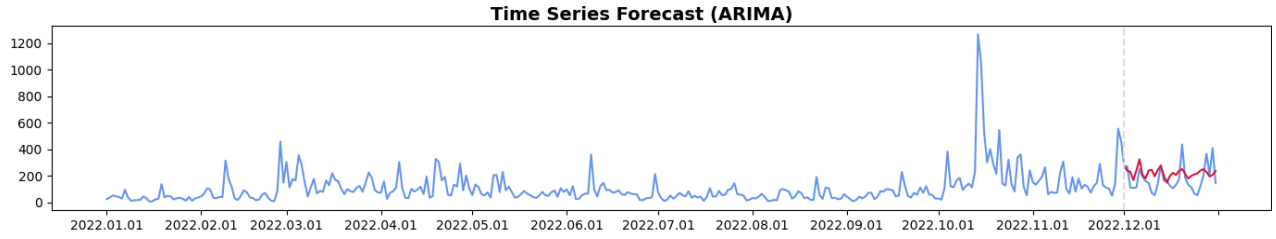
Результати

Згідно кожного методу було розроблено програмне забезпечення, та підлаштовано часовий ряд для подальшого отримання прогнозованих результатів. Було змодельовано графіки, зображені на Рисунку 1, актуальних (реальних) та прогнозованих значень відповідно до часового ряду, отримані результати було повторено задля отримання найбільш релевантних прогнозованих рядів.

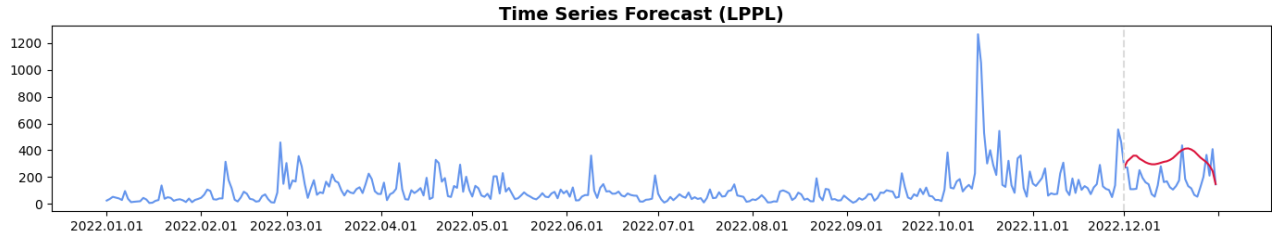
Найбільш результативними для прогнозів показали себе нейронно-мережевий та статистичний підходи, в свою чергу економетричний та лінгвістичний методи показали себе доволі лімітованим для використання в прогнозуванні.



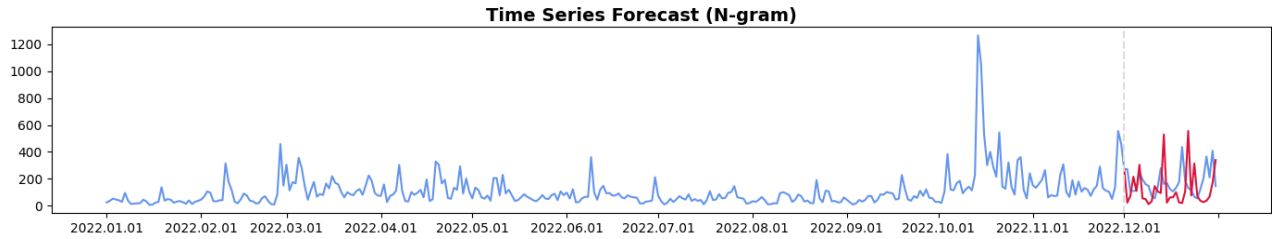
(а) LSTM



(б) ARIMA



(в) LPPL



(г) N-gram

Рис. 1. Ряди показують згадуваність систем старлінк по пошуковим запитам, відрізок за останній місяць – сірий колір, дійсний часовий ряд – блакитний колір, прогнозований часовий ряд – червоний колір

LSTM

Метод LSTM доволі варіативний, який легко можна підігнати під специфіку часового ряду, через його комплексність метод працює стабільно, безвідмовно, отримані прогнозовані результати доволі близькі до реальних (рис. 1(а)). Можливе також налаштування додаткових параметрів [6], так можливо створити багатошарову модель з сильнішим відсіюванням, яка буде прогнозувати більш точні результати.

ARIMA

Хорошим вибором стала ARIMA, вона менш гнучка в використанні, при правильному підборі параметрів p, d, q доволі точно будує свої прогнози відповідно до різного типу часових рядів, серед вибраних варіантів вона показала себе найкраще (рис. 1(б)).

LPPL

Виявилось, що LPPL показує себе доволі погано як метод для прогнозування саме часових рядів, що не дивно через його вузьку напрямленість на вирішення інших завдань (рис. 1(в)). Модель до сих пір еволюціонує з часом, частково у відповідь на обґрунтовану критику, так саме в процесі дослідження було виявлено, що стратегія еволюції адаптації коваріаційної матриці CMA-ES є хорошим вдосконаленням який дозволяє отримувати точніші результати, але попри це при використанні таких генеративних алгоритмів як CMA-ES для покращення прогнозу пропорційно виростає складність самого розрахунку. Виявилось, що проблематичним є й розрахунок окремих великих числових значень, через що доводиться брати їх логарифмічне відображення, що також може впливати на спотворення прогнозу.

N-gram

Модель N-gram представила доволі лімітований варіант прогнозування часових рядів через обмеженість кількістю попередніх можливих значень, згідно яких прогноз може їх приймати (рис. 1(г)). Тобто, розглядаючи цей метод в рамках нестаціонарних рядів прогноз лімітований пороговими значеннями часового ряду та не може виходити за його межі, що зменшує його точність. Тому саме при дослідженні часових рядів можна відмітити широкий простір для вдосконалення при використанні моделі з алгоритмом N-1, при якому спотворення прогнозу на коротких проміжках буде набагато меншим й поступово буде градуюватись відносно часу, та додаванням рекурентної складової який може збільшити точність вже й на більших часових проміжках. Потрібно зауважити й створення спільних або роздільних словників для різних рядів, при відповідній семантичній кореляції для спільних рядів це буде додавати точності, але при відсутності таких кореляцій – навпаки.

Для визначення точності моделей було вирахувано їх середньоквадратичну похибку (MSE), та корінь з середньоквадратичної похибки (RMSE).

Таблиця 1. Розрахунки похибок моделей

	LSTM	ARIMA	LPPL	N-gram
MSE	21009.85	10242.77	36911.94	33618.03
RMSE	144.9477	101.2065	192.1248	183.3522

З результатами можна ознайомитись в Таблиці 1, де нижче за значенням число відображає більшу точність прогнозу.

Висновки

Виходячи з практичної частини, можна зауважити, що кожен з розглянутих методів задовольняє поставлене завдання, не дивлячись на невелику точність до ефективного прогнозування часового ряду таких моделей як LPPL та N-gram, вони дають куди більший творчий простір для подальшого їх вивчення та оптимізування. В свою чергу моделі LSTM та ARIMA показали себе доволі результативно, не дивно, що саме ці моделі та їх підходи є пануючими в розрізі прогнозування часових рядів.

Завдяки проведеній роботі, надалі з'являється ґрунт для подальшого вивчення теми, прогнозування різних типів подій отриманих з відкритих джерел, та зокрема самих моделей. В площині же проробленого дослідження володіючи засобами автоматизованого колекціонування OSINT даних, можна затвердити ефективність їх застосовування для побудови прогнозованих варіантів майбутнього.

Перелік використаних джерел

1. *Sudriani Y.*, *Ridwansyah I.*, *Rustini H. A.* Long short term memory (LSTM) recurrent neural network (RNN) for discharge level prediction and forecast in Cimandiri river, Indonesia. — 2019. — DOI: 10.1088/1755-1315/299/1/012037.
2. *Brownlee J.* Introduction to Time Series Forecasting with Python. — 1st ed. — 2020. — 365 p.
3. *Fantazzini D.*, *Geraskin P.* Everything You Always Wanted to Know about Log Periodic Power Laws for Bubble Modelling but Were Afraid to Ask // European Journal of Finance. — 2011. — Jan. — Vol. 19. — P. 11–13. — DOI: 10.1080/1351847X.2011.601657.
4. *Shu M.*, *Zhu W.* Diagnosis and Prediction of the 2015 Chinese Stock Market Bubble. — 2019. — arXiv: 1905.09633 [q-fin.ST].
5. *Jurafsky D.*, *Martin J. H.* Speech and Language Processing. — 3rd ed. — 2023. — 636 p.
6. *Staudemeyer R. C.*, *Morris E. R.* Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks. — 2019. — arXiv: 1909.09586 [cs.NE].