Адаптация больших языковых моделей к задаче классификации рекламных текстов с множественными метками при экстремальном дисбалансе данных

Шестаков Антон Владимирович

Аспирант

Российский университет дружбы народов, Москва, Россия

E-mail: savspit@gmail.com

Научный руководитель *— Виноградов Андрей Николаевич*

В данной работе рассматривается задача классификации коротких рекламных текстов с множественными метками на предмет выявления нарушений законодательства. Эксперименты проводились на корпусе из более чем 50 000 таких текстов, ключевая особенность которого — экстремальный дисбаланс. Преобладающие категории на порядки превосходят по частоте редкие, но значимые классы.

Классические языковые модели архитектуры BERT демонстрируют ограниченную обобщающую способность на малопредставленных классах [3]. Современные большие языковые модели (LLM) в стандартном генеративном режиме (Generative SFT) также показывают недостаточную эффективность на малопредставленных категориях. Это обусловлено тем, что в процессе генерации текста оптимизация функции потерь ориентирована на наиболее частотные паттерны обучающей выборки, что приводит к доминированию высокочастотных классов при генерации меток [2].

В рамках исследования предлагается метод дискриминативной адаптации LLM для работы с данными, характеризующимися экстремальным дисбалансом классов. В качестве базовой архитектуры выбрана модель Qwen-2.5-7B, основанная на архитектуре Transformer [5]. Предложенный подход заключается в переходе от задачи генерации текста к задаче прямой классификации. Слой языкового моделирования заменяется на полносвязный классификационный слой. Для компенсации дисбаланса применяется модифицированная функция потерь с динамическим взвешиванием позитивных примеров (взвешенная бинарная кросс-энтропия):

$$L = -\frac{1}{C}\sum_{c=1}^{C}\left[w_c \cdot y_c \log(p_c) + (1 - y_c)\log(1 - p_c)\right]$$

где C — общее количество классов (категорий нарушений), $y_c \in$

$\{0, 1\}$ — истинная метка для класса c, p_c — предсказанная вероятность c, а w_c —весовой коэффициент, обратно пропорциональный частоте класса в обучающей выборке. Вычислительная сложность обучения снижается за счет использования метода низкоранговой адаптации (LoRA) [1].

Итоговое тестирование всех моделей осуществлялось на отложенной выборке, содержащей исключительно оригинальную экспертную разметку. В качестве базового решения была выбрана модель ruRoberta-large [4], показавшая наилучший результат среди моделей архитектуры BERT.

Результаты экспериментов (таблица 1) демонстрируют, что предложенная модель превосходит бейзлайн как по способности к обобщению редких категорий (Macro F1), так и по общей взвешенной точности (Weighted F1).

Таблица 1: Сравнение агрегированных метрик качества на тестовой выборке

Модель	Метод обучения	Macro F1	Weighted F1
ruRoberta-large (baseline)	Cross-Entropy	0.4629	0.87
Qwen-2.5-7B	Weighted BCE + LoRA	0.4871	0.89

Прирост качества обусловлен повышением точности и полноты на категориях с разной представленностью в обучающей выборке, представляющих сложность для классификации (таблица 2).

Таблица 2: Срез производительности предложенной модели по частотности классов

Категория	Support	Precision	Recall	F1-Score
Акции	323	0.99	0.92	0.95
Возрастной рейтинг	123	0.93	0.83	0.88
Слова на иностр. языке	64	0.95	0.89	0.92
МФО	38	0.94	0.89	0.92
Финансовые услуги	38	0.95	0.53	0.68
Медицинские услуги	30	0.79	0.90	0.84
Букмекеры	8	1.00	1.00	1.00
Консалтинговые услуги	5	0.56	1.00	0.71

Предложенная архитектура позволяет выявлять нарушения, требующие анализа контекста (категория «Возрастной рейтинг», F1 = 0.88). За счет использования весовых коэффициентов в функции потерь было достигнуто существенное повышение полноты обнаружения (Recall) на малопредставленных категориях («Букмекеры», «Консалтинговые услуги») при сохранении приемлемого уровня точности (Precision). Способность модели извлекать редкие примеры без критического роста ложных срабатываний свидетельствует о её устойчивости к дисбалансу данных. При этом взвешенная точность (Weighted F1) предложенной модели на всей выборке составила 0.89, что сопоставимо с результатами базовой модели (Weighted F1 = 0.87). Это подтверждает, что достижение устойчивости на малопредставленных классах не привело к деградации качества классификации на высокочастотных категориях.

Таким образом, дообучение LLM в режиме классификатора со взвешенной функцией потерь позволяет нивелировать проблему экстремального дисбаланса классов и повысить эффективность модерации рекламных текстов.

Литература

1. Hu E. J. et al. Lora: Low-rank adaptation of large language models // ICLR. – 2022. – Т. 1, № 2. – P. 3.
2. Kandpal N. et al. Large language models struggle to learn long-tail knowledge // International conference on machine learning. – PMLR, 2023. – P. 15696-15707.
3. Li X. et al. Dice loss for data-imbalanced NLP tasks // Proceedings of the 58th annual meeting of the association for computational linguistics. – 2020. – P. 465-476.
4. Shavrina T. et al. RussianSuperGLUE: A Russian language understanding evaluation benchmark // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2020. – P. 4717-4726.
5. Vaswani A. et al. Attention is all you need // Advances in neural information processing systems. – 2017. – Т. 30.