УДК 004.852

Р. Н. Квєтний
Д. А. Ткачик

АЛГОРИТМ ДЛЯ АНАЛІЗУ ЕМОЦІЙНОГО ЗАБАРВЛЕННЯ ТЕКСТУ В ПРОГНОЗУВАННІ ДАНИХ НА ФІНАНСОВИХ РИНКАХ

[1] Вінницький національний технічний університет;

Анотація
Описано алгоритм для аналізу емоційного забарвлення тексту в прогнозуванні даних на фінансових ринках, обґрунтовано актуальність та використання даного алгоритму в інтернет-трейдингу.
Ключові слова: інтернет-трейдинг, фінансові ринки, біржа, прогнозування даних, нейронні мережі.

Abstract
An algorithm for sentiment analysis of the text in forecasting data in financial markets is described, the relevance and using of this algorithm in online trading is substantiated.
Keywords: internet trading, financial markets, stock exchange, data forecasting, neural networks.

Вступ

Інтернет-трейдинг в сучасних реаліях набуває стрімкого розвитку, в першу чергу через швидкий доступ до даних завдяки мережі інтернет. Саме тому і з'являються нові методи та підходи до прогнозування даних, які з кожним разом дають більшу точність та ефективність.

Аналіз емоційного забарвлення тексту має безліч застосувань: наприклад такі дані дозволяють передбачити поведінку біржових трейдерів щодо конкретної компанії з відгуків в соціальних мережах чи при аналізі відповідних новин.

Попередня обробка даних

Будь-який робочий процес аналізу даних починається з їх завантаження. Далі ми повинні пропустити їх через конвеєр (pipeline) попередньої обробки:

- токенізувати текст - розбити текст на речення, слова і інші об'єкти;
- видалити стоп-слова;
- привести слова до нормальної форми;
- векторизувати тексти - зробити числові представлення текстів для їх подальшої обробки [1].

Всі ці кроки служать для зменшення шуму, властивого будь-якому звичайному тексту, і підвищення точності результатів класифікатора.

Токенізація - це процес розбиття тексту на більш дрібні частини. Причому вихідними токенами можуть бути як слова, так і знаки пунктуації.

Стоп-слова - це слова, які можуть мати важливе значення в людському спілкуванні, але не мають сенсу для машин, наприклад прийменники, сполучники.

Приведення до нормальної форми використовується лематизація, яка прагне вирішити зазначену проблему, використовуючи структуру даних, в якій всі форми слова зв'язуються з його найпростішою формою - лемою.

Векторизація - перетворення токена в числовий масив, який представляє його властивості. У контексті завдання вектор унікальний для кожного токена. Векторні представлення токенів використовуються для оцінки подібності слів, класифікації текстів і т. д. Вихідні дані представляються у вигляді щільних масивів, в яких для кожної позиції визначені ненульові значень. Це відрізняє використовуваний підхід від ранніх методів, в яких для тих же цілей застосовувалися розріджені масиви і більшість позицій були заповнені нулями [2].

Аналіз емоційного забарвлення

Тепер текст перетворений в форму, зрозумілу комп'ютеру, так що можна почати роботу над його класифікацією.

Для початку важливо зрозуміти загальний робочий процес будь-якого виду завдань класифікації:

1. поділяємо дані на навчальну і тестову вибірки (набори даних);
2. вибираємо архітектуру моделі;
3. використовуємо навчальні дані для налаштування параметрів моделі (цей процес і називається навчанням);
4. використовуємо тестові дані, щоб оцінити якість навчання моделі;
5. використовуємо навчену модель на нових, раніше не аналізованих вхідних даних для створення прогнозів.

Фахівці по машинному навчання зазвичай поділяють набір даних на три складових:

1. дані для навчання (training);
2. дані для валідації (validation);
3. дані для тесту (test).

Так як для кінцевого аналізу потрібно використовувати нейронну мережу, то вхідними даними для неї мають бути текстові дані з визначеними для них категоріями. Саме тому одним з компонентів конвеєра є спеціальний текстовий категоризатор. Щоб за допомогою цього інструменту навчити модель, необхідно виконати наступні дії:

- додати в категоризатор валідні мітки (імена категорій);
- завантажити, перемішати і розділити на частини дані, на яких проходить навчання;
- навчити модель, оцінюючи кожну ітерацію навчання;
- використовувати навчену модель, щоб передбачити тональність настроїв в текстах, що не входили в навчальну вибірку;
- зберегти навчену модель.

Конвеєр дозволяє створити і навчити згорткову нейронну мережу для класифікації текстових даних. Тут цей підхід використовується для сентимент-аналізу, але ніщо не заважає поширити його і на інші завдання класифікації текстів.

Оскільки модель для кожної мітки повертає оцінку від 0 до 1, ми визначаємо позитивний або негативний результат на основі цієї оцінки. На основі статистичних даних ми обчислюємо дві метрики: точність і повноту. Ці метрики є показниками ефективності моделі класифікації:

- точність (P) - відношення істинно позитивних результатів до всіх елементів, зазначеним моделлю як позитивні (справжні і помилкові спрацьовування). Точність 1.0 означає, що кожен відгук, зазначений нашою моделлю як позитивний, дійсно відноситься до позитивного класу;
-

$$P = \frac{TP}{TP+FP},$$

де TP – це істинно позитивні результати, FP – хибно позитивні.

- повнота (R) - це відношення істинно позитивних відгуків до всіх фактичним позитивним відгуками, тобто кількість істинно позитивних відгуків, виділених на сумарну кількість істинно позитивних і помилково негативні відгуків.

$$P = \frac{TP}{TP+FN},$$

де FN – хибно негативні результати.

По мірі навчання моделі будуть змінюватися показники втрат таточності для кожної ітерації навчання. Значення функції втрат стрімко зменшується з кожною ітерацією. Інші параметри також повинні змінюватися, але не так значно: зазвичай вони ростуть на найперших ітераціях, а після цього тримаються приблизно на одному рівні. Після навчання такої нейронної мережі можна використовувати її на реальних даних [3].

Висновки

В роботі було описано загальний алгоритм для аналізу емоційного забарвлення тексту, завдяки якому можна обробляти різні відгуки чи рецензії, що стосуються конкретних компаній на фінансових

ринках для прогнозування даних. Даний підхід можна комбінувати з технічним аналізом для зменшення похибки при прогнозуванні ринку.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc. — P.1

2. Іванов О. В. Класичний контент-аналіз та аналіз тексту: термінологічні та методологічні відмінності / Іванов Олег Валерійович // Вісник Харківського національного університету імені В. Н. Каразіна, Харків: Видавничий центр ХНУ імені В. Н. Каразіна, 2013. — № 1045. — С.72

3. Диковицкий В. В., Шишаев М. Г. Обработка текстов естественного языка в моделях поисковых систем // Сборник научных трудов. — 2010. — С.30

Денис Анатолійович Ткачик – аспірант кафедри АІІТ, факультет комп'ютерних систем і автоматики, Вінницький національний технічний університет, м. Вінниця, e-mail: true.tkachyk@gmail.com

Науковий керівник: ***Квєтний Роман Наумович*** – д-р. техн. наук, професор, завідувач кафедри АІІТ, Вінницький національний технічний університет, м. Вінниця.

Denys A. Tkachyk – AIIT graduate student, Department of Computer Systems and Automation, Vinnytsia national technical University, Vinnytsia, e-mail: true.tkachyk@gmail.com

Supervisor: ***Kvyetnyy Roman N.*** – Dr. Sc. (Eng.), Professor, Head of the Chair of Automation and Intelligent Information Technology, Vinnytsia National Technical University, Vinnytsia.