УДК 004.9

Методи автоматичного аналізу настроїв в соціальних мережах

Шингалов Д.В., аспірант; Тріщ О.В., аспірантка;
Минайленко Р.М., к.т.н., доцент
Центральноукраїнський національний технічний університет, м. Кропивницький

Аналіз настроїв користувачів віртуальних соціальних мереж, таких як Twitter або Facebook, відноситься до класу методів, заснованих на обчислювальній обробці, що використовується для ідентифікації, вилучення та характеристики суб'єктивної інформації, наприклад, думок, виражених в тій чи іншій частині тексту. Основна мета аналізу настроїв класифікувати ставлення автора до різних тем в позитивні, негативні або нейтральні категорії. Аналіз настроїв має безліч застосувань в різних галузях, включаючи бізнес-аналітику, політику, соціологію і т.д.

Основні два методи автоматичного аналізу настроїв – це метод на основі використання лексем (неконтрольований підхід) і метод, машинного навчання (контрольований підхід), які використовують спеціалізовані словники. При машинному навчанні застосовуються класифікатори на базі юніграмм або їх комбінацій (N-грам) в якості ознак. У лексемному методі основі лежать юніграмми, які знаходяться в словнику і мають відповідні бали поляності.

Перед застосуванням будь-якого з методів вилучення настрою, звичайною є практика попередньої обробки даних. Попередньо оброблені дані дозволяють забезпечити високу якість класифікації тексту і зменшити обчислювальну складність. Типова процедура попередньої обробки включає в себе наступні кроки:

- Розмітка за частинами мови. Цей процес дозволяє автоматично визначити кожне слово речення як частину мови.
- Зведення до кореня. Процедура відсікання суфіксів та закінчень від кореня. Розмірність слів зменшується, коли корінь схожих слів відображаються як одне слово.
- Видалення некорисних слів. Це слова, які несуть в собі сполучну функцію в реченнях, наприклад, прийменники, артиклі і т.д.
- Обробка заперечень. Заперечення відноситься до процесу перетворення настроїв тексту з позитивного на негативний або з негативного на позитивний, використовуючи спеціальні слова: "ні", "не" і т.д.
- Умовні оператори. Фрази на кшталт "але", "за винятком", "за винятком того, що", "за винятком для" взагалі можуть кардинально змінити полярність частини тексту, що іде слідом за ними.

- Токенізація в N-грами. Токенізація - це процес створення словнику зі слів тексту.

Лексемно-орієнтований підхід обчислює настрій заданого тексту в залежності від полярності слів або фрази в цьому тексті.

Методика розрахунку настрою [1] полягає у наступному: після попередньої обробки тексту, відбувається перевірка маркера кожного слова на його полярність в лексиконі. Якщо слово не знайдено у лексиконі, тоді його полярність вважається нульовою. Після призначення балів полярності W всім словам, що містяться у тексті, остаточна оцінка S настрою тексту розраховується діленням суми балів слів, які задають настрій тексту (крім нульових) на кількість m таких слів:

$$S = \frac{1}{m}\sum_{i=1}^{m} Wi$$

Усереднення балу дозволяє отримати значення балу настрою у діапазоні від -1 до 1, де 1 означає сильне позитивне почуття, -1 означає сильний негативний настрій і 0 означає, що текст є нейтральним. Якість класифікації багато в чому залежить від якості словника.

Словники можуть бути створені з використанням різних методів:

- Вручну побудовані словники [2] (простий, повільний метод);
- Словники з підготовлених даних.

Метод машинного навчання для аналізу текстів - це контрольований алгоритм, який аналізує дані, які раніше були помічені як позитивні, негативні або нейтральні.

У спрощеному вигляді, задача класифікації текстів може бути описана наступним чином – задано набір маркованих даних:

$$T_{data} = \{(t_1, L_1), \dots (T, n)\},$$

де кожен текст належить до набору даних T і мітка L_i є попередньо встановленим класом всередині групи класів L, мета полягає в тому, щоб побудувати алгоритм навчання, який буде приймати в якості вхідних даних навчальний набір T_{data} і створити модель, яка буде точно класифікувати немарковані тексти.

Найпопулярнішими алгоритмами навчання для класифікації текстів є метод опорних векторів [3], наївний класифікатор Бейєса [4], дерева прийняття рішень [5], метод максимальної ентропії.

Машинне навчання для класифікації текстів починається з аналізу навчальних даних за допомогою алгоритму класифікації. Тут атрибутом маркеру є клас настрою або класифікатор, представлений у вигляді правил класифікації. Тестові дані використовуються для оцінки точності правил класифікації. Якщо точність вважається прийнятною, то правила

можуть застосовуватися до класифікації нових кортежів даних. Точністю класифікатора для даного тестового набору є відсоток тестових наборів кортежів, які правильно класифіковані класифікатором, тому що клас – мітка кожного навчального кортежу забезпечує крок також відомий як «навчання з учителем». Настрій кожного слова з документа визначається за допомогою функції агрегації, загальний же настрій документа визначається різними алгоритмами.

Для цього найчастіше використовується наївний класифікатор Бейєса. Цей класифікатор передбачає, що вплив значення атрибута на даному класі не залежить від значень інших атрибутів. Це припущення називається класом умовної незалежності. Для підвищення якості класифікації застосовується метод максимальної ентропії. На відміну від наївного класифікатора Бейєса, він не припускає, що ознаки умовно незалежні одна від одної. Цей класифікатор засновано на принципі максимальної ентропії усіх моделей, які відповідають даним навчання.

Будь-який з методів автоматичної класифікації тексту не може дати беззаперечних результатів. Поліпшити результати автоматичного визначення тональності тексту можливо за допомогою використання декількох систем класифікації, застосуванням гібридних методів класифікації. Також важливу роль відіграє використання методів автоматичного виправлення орфографічних помилок, вдосконалення словників (для методів, заснованих на словниках) і навчальної вибірки (для методів машинного навчання).

Список літератури

1. Stone, P. J. and Hunt, E. B. (1963). A computer approach to content analysis: Studies using the general inquirer system. In Proceedings of the May 21-23, 1963, Spring Joint Computer Conference, AFIPS '63 (Spring), pages 241–256, New York, NY,USA. ACM.

2. Wiebe, J. (2000). Learning subjective adjectives from corpora. In Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, pages 735–740. AAAI Press.

3. Cortes, C. and Vapnik, V. (1995). Support-vector networks. In Machine Learning, volume 20, pages 273–297, Hingham, MA, USA. Kluwer Academic Publishers.

4. Narayanan, V., Arora, I., and Bhatia, A. (2013). Fast and accurate sentiment classification using an enhanced naive bayes model. In Yin, H., Tang, K., Gao, Y., Klawonn, F., Lee, M., Weise, T., Li, B., and Yao, X., editors, Intelligent Data Engineering and Automated Learning IDEAL 2013, volume 8206 of Lecture Notes in Computer Science, pages 194–201. Springer Berlin Heidelberg.

5. Mitchell, T. M. (1996). Machine Learning. McGrwa Hill, New York, New York, NY, USA.