УДК 004.89 DOI: https://doi.org/10.30837/IYF.IIS.2024.418

ДОСЛІДЖЕННЯ МЕТОДІВ ОЦІНКИ СЕНТИМЕНТУ ДІАЛОГОВИХ ПОВІДОМЛЕНЬ

Штанько О.
Науковий керівник – д.т.н., проф. Турута О. П.
Харківський національний університет радіоелектроніки, каф. ПІ
м. Харків, Україна
e-mail: oleksii.shtanko@nure.ua

In this theses proposes the task of text sentiment analysis or sentiment analysis involves determining the emotional attitude of the author towards a specific object described in the text. This task is one of the most relevant tasks in natural language processing (NLP). Sentiment analysis is used to evaluate the quality of goods and services based on Internet user reviews, to detect criminally significant content, to determine the authorship of texts, to forecast various economic indicators, and to generate texts with predetermined emotional coloring. The amount of information in electronic form is growing exponentially. Therefore, manual analysis is impossible, leading to the need for automatic methods and tools for analyzing textual information, including methods and tools for automated sentiment analysis.

Огляд останніх досліджень та публікацій. Аналітичний огляд різних джерел показав великий інтерес дослідників до завдання аналізу сентименту [1]. У базовому варіанті це завдання – це завдання класифікації текстів. Результатом завдання є набір текстів, де тексти або елементи поділені на два (позитивний, негативний), три (позитивний, нейтральний, негативний), п'ять (позитивний, досить позитивний, нейтральний, досить негативний, негативний) або більше класів. Існує багато методів, які можна використовувати для вирішення цього завдання. Їх можна розподілити на кілька груп.

Перша група включає методи на основі правил та словників, які використовують попередньо скомпільовані емотивні словники та лінгвістичні правила для пошуку емотивних слів. Першим кроком процесу призначення тексту певному класу є пошук слів у емотивних словниках. Другим кроком є призначення всім знайденим словам їх тональності або ваги зі словника. Потім загальна тональність тексту обчислюється шляхом сумування значень тональності кожного знайденого слова.

Друга група включає методи машинного навчання з учителем, які використовують попередньо навчений класифікатор для визначення тональності нових текстів. Класифікатор навчається на спеціально відібраній колекції текстів з певним типом тональності.

Третя група включає методи машинного навчання без вчителя. У цьому випадку методи визначають тональність термінів, які мають найбільшу вагу. Частота цих термінів повинна бути найбільшою у певному

тексті і в той же час вони повинні бути присутні в невеликій кількості в текстах у всій колекції. Потім тональність всього тексту визначається за допомогою тональності термінів. Комбінація різних методів з різних груп є перспективним шляхом отримання кращого результату. Зазначені методи широко використовуються у відповідному програмному забезпеченні для аналізу сентименту текстів, такому як "Аналітичний Кур'єр" [2], "RCO Fact Extractor SDK" [3], "VAAL" [4], "Eureka Engine", SentiStrength та інші. Вони мають досить хорошу функціональність, але не лишені певних недоліків, особливо щодо аналізу інфлективних мов з багатою морфологією. Тому метою роботи є перевірка ефективності різних методів аналізу сентименту повідомлень у соціальних мережах.

Результати: Для оцінки якості отриманих результатів класифікації використовувалися загальноприйняті метрики: $Recal$, $Precision$, F-міра, Accуратність. Для розрахунку метрик були обчислені значення наступних параметрів:

- TP – кількість правильно позитивних результатів;
- TN – кількість правильно негативних результатів;
- FP – кількість неправильно позитивних результатів;
- FN – кількість неправильно негативних результатів.

$Precision$ – це відношення об'єктів, класифікованих як X, які дійсно належать до класу X:

$$Precision = \frac{TP}{TP + FP}$$

$Recall$ – це відношення всіх об'єктів класу X, класифікованих як належні до класу X:

$$Racall = \frac{TP}{TP + FN}$$

$F\text{-}measure$ – гармонічне середнє між $Precision$ і $Recall$:

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$Accuracy$ – це відношення правильно класифікованих об'єктів до всіх класифікованих об'єктів:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Результати оцінки ефективності класифікаторів після навчання на двох корпусах представлені в Таблиці 1. У цій таблиці NBC означає класифікатор "Наївний Баєсівський", а RNNC – класифікатор, що базується на рекурентній нейронній мережі.

Таблиця 1 – Оцінка ефективності результатів класифікації

	Corpus RuTweetCorp		Slang corpus	
Recall	0.853	0.853	0.948	0.965
Precision	0.875	0.869	0.975	0.982
F-Measure	0.861	0.861	0.961	0.973
Accuracy	0,861	0.855	0.960	0.973

Як показує аналіз результатів оцінки ефективності, після додаткового навчання на другому корпусі жаргону ефективність класифікаторів зросла на 10–11%, що підтверджує раніше запропоновану гіпотезу дослідження.

Список використаних джерел:

1. Ameur H., Jamoussi S., Hamadou A.B.A New Method for Sentiment Analysis Using Contextual Auto-Encoders. Journal of Computer Science and Technology.2018.Volume33, issue 6.P.1307–1319. DOI: https://doi.org/10.1007/s11390-018-1889-1.

2. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. Available at: https://scholar.google.com.ua/citations?view_op=view_citation&hl=de&user=_0Gh01QAAAAJ&citation_for_view=_0Gh01QAAAAJ:p2g8aNsByqUC (accessed 04.06.2022).

3. RCO Fact Extractor SDK. Available at: page_id=3554. (accessed15.09.2019)

4. VAAL project. Available at: http://www.vaal (accessed 05.03.2024).