УДК 004.93 DOI: https://doi.org/10.30837/IYF.IIS.2024.355

ОЦІНКА ЕФЕКТИВНОСТІ МЕТОДІВ АНАЛІЗУ ЕМОЦІЙНОГО ЗАБАРВЛЕННЯ КОМЕНТАРІВ

Фролов М. В.
Науковий керівник: к.т.н., доцент Валенда Н. А.
Харківський національний університет радіоелектроніки, ПІ
м.Харків, Україна
e-mail: maksym.frolov@nure.ua

This thesis is devoted to the study and analysis of various methods of identifying and classifying the emotional coloring of comments in modern Internet resources. The research is aimed at revealing effective approaches to automatic detection of emotional responses in textual content, taking into account the specificity of detecting different shades of emotions, such as positive, negative and neutral. The analysis and comparison of different methodologies and algorithms will allow to determine the optimal approaches for the development of tools for the automated analysis of large volumes of textual information, taking into account its emotional component.

Актуальність та постановка проблеми. В умовах всеосяжної віртуалізації інтеграцій, емоційний вимір текстових комунікацій стає надзвичайно важливим для розуміння та взаємодії в онлайн-середовищах. Кількість коментарів, що публікуються щоденно в Інтернеті, надзвичайно велика, і ефективний аналіз їх емоційного забарвлення має значущий потенціал для покращення якості комунікації, розвитку соціальних мереж, а також для виявлення тенденцій в громадській думці [1]. Здатність автоматично класифікувати емоції в текстових коментарях стає необхідною для ефективного використання цієї інформації в різних сферах, включаючи маркетинг, політику та соціологію. З урахуванням розмаїття виразів та контекстуальних варіацій, що характерні для емоційного висловлення в мовленні, виникає необхідність в розробці та оптимізації алгоритмів, які здатні ефективно виявляти та класифікувати різні емоційні стани в текстових коментарях.

Основні матеріали дослідження. Аналіз тональності тексту – це процес визначення емоційного тону або відчуттів, які виражені в текстовому матеріалі [2]. Цей аналіз спрямований на визначення того, чи текст має позитивний, негативний чи нейтральний характер. Основним завданням в аналізі тональності є класифікація полярності документа, тобто визначення, чи є виражена думка в документі або реченні позитивною, негативною або нейтральною. Методи класифікації тональності в текстах використовуються для визначення емоційного забарвлення висловлювань чи текстового матеріалу. Для оцінки ефективності класифікатора та визначення рівня точності моделі в аналізі емоційного забарвлення часто використовується перехресна перевірка.

Основою цього процесу є тестова вибірка, в якій визначена відповідність між документами та їх класами [3].

Під час перевірки використовується результат, який класифікатор надав для документів у тестовій вибірці, і порівнюється з відомим правильним рішенням. Однак для об'єктивної оцінки ефективності алгоритму потрібна чисельна метрика його якості. У найпростішому випадку такою чисельною метрикою може бути точність (accuracy), яка визначає частку документів, для яких класифікатор прийняв правильне рішення.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}, \quad (1)$$

де Accuracy – точність;
TP – істинно позитивні рішення;
TN – істинно негативні рішення;
FP – помилково позитивні рішення;
FN – помилково негативні рішення.

Однак у цій метриці важливо враховувати особливість: усі документи надаються однаковій вазі, що може бути некоректним, особливо у разі значного усунення розподілу документів у навчальній вибірці на користь одного класу. У такому випадку класифікатор має більше інформації про цей клас і його рішення в межах цього класу може бути адекватнішим. Насправді це може призвести до високої точності, але при цьому класифікатор може показувати слабкі результати в рамках конкретного класу. Для вирішення цієї проблеми рекомендується використовувати збалансований набір даних (датасет).

Точність (precision) та повнота (recall) використовуються як метрика для оцінки алгоритмів вилучення інформації. Іноді вони використовуються самостійно, а іноді служать базою для похідних метрик, таких як F-міра. Система зберігає інформація про те, скільки разів за документами заданого класу прийняте вірне і скільки разів невірне рішення:

$$Precision_p = \frac{TP}{TP + FP}, \quad (2)$$

$$Precision_n = \frac{TN}{TN + FN}, \quad (3)$$

$$Recall_p = \frac{TP}{TP + FN}, \quad (4)$$

$$Recall_n = \frac{TN}{TN + FP}, \quad (5)$$

де $Precision_p$ – точність позитивних рішень;

$Precision_n$ – точність негативних рішень;
$Recall_p$ – повнота позитивних рішень;
$Recall_n$ – повнота негативних рішень.

Точність системи в межах класу визначає, яка частина документів, вірно віднесених системою до даного класу, відноситься до всіх документів, які система визначила як цей клас. Повнота системи вказує на те, яка частина документів, коректно визначених класифікатором як належачих до певного класу, становить відсоток всіх документів цього класу в тестовій вибірці.

Хоча вищі значення точності і повноти є бажаними, в реальних умовах досягнення максимальних значень обох метрик одночасно є складним завданням. Тому необхідно шукати баланс між цими параметрами. В цьому контексті F-міра стає важливою метрикою, яка об'єднує інформацію про точність і повноту алгоритму. F-міра представляє собою гармонійне середнє між точністю і повнотою та тендує до нуля, якщо або точність, або повнота наближається до нуля.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}, \quad (6)$$

де Precision – точність класифікації;
Recall – повнота.

Дана формула надає однакову вагу точності і повноти, тому F-міра буде падати однаково при зменшенні і точності і повноти.

Оскільки метрики використовують різні шкали оцінювання, необхідно провести їхнє стандартизування для подальшого використання та підвищення точності оцінювання. Зважаючи на те, що природна мова вимагає попередньої обробки, необхідно здійснити очищення та векторизацію тексту. Етап обробки коментарів включає такі кроки, як токенізація речень, видалення "стоп-слів", нормалізація слів та їх перетворення в числове представлення для класифікації.

На етапі токенізації речень проводиться розбиття тексту на менші атомарні одиниці, такі як окремі речення або слова. Під час видалення "стоп-слів" очищаються дані від слів, які не несуть семантичного або емоційного навантаження і не мають важливості при класифікації. На етапі нормалізації слів забезпечується однаковий вигляд всіх форм слів за допомогою методу лематизації.

Використовуючи класифікатори машинного навчання, формується резюме, яке містить інформацію про об'єкти висловлювання та відповідну їм тональну лексику. Для ефективної роботи класифікатора та точного визначення тональності тексту, модель потрібно навчати на збалансованих прикладах, які беруться з відкритих джерел.

Висновки. Виявлено, що різні методології та алгоритми мають свої переваги та обмеження у визначенні емоційного тону в коментарях. Важливим аспектом є необхідність розробки гнучких систем, які враховують контекст, семантичні особливості та можливість виявлення нюансів у висловленні емоцій. Дослідження підтверджує, що точні та надійні методи аналізу емоційного висловлення можуть знайти застосування в різних сферах, включаючи моніторинг громадської думки, покращення обслуговування клієнтів, аналіз ринкових тенденцій та управління репутацією.

Список використаних джерел:

1. Задача аналізу тональності тексту Шуляк С.М, Валенда Н.А. Topical issues of the development of modern science // Abstracts of the 9th International scientific and practical conference. Sofia, Bulgaria: ACCENT, 2020. с. 951-956 URL: https://sci-conf.com.ua/ix-mezhdunarodnaya-nauchno-prakticheskaya-konferentsiya-topical-issues-of-the-development-of-modern-science-6-8-maya-2020-goda-sofiya-bolgariya-arhiv.

2. The International Journal of Research on Intelligent Systems for Real Life Complex Problems, TERMS: textual emotion recognition in multidimensional space Yusra Ghafoor, Shi Jinping, Fernando H. Calderon, pages 2673–2693, (2023).

3. International Journal of Computational Intelligence Systems, An Intelligent Hybrid System for Forecasting Stock and Forex Trading Signals using Optimized Recurrent FLANN and Case-Based Reasoning, Luis Martínez Lopez, Jie Lu, ISSN (Print): 1875-6891 (2023).