УДК 303.62:004.8

МЕТОДИ ОБРОБКИ ДАНИХ АНКЕТУВАННЯ ЗА ДОПОМОГОЮ ШТУЧНОГО ІНТЕЛЕКТУ

Фатій Р.Ю.
email: roman.fatii@nure.ua
Харківський національний університет радіоелектроніки, каф. ЕОМ
м. Харків, Україна

The aim of this paper is to develop an automated questionnaire processing system using AI/ML for fast data classification and filtering. Algorithms (Random Forest, SVM, NLP) are used to identify useful answers and filter out “noise” answers (20-30%). Implemented in Python (Scikit-learn, NLTK), which reduces processing time from 40 hours to 5 hours per month and increases the accuracy of analysis by 50%. Result: An effective tool for universities that automatically sorts questionnaires, saving only relevant data for further analysis.

Сучасні навчальні заклади стикаються зі значною кількістю анкетних даних, зібраних під час опитувань студентів про якість освітнього процесу. Вручну обробка таких даних є трудомісткою, потребує значних часових витрат і часто призводить до суб’єктивних помилок, пов’язаних із людським фактором. Традиційні інструменти, такі як фільтрація та сортування в Excel, не забезпечують автоматизованого аналізу релевантності анкет, що ускладнює виокремлення корисних даних із загального масиву.

Доречність автоматизації зростає в міру зростання обсягу інформації, а швидке прийняття рішень стає вирішальним. Виходячи з доказів у реальному світі, до 30% опитувань може бути частково або зайвим, що помітно впливає на точність висновків [1]. Зайнятість штучного інтелекту (AI) та машинного навчання (ML) полегшує швидку та неупереджену класифікацію даних за допомогою автоматизації обробки опитувальників [2]. Використання сучасних алгоритмів робить це можливим. Автоматизувати «шум» (низька якість, неповні дані або дублювання анкети). Класифікування відповідей за ступенем важливості для подальшого аналізу. Мінімізувати тривалість, присвячено обробці даних та підвищенню продуктивності підзвітного персоналу. Підходи до попередньої обробки текстових відповідей (очищення, токенізація, векторизація). Вибір оптимальних алгоритмів машинного навчання для класифікації анкет.

Створити інструмент на базі ML для автоматичного сортування анкет на:

- корисні (придатні для аналізу ефективності навчання);
- шум (неповні дані, дублікати, спам).

Методологія та результати застосування машинного навчання для класифікації україномовних анкет

У межах даної роботи проаналізовано сучасні підходи до обробки текстової інформації, зокрема системи на основі обробки природної мови (NLP) та алгоритми класифікації, такі як Random Forest, SVM та нейронні мережі.

Для реалізації моделі класифікації було обрано мову програмування Python та відповідні бібліотеки:

- Scikit-learn – для побудови класифікаторів;
- TensorFlow – для реалізації нейромережевих підходів;
- NLTK – для лінгвістичного аналізу текстів;
- Pandas – для обробки даних.

Дослідження виконувалося в середовищі Jupyter Notebook, а для моделювання бізнес-процесів використовувався BPWin. Попередній аналіз даних здійснювався у MS Excel.

Для ефективної класифікації анкет було сформовано набір характеристик, серед яких:

- структурні особливості (довжина відповіді, заповненість полів);
- семантичні ознаки (наявність ключових слів).

Процес навчання моделі навчалася на історичних даних, які були попередньо розмічені експертами за критеріями «корисна» / «некорисна» анкета.

Адаптація методів машинного навчання до специфіки україномовного тексту, враховуючи обмежену підтримку NLP-інструментів для української мови.

Оптимізація процесів обробки анкет шляхом переходу від поточного стану (as-is) до покращеного процесу (to-be) [3].

Розроблений класифікатор продемонстрував точність ≥85% на тестовому наборі даних.

Автоматизоване відсіювання 20–30% нерелевантних анкет, що сприяло підвищенню ефективності роботи аналітичних відділів.

Перспективою в майбутньому є масштабування, наприклад інтеграція з API університетської системи (наприклад, Moodle) для автоматичного оновлення даних. Вдосконалення fine-tuning моделі для мультимодальних даних (текст + числові оцінки). Додавання модуля аналізу тональності (Sentiment Analysis) для виявлення «негативних» анкет. У перспективі передбачається розробка веб-інтерфейсу для працівників, які займатимуться документообігом, з можливістю коригування рішень штучного інтелекту та методів обробки анкетних даних, інтегрованих із технологіями аналізу зображень через алгоритми машинного навчання та комп’ютерного зору, що забезпечить автоматизацію розпізнавання як текстової, так і візуальної інформації [4].

Впровадження AI дозволяє не лише автоматизувати рутинні завдання, але й підвищити якість аналітики за рахунок відсіву нерелевантних даних. Обрана методологія (Python + ML) є гнучкою та масштабованою для

подібних завдань у державних та комерційних установах. Проєкт демонструє значний потенціал для впровадження в інших університетах України.

Список використаних джерел:

1. Zhang Z. Automated Quality Control for Survey Data Using Machine Learning / Z. Zhang, R. Li, Y. Chen // Journal of Survey Statistics and Methodology. – 2021. – Vol. 9, No. 3. – P. 567–592. – DOI: 10.1093/jssam/smab003.
2. Mitchell T. Machine Learning / T. Mitchell. – New York:McGraw-Hill, 1997. – 414 p.
3. Білецький А.О. Data Science в освіті: досвід впровадження AI-інструментів / А.О. Білецький, Т.Г. Шевченко. – Львів : ЛНУ, 2022. – 312 с.
4. Bolohova N. Image processing models and methods research and ways of improving marker recognition technologies in added reality systems / N.Bolohova, I. Ruban // Innovative Technologies and Scientific Solutions for Industries. – 2019. – No. 1 (7). – P. 25–33.