Создание персонализированных классификаторов контента конечными пользователями: сравнение маркировки примеров, написания правил и LLMPrompting

Дата: 2025-03-01 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2409.03247

Рейтинг: 85
Адаптивность: 90

Ключевые выводы:

Основная цель исследования - сравнить три стратегии создания персонализированных классификаторов контента: маркировку примеров, написание правил и промптинг LLM. Главные результаты показали, что написание промптов обеспечивает лучшую производительность, но пользователи предпочитают разные стратегии в зависимости от контекста, а все стратегии сталкиваются с трудностями при итеративном улучшении.

Объяснение метода:

Исследование предлагает практические рекомендации по выбору оптимальной стратегии взаимодействия с LLM для создания персонализированных классификаторов контента. Оно выявляет, что разные подходы (маркировка примеров, правила, промпты) эффективны в разных контекстах и демонстрирует преимущества гибридных стратегий. Результаты напрямую применимы широкой аудиторией без технических знаний и дают понимание ограничений каждого метода.

Ключевые аспекты исследования: 1. **Сравнение трех стратегий создания персонализированных классификаторов контента**: исследование сравнивает маркировку примеров (labeling examples), написание правил (rule writing) и формулирование промптов для LLM.

Оценка скорости инициализации и итеративного улучшения: анализируется, насколько быстро пользователи могут создать начальный классификатор и как легко его улучшать со временем.

Выявление предпочтений пользователей в разных контекстах: исследование показывает, что пользователи предпочитают разные подходы в зависимости от характера их предпочтений (интуитивные, хорошо определенные общие или конкретные).

Анализ производительности классификаторов: LLM-промпты показали наилучшую общую производительность, особенно по полноте (recall), хотя точность (precision) была сопоставима с системой на основе правил.

Исследование гибридных подходов: пользователи естественным образом комбинировали разные стратегии для улучшения своих классификаторов, например, добавляя примеры в промпты или создавая промпты, похожие на правила.

Дополнение:

Применимость методов исследования в стандартном чате

Методы, исследованные в данной работе, **не требуют дообучения или специального API** для основного применения. Хотя авторы использовали специализированные интерфейсы для проведения эксперимента, ключевые концепции и подходы могут быть адаптированы для использования в стандартном чате с LLM.

Концепции и подходы, применимые в стандартном чате:

Выбор стратегии в зависимости от типа предпочтений: Для интуитивных, но плохо определенных предпочтений: предоставление примеров Для хорошо определенных общих предпочтений: формулирование промптов Для конкретных тем или событий: создание правил (списков ключевых слов)

Гибридные подходы:

Включение примеров в промпты (few-shot learning) Написание промптов в стиле правил с перечислением ключевых слов Итеративное уточнение промптов на основе результатов

Эффективная инициализация:

Начало с простых промптов для быстрого получения базовых результатов Последующее уточнение на основе ошибок #### Ожидаемые результаты от применения:

Повышение эффективности взаимодействия: Пользователи смогут быстрее достигать желаемых результатов, выбирая оптимальную стратегию.

Улучшение качества персонализации: Комбинируя подходы (например, добавляя примеры в промпты), можно достичь лучшего понимания нюансов пользовательских предпочтений.

Снижение когнитивной нагрузки: Выбор правильной стратегии снижает усилия, необходимые для объяснения своих предпочтений LLM.

Более предсказуемые результаты: Понимание ограничений каждого подхода помогает формировать реалистичные ожидания и выбирать стратегии в зависимости от приоритета точности или полноты.

Таким образом, хотя исследование проводилось с использованием специальных интерфейсов, его основные выводы и рекомендации могут быть непосредственно применены в стандартном чате с LLM без необходимости в дообучении или специальном API.

Анализ практической применимости: **Сравнение трех стратегий создания классификаторов**: - Прямая применимость: Высокая. Исследование предлагает конкретные рекомендации по выбору подхода в зависимости от задачи. Пользователи могут выбрать маркировку примеров для интуитивных предпочтений, промпты для общих хорошо определенных критериев и правила для конкретных тем. - Концептуальная ценность: Высокая. Исследование демонстрирует ограничения каждого подхода и помогает понять, когда какой метод наиболее эффективен. - Потенциал для адаптации: Высокий. Выводы легко переносятся на взаимодействие с любыми LLM-системами для персонализированной фильтрации контента.

Оценка скорости инициализации и итеративного улучшения: - Прямая применимость: Средняя. Пользователи могут применять стратегию написания промптов для быстрого создания начального классификатора, но должны учитывать сложности с дальнейшей итерацией. - Концептуальная ценность: Высокая. Понимание, что LLM быстро достигают 95% своей максимальной производительности, но затем сложнее улучшаются, помогает формировать реалистичные ожидания. - Потенциал для адаптации: Высокий. Эти наблюдения применимы к широкому спектру задач с использованием LLM.

Выявление предпочтений пользователей в разных контекстах: - Прямая применимость: Высокая. Пользователи могут сразу применить рекомендации по выбору подхода в зависимости от характера своих предпочтений. - Концептуальная ценность: Высокая. Исследование помогает понять, как тип предпочтений влияет на эффективность разных подходов коммуникации с LLM. - Потенциал для адаптации: Высокий. Эти выводы можно использовать для любых задач персонализации с помощью LLM.

Анализ производительности классификаторов: - Прямая применимость: Средняя. Результаты показывают, что промпты LLM обеспечивают лучшую полноту, но для некоторых задач пользователи могут предпочесть точность, которую лучше обеспечивают правила. - Концептуальная ценность: Высокая. Исследование демонстрирует компромисс между разными метриками производительности. - Потенциал для адаптации: Высокий. Эти знания помогают выбрать подход в зависимости от приоритета точности или полноты.

Исследование гибридных подходов: - Прямая применимость: Высокая.

Пользователи могут сразу применять гибридные стратегии, например, добавлять примеры в промпты или писать промпты, похожие на правила. - Концептуальная ценность: Очень высокая. Понимание, что комбинирование подходов часто даёт лучшие результаты, чем использование одного метода. - Потенциал для адаптации: Очень высокий. Гибридные подходы могут быть адаптированы для различных задач взаимодействия с LLM.

Prompt:

Использование исследования о персонализированных классификаторах в промптах для GPT ## Ключевые выводы исследования для промптинга

Исследование показывает, что: - Промпты для LLM обеспечивают лучшую производительность классификаторов контента - Разные стратегии эффективны для разных типов задач - Включение конкретных примеров (few-shot) улучшает точность - Структурирование промптов по образцу правил повышает эффективность

Пример промпта для классификации контента

[======] # Задача: Классификация комментариев как оскорбительных/неоскорбительных

Контекст и инструкции Я хочу создать персонализированный фильтр контента для выявления оскорбительных комментариев. Исследования показывают, что для хорошо определенных общих предпочтений LLM-промпты наиболее эффективны.

Определение оскорбительного контента Оскорбительным считается контент, который: - Содержит явные ругательства - Включает унижающие достоинство выражения - Содержит угрозы или агрессивные высказывания - Демонстрирует дискриминацию по любому признаку

Few-shot примеры для повышения точности Примеры оскорбительных комментариев: 1. "Ты полный идиот, если думаешь иначе" 2. "Люди из этой страны всегда такие тупые"

Примеры неоскорбительных комментариев: 1. "Я не согласен с этой точкой зрения" 2. "Этот продукт имеет серьезные недостатки"

Структура правил для улучшения точности ЕСЛИ комментарий содержит прямые оскорбления личности ИЛИ комментарий включает дискриминационные высказывания ИЛИ комментарий содержит угрозы ТОГДА классифицировать как "оскорбительный" ИНАЧЕ классифицировать как "неоскорбительный"

Задание Проанализируй следующие комментарии и классифицируй их как оскорбительные или неоскорбительные, объясняя свое решение: [Комментарии для классификации] [======]

Почему это работает

Структура промпта включает элементы, которые исследование определило как эффективные: Четкое определение критериев классификации Few-shot примеры для повышения точности Структурирование по образцу правил для лучшего понимания

Учет контекста задачи - промпт указывает, что мы работаем с хорошо определенными общими предпочтениями, для которых LLM-промпты особенно эффективны.

Итеративное улучшение - промпт можно легко модифицировать, добавляя примеры неправильно классифицированных случаев, что согласно исследованию повышает эффективность.

Такой подход позволяет достичь высокой эффективности классификации уже в первые минуты использования, что соответствует выводам исследования о быстром достижении 95% пиковой производительности.