УДК 004.8:004.93

Чайковський П. А., здобувач вищої освіти,
Штовба С. Д., д-р техн. наук, професор,
професор кафедри інформаційних технологій,
Донецький національний університет імені Василя Стуса

АВТОМАТИЧНА ГЕНЕРАЦІЯ НАВЧАЛЬНИХ ДАНИХ ЗА ДОПОМОГОЮ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ДЛЯ ЗАДАЧ КЛАСИФІКАЦІЇ ТЕКСТОВИХ ПОВІДОМЛЕНЬ

Анотація. Запропоновано та досліджено методологію автоматичної генерації навчальних наборів даних для класифікації текстових повідомлень з використанням великих мовних моделей.

Ключові слова: великі мовні моделі, генерація синтетичних даних, класифікація текстів, машинне навчання, навчальні набори даних, GPT-4.

Задача створення репрезентативних навчальних наборів даних є фундаментальною проблемою у розробці систем машинного навчання [1]. Особливої актуальності ця проблема набуває в контексті класифікації текстових повідомлень, де для досягнення високої точності потрібні значні обсяги якісно розмічених даних [2]. Згідно з дослідженнями [3], процес збору та розмітки даних може займати до 80 % часу розробки моделей машинного навчання. Додаткові складнощі виникають через обмеження доступу до реальних даних, пов’язані з питаннями конфіденційності та захисту персональної інформації [4].

Сучасні великі мовні моделі, як-от GPT-4, демонструють значний потенціал для генерації високоякісних синтетичних даних [5]. Під час розробки методу генерації навчальних даних використовується модель GPT-4o, яка демонструє значно покращені можливості генерації тексту, порівняно з попередніми версіями. Розроблений процес (рис. 1) включає формування структурованого запиту з детальним описом категорій та їх характеристик, після чого відбувається ітеративна генерація текстових повідомлень з використанням технік few-shot learning для покращення якості.

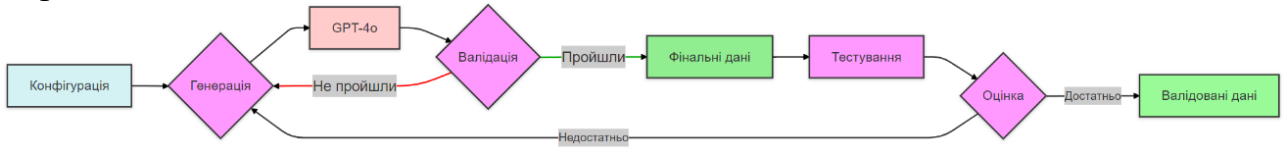

Рисунок 1 – Процес генерації та валідації синтетичних даних

Особлива увага під час формування навчального набору приділяється збалансованості даних між категоріями. Експериментально встановлено, що оптимальним є розподіл приблизно 200 прикладів на кожну категорію важливості. У разі виявлення дисбалансу система автоматично корегує набір даних шляхом видалення надлишкових прикладів або генерації додаткових повідомлень для недопредставлених категорій.

Експериментальна верифікація підходу проводилась на задачі класифікації важливості повідомлень з використанням 5-рівневої шкали категоризації [6].

Навчальний набір включав 1 000 згенерованих повідомлень, а валідаційний – 100 реальних повідомлень, розмічених експертами.

Стабільність запропонованого підходу підтверджується серією експериментів з різними параметрами генерації [7]. Було продемонстровано стійкість результатів класифікації під час проведення серії з п’яти експериментів. Середня точність класифікації становить 92,3 % (±1,2 %), що свідчить про високу надійність розробленого підходу.

Аналіз матриці помилок показав очікувану концентрацію неправильних класифікацій між сусідніми категоріями важливості, що відповідає природі задачі [8]. Це підтверджує, що згенеровані дані зберігають природну структуру та взаємозв’язки між категоріями, характерні для реальних повідомлень.

Висновки. Подальші дослідження спрямовані на розробку метрик оцінки якості згенерованих даних та методів виявлення потенційних упереджень. Особливу увагу планується приділити оптимізації процесу генерації для специфічних предметних областей та дослідженню впливу розміру навчальної вибірки на якість класифікації [9, 10].

Список використаних джерел

1. Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations / Zh. Li et al. URL: https://arxiv.org/abs/2310.07849 (дата звернення: 26.09.2024).

2. The Cost of Down-Scaling Language Models: Fact Recall Deteriorates before In-Context Learning. URL: https://arxiv.org/abs/2310.04680 (дата звернення: 26.09.2024).

3. Mansour Kh. A Survey of Synthetic Data Generation for Machine Learning. URL: https://www.researchgate.net/publication/357907999_A_Survey_of_Synthetic_Data_Generation_for_Machine_Learning (дата звернення: 26.09.2024).

4. Efstathia Soufler, Synthetic Dataset Generation for Privacy-Preserving Machine Learning. URL: https://arxiv.org/abs/2210.03205 (дата звернення: 26.09.2024).

5. Song Y. A Comprehensive Survey of Few-shot Learning: Evolution, Applications, Challenges, and Opportunities. URL: https://dl.acm.org/doi/10.1145/3582688 (дата звернення: 26.09.2024).

6. Pratvina Talele. Classification and Prioritisation of Software Requirements using Machine Learning – A Systematic Review. URL: https://ieeexplore.ieee.org/document/9377190 (дата звернення: 26.09.2024).

7. Chancellor R. Woolsey, Prakash Bisht, Joshua Rothman, Gondy Leroy. Utilizing Large Language Models to Generate Synthetic Data to Increase the Performance of BERT-Based Neural Networks. URL: https://arxiv.org/abs/2405.06695 (дата звернення: 26.09.2024).

8. White M., Rozovskaya A. A Comparative Study of Synthetic Data Generation Methods for Grammatical Error Correction. URL: https://proceedings.mlr.press/v202/anderson23a.html (дата звернення: 26.02.2024).

9. Du Ch., Tian J. Task-Level Thinking Steps Help Large Language Models for Challenging Classification Task. URL: https://aclanthology.org/2023.emnlp-main.150/ (дата звернення: 26.09.2024).

10. Sun X., Li X. Text Classification via Large Language Models. URL: https://aclanthology.org/2023.findings-emnlp.603.pdf (дата звернення: 26.02.2024).