РОЗРОБКА WEB-ДОДАТКА ДЛЯ АНАЛІЗУ ТОНАЛЬНОСТІ ТЕКСТОВОЇ ІНФОРМАЦІЇ

Іванов Є.М., Коваленко С.В.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

Метою даної роботи є розробка системи, яка буде використовувати бінарну оцінку «позитивна» тональність або «негативна» тональність для дослідження алгоритму аналізу тональності змісту природно-мовних текстів з соціальних мереж [1].

У мережі Інтернет міститься величезна кількість різноманітних текстів. Це можуть бути статті у блогах, відгуки на продукти, повідомлення в соціальних мережах та інше. У даному контенті міститься велика кількість інформації.

Для досягнення поставленої мети було вирішено наступні завдання:

- розглянуто існуючі методи аналізу тональності тексту;
- проаналізовано існуючі алгоритми аналізу тональності тексту;
- визначено вимоги до розробки алгоритмічного забезпечення інтелектуального модуля аналізу емоційного змісту;
- розроблено модуль для аналізу тонального змісту природно-мовних повідомлень соціальних мереж Facebook та Twitter.

У розробленій системі, як один з алгоритмів обробки даних, використовується наївний байєсівський класифікатор [2], який виглядає наступним чином:

$$C = \text{argmax}_{c \ni C} P(C|o_1, o_2, .. o_n) = \text{argmax}_{c \ni C} P(c) \prod_{i=1}^{n} P(o_i|c), \quad (1)$$

де C– набір класів, $o_1, o_2, \dots o_n$– набір ознак. Класифікація зводиться до обчислення максимального значення аргументу при відомому наборі незалежних ознак $o_1, o_2, \dots o_n$.

При цьому: $P(C) \prod_{i=1}^{n} P(o_i|c) = P(C)P(P(o_1|c)P(o_2|c) \dots P(o_n|c))$.

Обчислення ймовірності класу $P(C)$ при відомих ознаках $o_1, o_2, \dots o_n$ зводиться до наступного:

$$P(C|o_1, o_2, \dots o_n) = \sum_{i=1}^{n}(o_1, o_2, \dots o_n) + {}^{1}/_{\sum_{i=1}^{n}(C|A)} + \sum_{i=1}^{n} A, \quad (2)$$

де A – набір відомих ознак, отриманих при навчанні класифікатора

Класифікація тексту при цьому виглядає наступним чином:

$$C(T) = \max \sum_{i=1}^{n}(t_1, t_2 \dots t_n|C), \quad (3)$$

де T – текст, що класифікується, а $t_1, t_2 \dots t_n$– набір речень тексту.

Результатом проведеної роботи є система, яка визначає тональність змісту текстів з соціальних мереж Facebook та Twitter. Система пройшла тестування на різних соціальних профілях. Розроблений додаток може бути корисним у багатьох сферах, наприклад, при управлінні персоналом компанії.

Література:

1. *Котельников Е. В.*Автоматический аналіз тональности текстов на основе методов машинного обучения. Вып. 11 (18), М.: Изд–во РГГУ, 2012. С. 27–36.

2. *Lewis D. D*. Naive (Bayes) atforty: Thein dependence assumptioninin format ionretrieval. Proceedingsof 10th European Conferenceon Machine Learning, Chemnitz, Germany, 1998, pp. 4–15.