Методы классификации обращений пациентов для автоматизации поддержки принятия управленческих решений в сфере здравоохранения

Калинин П.С., Волгин Л.А., Седельников П.В, Федоров Д.А. (ИТМО).
Научный руководитель: доцент Фёдоров Д.А. (ИТМО)

Введение. Обработка обращений пациентов является критически важной задачей для системы здравоохранения, поскольку жалобы отражают неудовлетворенность граждан и системные недостатки в оказании медицинской помощи [1]. Процесс ручной обработки крайне трудоемок: в период пандемии COVID-19 объем обращений в отдельных регионах вырос втрое – с менее чем 5 000 в 2019 году до более 18 000 в 2021 году [2]. Ручная классификация, сбор информации и подготовка ответов занимают значительное время и приводят к задержкам, а качество ответов зависит от опыта конкретного специалиста.

Основная часть. В рамках ранее проведенного исследования [1] была спроектирована мультиагентная система поддержки принятия решений (СППР), включающая семь специализированных ИИ-агентов: (1) агент предобработки текста – очистка, деперсонализация и структурирование текста обращения; (2) агент классификации – автоматическое определение тематики обращения; (3) агент поиска в базе знаний – извлечение релевантной информации из базы ретроспективных обращений и нормативных документов с использованием технологии RAG (Retrieval-Augmented Generation) [3, 4]; (4) агент проверки ЕГИСЗ – верификация данных через API Единой государственной информационной системы в сфере здравоохранения; (5) агент-чатбот для взаимодействия с ответственными организациями; (6) генеративный агент – подготовка проекта ответа на основе большой языковой модели с принципами объяснимого ИИ (xAI); (7) агент статистического анализа – мониторинг тенденций и формирование аналитических отчётов.

В текущей работе внимание уделено агенту классификации как ключевому компоненту системы, от которого зависит маршрутизация обращений и качество работы всех последующих агентов. Проведён сравнительный анализ трёх методов классификации на наборе данных из более чем 25 000 обращений по 34 тематическим категориям.

Первый метод – использование больших языковых моделей (LLM) с загрузкой контекста [5]. Были протестированы пять моделей: ChatGPT 5.1, Claude Sonnet 4.5, Qwen3-Max, DeepSeek V3 и Яндекс Алиса. Были сформированы несколько файлов: (1) тематики (категории) обращений (7 штук), (2) неразмеченный (test) набор обращений (220+ штук), (3) размеченный (train) набор обращений (900+ штук), 4) структурированный промпт. Наилучший результат показала модель Claude Sonnet 4.5 – точность 73,45%. Однако практическое применение облачных LLM в государственных медицинских учреждениях ограничено невозможностью передачи персональных данных за пределы защищённого контура региональной инфраструктуры.

Второй метод – многоклассовая классификация на основе дообучения модели RuBERT [2] на ~19 000 однотемных обращений. Достигнута точность 69,5% и точность в топ-3 предсказаниях 91,8%, что критически важно для сценариев с участием человека. Третий метод – многометочная классификация на полном наборе данных (25 248 обращений) с использованием бинарного вектора меток, которая показала F1-micro 64,4%. Снижение на 5 процентных пунктов по сравнению с многоклассовой моделью отражает сложность многотемных обращений.

Анализ ошибок выявил четыре основных фактора, ограничивающих производительность: низкое качество текста (около 30% обращений содержат

орфографические и грамматические ошибки); отсутствие семантической сегментации при наличии нескольких тем в одном обращении; пересечение категорий в плоской таксономии; переобучение при дообучении всех 180 млн параметров RuBERT.

На основании выявленных ограничений предложен интегрированный конвейер классификации: (1) предобработка текста – проверка орфографии, грамматики, сегментация предложений; (2) семантическая сегментация – разделение обращения на тематические блоки на основе BERT-эмбеддингов; (3) иерархическая классификация с замороженными весами RuBERT – на первом уровне определяется тематическая группа (5–7 групп), на втором – конкретная тема (34 категории); (4) агрегация и оценка уверенности. Ожидаемое улучшение составляет 20–35 процентных пунктов, что потенциально позволит достичь точности 85–90% на однотемных обращениях.

Выводы. Предложенная мультиагентная СППР позволяет автоматизировать обработку обращений пациентов, сохраняя необходимый контроль человека для сложных случаев. Проведённый эксперимент подтвердил возможность эффективной классификации обращений как с помощью LLM, так и с помощью дообученных трансформерных моделей. Разработанный интегрированный конвейер классификации с семантической сегментацией и иерархическим методом закладывает основу для существенного повышения качества автоматической маршрутизации обращений. Дальнейшая работа будет направлена на реализацию предложенного конвейера, интеграцию с медицинскими информационными системами и пилотное внедрение в одном из регионов России.

Список использованных источников:

1. Kalinin P., Fedorov D., Sedelnikov P., Volgin L. Design of a Multi-Agent Decision Support System for Handling Patient Complaints // Lecture Notes in Computer Science. – 2026.

2. Reshetnikov A.V., Abaeva O.P., Berdutin V.A. et al. Experience in developing a neural network dialogue system for responding to written requests of the population to a large federal health care institution // Med. Technol. Assess. Choice. – 2025. – Vol. 47(2). – P. 48–57.

3. Amugongo L.M., Mascheroni P., Brooks S. et al. Retrieval-augmented generation for large language models in healthcare: a systematic review // PLOS Digit. Health. – 2025. – Vol. 4(6): e0000877.

4. Liu S., McCoy A.B., Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines // J. Am. Med. Inform. Assoc. (JAMIA). – 2025. – Vol. 32(4). – P. 605–615. DOI: 10.1093/jamia/ocaf008.

5. Koh S.W.C., Wong E.R.N., Tan J.C.M. et al. Classifying Patient Complaints Using Artificial Intelligence–Powered Large Language Models: Cross-Sectional Study // J. Med. Internet Res. (JMIR). – 2025. – Vol. 27: e74231. DOI: 10.2196/74231.