УДК 004.04

ІНФОРМАЦІЙНА СИСТЕМА ДЛЯ АНАЛІЗУ ПОВІДОМЛЕНЬ ВПЛИВОВИХ ЛЮДЕЙ В СОЦІАЛЬНІЙ МЕРЕЖІ TWITTER

Сарафанов М. Б.
ст. викладач каф. ІС Манікаєва О. С.
Одеський національний політехнічний університет, УКРАЇНА

АНОТАЦІЯ. **На основі аналізу повідомлень впливових політиків, діячів, лідерів в соціальній мережі Twitter запропонована інформаційна система для прогнозування середньотижневої ціни бензину та максимальної денної вартості акцій компанії General Motors.**

Вступ. В останній час все більше посилюється вплив публічних повідомлень, висловлювань відомих політиків, діячів, лідерів в соціальних мережах таких як Twitter, Facebook та ін. Twitter [1] представляє собою не тільки соціальну мережу (Social Media, SM) але є засобом масової інформації (ЗМІ). Тому дослідивши повідомлення впливових людей та світові макротенденції можна спостерігати їх залежність. В роботі запропоновано провести аналіз повідомлень впливового бізнесмена, мера Нью-Йорка (2002-2013) та 44-го президента США Барака Обами. Вибір обраних діячів обумовлений найбільшою кількістю читачів в SM Twitter.

Мета роботи. Метою роботи є прогнозування середньотижневої ціни бензину та максимальної денної вартості акцій компанії General Motors на основі розробленої інформаційної системи аналізу повідомлень Майкла Блумберга та Барака Обами в SM Twitter.

Основна частина роботи. Робота інформаційної системи для аналізу повідомлень в SM Twitter складається із наступних етапів:

1. Попередня обробка даних. Спочатку відбувається завантаження даних із SM Twitter та видалення усіх ретвітів. Ретвіт – це функція SM Twitter, яка дозволяє поділитися зі своїми читачами чужою публікацією. Наступним кроком є видалення керуючих символів (перенос рядка, амперсанд тощо) з тексту публікацій.

2. Виділення ознак. Для виділення ознак було обрано стандартний підхід – робота з табличним відображенням даних за допомогою Python бібліотеки Pandas [2], яка дозволяє працювати як з таблицями (об’єкт DataFrame) так і з колонками/строками (об’єкт Series), підтримуючи всі стандартні операції: сортування, фільтрацію, агрегацію, операцію join та інші. Окрім цього була використана бібліотека Spacy, яка має готові моделі для використання. Завдяки функціоналу цієї бібліотеки було розроблено алгоритм, який дозволяє виділити змістові токени з тексту публікацій. Після цього для кожної публікації нараховувалася кількість надходжень тих чи інших змістових токенів.

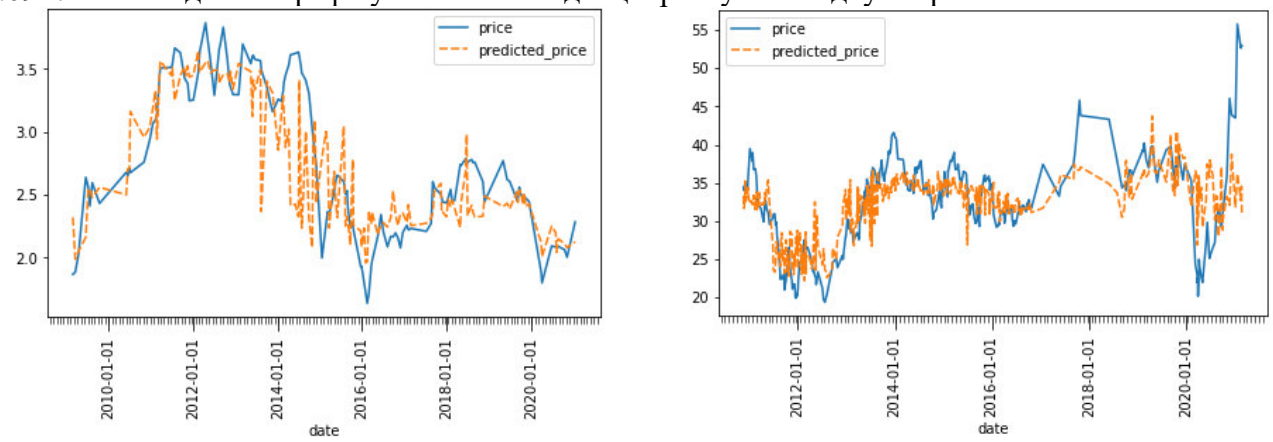
В результаті виконання попередніх етапів отримаємо датасет з ознаками, який необхідно зіставити з датасетом цільової змінної, після чого можна перейти до безпосереднього моделювання, обрання алгоритму, виділення найважливіших ознак та аналізу результату роботи моделі.

На основі попереднього аналізу бібліотек для вирішення задач машинного навчання зі вчителем була обрана бібліотека Scikit-learn, яка надає широкий вибір алгоритмів навчання з учителем і без вчителя. Якість побудованої моделі перевіряється за метрикою MSE (середньо квадратична помилка).

Для прогнозування середньотижневої ціни бензину на основі повідомлень Майкла Блумберга був побудований датасет, який складається із 18365 ознак. Для вибору найбільш значимих ознак був проведений регресійний аналіз, на основі бібліотеки Scikit-learn. В результаті чого, було отримано 400 найбільш значимих ознак, які пов’язані зі світовими новинами/проблемами, а саме пандемією COVID-19: “PANDEMIC”, “COVID-19”, “VIRUS”, та ознаки пов’язані з політикою та нещодавніми виборами президента у США: “TRUMP”, “IMIGRATION REFORM” та актуальними темами у світі: “CLIMATE CHANGE”.

Додаткові ознаки для навчання моделі були отримані із публікацій SM Twitter, що характеризуються наступними метриками: кількість ретвитів, відповідей, лайків та цитувань. Ці метрики дозволяють отримати інформацію про розповсюдженість повідомлень та ступінь реакції суспільства на ці чи інші повідомлення. Важливо зазначити, що прогнозовані змінні мають усереднене значення (на день, на тиждень тощо). Проте, повідомлення у соціальних мережах можуть робитися декілька раз на день. Тому під час формування датасету для навчання моделі проводилась агрегація значень стосовно до прогнозованій змінній. Для виділених токенів з тексту повідомлень була обрана функція суми, а для метрик – середнє значення. Тобто, якщо фінальна змінна характеризується максимальним значенням на день, то для ознак з публікацій була зроблена агрегація теж на день. Таким чином, рядки датасету мають наступний сенс: кількість згадувань змістових токенів на день та середня їх розповсюдженість у вигляді кількості лайків, цитувань тощо. Отримавши датасети з даними публікацій та значенням фінальної змінної, було побудовано датасет для моделювання через виконання операції INNER JOIN. На навчальну вибірку припадає 75% вихідного датасета, а на тестову – 25%

Для прогнозування середньотижневої ціни бензину на основі повідомлень Майкла Блумберга, була обрана модель «випадковий ліс» - RandomForestRegressor. На рис. 1, *а* показано графік прогнозування ціни бензину, де синій графік – реальна середньотижнева ціна бензина, жовта – прогнозована ціна. На навчальній вибірці MSE складає 0.0139, на тестовій вибірці – 0.0947 та як видно на графіку загальна тенденція росту чи спадку зберігається.



а б

Рис. 1. – Результати аналізу повідомлень: а) прогнозування середньотижневої ціни бензину; б) прогнозування максимальної денної вартості акцій компанії General Motors

На основі аналізу повідомлень Барака Обами в MS Twitter прогнозується максимальна денна вартість акцій компанії акцій компанії General Motors, одного з найбільших роботодавців у США. Етап обрання ознак включає виділення тільки тих ознак. які згадувалися хоча б один раз у три-річний період. Завдяки цьому були обрані тільки ті ознаки, які розподілені по часовому протягу датасету. Після цього використовувалась функція SelectFromModel бібліотеки Scikit-learn. Ця функція дозволяє побудувати модель за обраним алгоритмом, проте залишити опорними тільки ті ознаки, вага яких більше ніж середня вага ознак (параметр threshold=’median’). Таким чином, вибірка складалась з 454 найважливіших ознак. На рис. 1, *б* наведено графік реальної та прогнозованої ціни. На навчальній вибірці MSE складає 2.6123, на тестовій вибірці – 17.7904

Висновок. В роботі розроблено інформаційну систему, яка дозволяє прогнозувати середньотижневу ціну бензину та максимальну денну вартість акцій компанії General Motors на основі повідомлень Майкла Блумберга та Барака Обами в SM Twitter.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Social Media Twitter [Електронний ресурс]. – Режим доступу: URL: https://twitter.com/
2. Pandas [Електронний ресурс] / Офіційна документація бібліотеки. – Режим доступу: URL: https://pandas.pydata.org/