УДК 004.8

Козьма Александра Петровна

Бакалавр

Технологии защиты информации

Университет ИТМО

г. Санкт-Петербург

ИССЛЕДОВАНИЕ ИСПОЛЬЗОВАНИЯ ТЕКСТОВЫХ ЭМБЕДДИНГОВ ДЛЯ РЕКОМЕНДАЦИИ ЛЕКАРСТВЕННЫХ ПРЕПАРАТОВ

Аннотация. В работе представлен прототип системы рекомендации лекарственных препаратов, основанной на анализе текстовых описаний назначения и действия лекарственных препаратов. Модель SentenceTransformer используется для получения эмбеддингов, которые затем сравниваются с помощью индекса ближайших соседей. Система не требует ручной разметки и демонстрирует высокую точность на открытом медицинском наборе данных. Обсуждаются архитектурные решения, возможные применения в медицинских системах и направления дальнейшего развития.

Annotation. This paper presents a prototype of a drug recommendation system based on text descriptions of drug use and effects. SentenceTransformer is used to generate vector representations, which are then compared using a nearest neighbor index. The system operates without supervised learning and demonstrates high accuracy on open medical data. Architectural solutions, potential applications in healthcare systems, and future directions are discussed.

Ключевые слова: рекомендации, эмбеддинги, трансформеры, информационная система, медицина, NLP, машинное обучение, нейросеть, валидация, предобработка текста.

Key words: recommendations, embeddings, transformers, information system, medicine, NLP, machine learning, neural network, validation, text preprocessing.

ВВЕДЕНИЕ

Современная медицина предлагает огромное количество лекарств с разным составом и действием. В потоке информации пациентам и даже врачам порой трудно быстро выбрать подходящее лекарство, когда есть множество вариантов. Особенно это заметно в телемедицине, где нужно дать предварительные советы, имея только краткое описание жалоб, без полного осмотра пациента.

Умные системы поиска, основанные на глубоком обучении и обработке текста (NLP), уже хорошо себя показали в работе с языком. Они помогают искать, кратко излагать и делить на группы медицинские тексты [3]. Использование готовых трансформеров и семантических векторов – один из самых многообещающих способов. Он позволяет представить сложные описания в виде векторов, которые легко сравнивать и искать по смыслу.

В данной работе реализована система семантического поиска лекарственных препаратов по их текстовому описанию и показаниям к применению. Система позволяет находить наиболее релевантные препараты по симптомам, введенным в свободной текстовой форме, и возвращает список подходящих препаратов с кратким описанием и принадлежностью к терапевтическому классу.

Особенностями предлагаемого подхода являются: использование предобученной модели-трансформера SentenceTransformer (MiniLM), адаптированной для быстрого получения эмбеддингов; использование поиска ближайших соседей по косинусному расстоянию для получения рекомендаций; валидация модели на реальном наборе данных лекарственных препаратов из открытого источника [1].

Таким образом, работа демонстрирует возможность использования современных методов обработки текста в практической медицинской задаче и закладывает основу для дальнейшего развития интеллектуальных рекомендательных систем в фармацевтике и телемедицине.

1 Обзор предметной области

В последние годы растет интерес к использованию интеллектуальных систем поддержки принятия решений в медицине. Такие системы, интегрированные в электронные медицинские карты (ЭМК), помогают врачам принимать обоснованные решения, предлагая планы лечения, диагностические данные и, в некоторых случаях, выбор лекарств. Однако большинство из них построены на жёстких правилах, основанных на экспертных системах, и не способны гибко адаптироваться к свободному вводу текста пользователем или учитывать тонкие семантические различия в симптомах.

Параллельно с развитием систем поддержки принятия решений в медицине наблюдался бурный рост методов обработки естественного языка. Особенно значительный прорыв произошёл с появлением преобразователей и их модификаций, таких как BERT, RoBERTa и их производных. Эти модели, обученные на масштабных корпусах, продемонстрировали высокую способность улавливать контекст и семантику текста и активно используются в анализе клинических данных, например, для автоматической классификации диагнозов, анализа историй болезни и извлечения признаков из медицинской литературы.

Несмотря на эти успехи, задача подбора лекарственных препаратов на основе текстовых описаний остается сложной. Во-первых, тексты, описывающие действие и применение препаратов, различаются по стилю, объему и полноте: некоторые содержат лишь краткое указание диагноза, в то время как другие содержат подробное объяснение механизма действия и

побочных эффектов. Во-вторых, препараты со схожими терапевтическими свойствами могут иметь разные названия и использоваться в разных контекстах, что затрудняет простой поиск по ключевым словам. В-третьих, один препарат может применяться при множестве различных заболеваний, а одно и то же заболевание может лечиться принципиально разными способами в зависимости от индивидуальных факторов.

В связи с этим использование эмбеддингов и трансформеров представляется особенно перспективным: они позволяют перейти от синтаксического анализа к семантическому сравнению. Вместо буквального поиска совпадений система может опираться на глубокое представление смысла текстов, сравнивая их в многомерном векторном пространстве. Это открывает возможности для создания более гибкой, интуитивно понятной и масштабируемой рекомендательной системы, не требующей ручной настройки для каждого случая.

2 Данные

Исследование основано на открытом датасете, размещенном на платформе Kaggle и содержащем обширную информацию о лекарственных препаратах. Источником данных послужил набор медицинских данных (Medical Information Dataset), в котором собраны сведения о составе, применении, механизме действия, побочных эффектах и других характеристиках различных медицинских препаратов. Каждая запись в наборе данных представляет собой текстовую карточку препарата с полями, содержащими как структурированную, так и неструктурированную информацию. Особое внимание в работе уделено столбцам ProductUses, ProductBenefits, HowWorks и Therapeutic_Class, поскольку они несут основную смысловую нагрузку, подходящую для семантического анализа.

Однако исходные данные далеки от чистоты. Они содержат фрагменты HTML-разметки, вложенные теги, спецсимволы, несоответствующие отступы

и явно сгенерированные машинным способом вставки, что затрудняет непосредственное использование текста в модели. Например, некоторые записи содержат хаотичные вставки тегов \n, ul,   и других элементов, не несущих информационной нагрузки. Это делает задачу предобработки критически важной: от качества очистки напрямую зависит корректность векторизации текста, а следовательно, и успешность последующего поиска.

Для повышения однородности и информативности описания был сформирован объединенный текст full_description, включающий очищенные и нормализованные фрагменты из указанных выше столбцов. Такой подход позволил создать единое текстовое представление действия препарата, пригодное для подачи на вход языковой модели.

Пример строки после объединения может выглядеть следующим образом: “Andol 0.5mg Tablet helps restore the chemical imbalances in the brain that are responsible for schizophrenia. It improves thoughts, behavior and enhances the quality of life. It is less likely to cause weight gain compared to similar medicines and should not be stopped without doctor’s advice.” При этом в исходной записи, из которой строился данный текст, содержались вложенные списки, HTML-форматирование и шум, требующие тщательной фильтрации.

Подводя итог, на этапе подготовки данных был решен ряд задач: от фильтрации нерелевантных и пустых записей до объединения текстовых фрагментов в содержательные описания, что стало основой для построения системы семантического поиска.

3 Методология

Разработанная система рекомендации лекарственных препаратов основана на архитектуре, сочетающей простоту реализации и эффективность глубоких языковых представлений. В ее основе лежит идея преобразования текстовых описаний лекарственных препаратов в векторное пространство, где семантически схожие препараты расположены ближе друг к другу. Такой

подход позволяет находить аналоги или релевантные препараты по произвольному текстовому запросу, описывающему, например, симптомы, диагноз или механизм действия.

На первом этапе исходный набор данных подвергается предобработке. Для этого из карточки каждого препарата извлекаются ключевые текстовые поля, очищаются от HTML-шума, пробелов и некорректной разметки и формируются в связное текстовое описание, пригодное для загрузки в модель. Это описание затем используется для построения эмбеддинга с помощью модели SentenceTransformer на основе архитектуры BERT [2].

В качестве ядра для встраивания был выбран вариант MiniLM-L6-v2, демонстрирующий хороший компромисс между качеством векторизации и вычислительной эффективностью. Эта модель способна за короткое время преобразовать длинный медицинский текст в компактное числовое представление, отражающее его смысл. Эмбеддинги имеют фиксированную размерность и содержат информацию о контексте и семантике описания лекарственного препарата.

После получения векторов всех описаний строится индекс ближайших соседей на основе косинусной метрики. Он позволяет эффективно находить k наиболее похожих препаратов по расстоянию между эмбеддингами. Для реализации использовалась реализация NearestNeighbors из библиотеки Scikit-learn. Индекс сохраняется на диск и может быть использован повторно без необходимости повторной векторизации всего корпуса.

Весь процесс построения системы включает последовательную загрузку данных, их очистку и агрегацию, эмбеддинг с помощью предварительно обученной модели преобразователя, построение индекса ближайших соседей и сохранение полученной модели. После этого пользователь может ввести свободный текстовый запрос, например, «anxiety and insomnia», и система в

режиме реального времени предоставит список препаратов, описания которых максимально приближены к введенному запросу (рисунок 1).

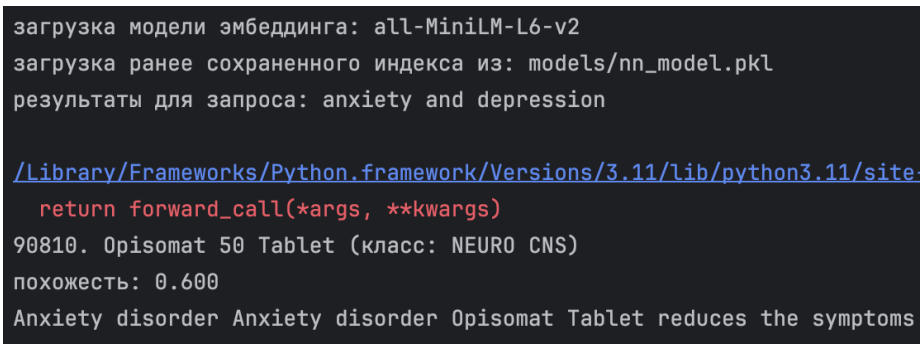

Рисунок 1 — Пример работы системы

4 Валидация

Для оценки качества построенной рекомендательной системы была проведена внутренняя валидация на том же наборе данных, который использовался для обучения. В качестве имитации пользовательского запроса использовалось поле ProductBenefits, содержащее краткое текстовое описание терапевтического эффекта препарата. Этот текст подавался на вход модели, после чего список лучших рекомендаций сравнивался с исходным названием препарата. Если название препарата, для которого был сформирован запрос, попадало в число рекомендуемых, это считалось успешным совпадением.

Основной метрикой качества был Recall@k, отражающий долю таких успешных случаев среди общего числа проверенных примеров. Это позволило оценить способность модели возвращать корректные или релевантные препараты в ограниченном списке наиболее близких кандидатов. Поскольку набор данных содержал тысячи записей, для ускорения проверки использовалась ограниченная подвыборка, включающая 1000 примеров, отобранных на основе наличия содержательного описания в поле ProductBenefits.

Результаты валидации показали высокую эффективность предложенного подхода: при k=5 модель достигла значения Recall@5, равного

0.971. Это означает, что в 971 случаях из 1000 система правильно сопоставила запрос и исходный препарат, что подтверждает релевантность векторных представлений, построенных с помощью SentenceTransformer, и корректность алгоритма поиска ближайших соседей.

5 Анализ результатов и ограничений

Разработанная система продемонстрировала высокую эффективность и простоту использования. Одним из ключевых преимуществ является отсутствие необходимости в контролируемом обучении: модель использует предобученные встраиваемые модели, что позволяет получать быстрые и содержательные рекомендации без затрат на разметку. Кроме того, поиск наиболее близких препаратов осуществляется практически мгновенно, что делает метод пригодным для интеграции в реальные медицинские справочные системы.

Однако, несмотря на всю свою гибкость, предлагаемый подход имеет ряд ограничений. В текущей реализации модель не способна учитывать важные клинические аспекты, такие как противопоказания, побочные эффекты и взаимодействия между действующими веществами. Она опирается исключительно на текстовые описания назначения препарата, что создаёт зависимость от полноты, ясности и стандартизации этих описаний. Кроме того, поскольку вложения были получены с помощью англоязычной модели, система не может быть напрямую применена к русскоязычным данным без предварительной адаптации.

Несмотря на это, существует потенциал для дальнейшего развития данного подхода. В будущем можно рассмотреть возможность добавления дополнительного уровня классификации, например, на основе модели BERT, для отсеивания нерелевантных кандидатов или оценки контекста запроса. Также можно включить генерацию описаний действия лекарственных препаратов, что особенно полезно для создания систем поддержки принятия

решений в телемедицине. Расширение подхода в сторону мультимодальности, например, с учётом химических структур или изображений упаковок, также открывает интересные направления для будущих исследований.

Заключение

В рамках данной работы была реализована система, способная генерировать релевантные рекомендации по лекарственным препаратам на основе текстового описания симптомов или рецепта. Использование предобученных трансформеров для построения семантических векторов в сочетании с индексированием по косинусному расстоянию позволило добиться высокой точности рекомендаций без необходимости ручной разметки данных или обучения сложных моделей.

Разработанное решение может быть использовано в реальных медицинских сценариях, включая телемедицину, чат-боты для первичного информирования пациентов, а также в качестве компонента более сложных систем поддержки принятия клинических решений. Гибкость архитектуры и модульность кода открывают возможности для дальнейших улучшений, включая интеграцию более сложных моделей, многоязычность и подключение дополнительных источников знаний. Работа демонстрирует, что даже при ограниченных ресурсах возможно создание исследовательски значимого прототипа на стыке медицины и искусственного интеллекта.

Список используемой литературы

Электронные ресурсы

1. Medical Information Dataset [Электронный ресурс]. – URL: https://www.kaggle.com/datasets/imtkaggleteam/medical-information-dataset?select=Therapeutic_class_counts-fXyUUS.xlsx (дата обращения: 10.07.2025).
2. SentenceTransformers Documentation [Электронный ресурс]. – URL: https://sbert.net/ (дата обращения: 09.07.2025).

3. The Growing Impact of Natural Language Processing in Healthcare and Public Health [Электронный ресурс]. – URL: https://pubmed.ncbi.nlm.nih.gov/39396164/ (дата обращения: 08.07.2025).
4. BERT [Электронный ресурс]. – URL: https://huggingface.co/docs/transformers/model_doc/bert (дата обращения: 14.07.2025).
5. K-Nearest Neighbor(KNN) Algorithm [Электронный ресурс]. – URL: https://www.geeksforgeeks.org/machine-learning/k-nearest-neighbours/ (дата обращения: 15.07.2025).