РЕЦЕНЗИЯ

На выпускную квалификационную работу студента 4 курса кафедры системного программирования СПбГУ Выродова Михаила Владимировича, направление 09.03.04 «Программная инженерия».

Тема выпускной квалификационной работы: Разработка системы для универсальной классификации текстов

Выпускная квалификационная работа посвящена разработке универсального классификатора текстов на основе машинного обучения для продуктовых задач компании «ВК». В главе 1 автор формулирует цель — создать компактную модель, способную классифицировать любой новый сценарий без дообучения, и создать сервис для использования модели. В главе 2 представлен обзор актуальных методов классификации текстов, а также рассмотрен алгоритм распознавания именованных сущностей GLiNER. В главе 3 предложен собственный подход, переносящий идеи GLiNER на задачу классификации: для текста и описания классов создаются векторы признаков некоторой моделью-кодировщиком, после чего вектор признаков текста сравнивается с представлениями классов; после чего происходит классификация — это упрощает архитектуру и снижает вычислительные затраты. В главе 4 описаны процессы подготовки данных: сегментация корпуса allenai/c4, автоматическая генерация названий классов (пункт 4.5) и четыре варианта разметки с помощью больших языковых моделей (пункт 4.6). Итоговая «смешанная» стратегия дала наибольший прирост качества. Далее автор обучил две собственные модели — GeRaCl-USER-bge-m3 и GeRaCl-USER2-base — и сопоставил их с четырьмя популярными альтернативами (NLI-кодировщики и БЯМ). По точности на пяти задачах RU-MTEB и промышленном корпусе получен выигрыш до 20 п.п. при сопоставимом размере весов. Глава 5 описывает развертывание: распределённую систему на FastAPI, балансировщик нагрузки nginx и сервер вывода, который распределяет запросы между vLLM или собственным конвейером для получения предсказаний.

Сильные стороны работы

- **Чёткая и обоснованная постановка задачи.** Во введении и главе 1 автор напрямую связывает потребность в универсальном классификаторе с изменчивостью пользовательского контента и ресурсными ограничениями при обучении отдельных моделей. Данные ограничения являются одними из самых часто встречающихся в индустрии.

- **Содержательный набор экспериментов.** Автор сопоставил три принципиально разные парадигмы классификации: Модели-кодировщики, обученные на задаче логического вывода (NLI); Большие языковые модели (БЯМ); Собственные компактные модели GeRaCl-USER-bge-m3 и GeRaCl-USER2-base. Эксперименты выполнены на пяти общедоступных датасетах набора RU-MTEB, а также на закрытом производственном корпусе (табл. 2 и 3 главы 5). Проведенный широкий экспериментальный анализ подтверждает способность предложенного метода обобщаться на широкий круг сценариев.
- **Следование промышленным стандартам.** Реализация включает пакеты PyTorch Lightning и WandB, nginx для балансировки, vLLM для эффективного получения предсказаний от LLM, контейнеризацию с Docker-compose (глава 5). Все эти инструменты активно применяются в промышленной разработке. Автор показал, что хорошо разбирается в нюансах использования и работы инструментов.
- **Высокое качество инженерии и открытость.** Все результаты ВКР выложены в открытый доступ на github.com и huggingface.co. Код организован по модулям с подсказками типов, YAML-конфигурациями и хорошо задокументирован.

Недостатки работы

- **Отсутствует сравнительный анализ по времени получения предсказаний.** Таблицы 2 и 3 содержат лишь размеры моделей и показатели качества, тогда как заявленная эффективность решения подразумевает также сравнение скорости.
- **Недостаточно числовых обоснований при составлении обучающей выборки.** В пункте 4.6 описаны эксперименты по разметке, однако не приведены метрики, подтверждающие оптимальность выбора смешанной стратегии.

Диплом демонстрирует высокий уровень как исследовательской, так и инженерной проработки: крупный объём экспериментов, аргументированный выбор архитектур, соблюдение индустриальных стандартов и открытость результатов. Недостатки касаются лишь глубины отчётности по некоторым метрикам и не уменьшают общей ценности проделанной работы.

Рекомендую поставить оценку “А” (отлично).

08.06.2025

Разработчик АО “Точка” А.Д. Земеров

