УДК 004.85

МЕТОДИ ГЛИБИННОГО НАВЧАННЯ ТА ОБРОБКИ ПРИРОДНОМОВНИХ ТЕКСТІВ В ЗАДАЧАХ ПСИХОЛОГІЧНОГО ПРОФІЛЮВАННЯ ЛЮДЕЙ

Шаталов О.В.
Науковий керівник – к.т.н., проф. Рябова Н.В.
Харківський національний університет радіоелектроніки
61166, Харків, просп. Науки, 14, каф. Штучного Інтелекту
e-mail: oleksii.shatalov@nure.ua

This thesis is related to the field of natural language processing (NLP) and psychology, namely personality classification by the Myers-Briggs type indicator (MBTI). The main task that has been resolved during the research is methods for text classification for such subject area comparison and data preprocessing. Several common approaches for text classification and dataset balancing have been used to retrieve and compare the metrics of the models. The main problem was a totally unbalanced dataset that had provoked a decrease in metrics in terms of quality. The results of the research described here show the effectiveness of using deep neural network architectures for the task of text classification in English inside the presented subject area.

Особистість людини – це сукупність відносно стійких патернів поведінки індивіда [1]. Існує декілька моделей класифікації таких патернів. Найпоширенішими з них, які застосовуються в реальному житті, є BigFive та Індикатор типів Майерс-Бріггс (MBTI) [2]. З точки зору обробки природної мови (NLP), такі моделі патернів використовуються для прогнозування рис особистості за вживанням слів, стилем тексту тощо. Такі рішення, що дозволяють передбачати поведінку та спосіб сприйняття світу людиною, можуть бути використані в плануванні маркетингових стратегій, рекомендаційних системах, комунікаційних пропозиціях та в багатьох інших сферах, пов'язаних з електронною комерцією та взаємодією з аудиторією.

Таким чином, завдання полягає в тому, щоб порівняти різні підходи до розпізнавання особистості за текстом, написаним людиною, на основі як класичних підходів машинного навчання, так і моделей глибинного навчання. Подібні дослідження вже проводилися щодо ідентифікації особистості за текстом, написаним людиною [3].

Для проведення подальших дослідницьких операцій необхідно знайти та верифікувати набір даних. Було вирішено обрати в якості джерела набір даних під назвою «MBTI Personality Types 500 Dataset». Він містить дані, зібрані з Reddit та PersonalityCafe. Загалом сама колекція містить близько 106 тисяч записів. Розмір кожного запису обмежений 500 словами. Набір даних має ліцензію CC0, що дозволяє використовувати його з будь-якою метою, включаючи наукові дослідження.

Характеристика MBTI кожного типу складається з 4 описових рис характеру. Кожна з рис має бінарне значення в залежності від того, яка з них має більш виражений прояв у поведінці людини. Таким чином, маємо наведений список цих рис:

– Екстраверти/Інтроверти (E/I);
– Сенсорики/Інтуїти (S/N);
– Мислителі/Почуття (M/П);
– Оцінювачі/Сприймачі (J/P).

Така задача відноситься до задачі класифікації в галузі обробки природної мови. Існує декілька підходів до розв'язання такої задачі, які вже були використані в попередніх роботах на цю тему. Тут будуть розглянуті лише деякі з них:

– на основі BERT;
– логістична регресія;
– лінійна машина опорних векторів (SVM);
– мультиноміальний наївний Байєс;
– дерево рішень;
– випадковий ліс.

Для роботи з текстовими даними, окрім загального балансування набору даних, необхідно виконати деякі стандартні процедури обробки для покращення чутливості моделей до критично важливих параметрів. Так, під час попередньої обробки тексту ми відмовилися від ідеї фільтрації за стоп-словами, віддавши перевагу лише стандартній лематизації та фільтрації непотрібного сміття з тексту (посилання, спецсимволи і т.д.). Найкращою моделлю за показниками wighted F1-score та accuracy стала BERT із значеннями 0.92 та 0.93 відповідно на тестових даних.

Таким чином, був проведений порівняльний аналіз між різними підходами, що дає нам наглядну ілюстрацію працездатності подібного підходу психологічного профлювання людей вцілому, а також демонструє якість використання моделей архітектури Transformer для подібних завдань.

Список використаних джерел:

1. F.H. Allport and G. W. Allport, Personality traits: Their classification and measurement // Journal of Abnormal And Social Psychology, 16, pp. 6–40, 1921.
2. F. Celli, B. Lepri, Is Big Five better than MBTI? A personality computing challenge using Twitter data // Proc. of the Fith Italian Conf. on Computational Linguistics (CLiC-it 2018), Volume 1: Main Conference. CEUR Workshop Proceedings, 2018, 2253.
3. S.S. Keh, and I. Cheng, Myers-Briggs personality classification and personality-specific language generation using pre-trained language models, CoRR, abs/1907.06333, 2019.