© 2024 г. Д.А. ЛЮТКИН (dalyutkin@gmail.com),
Д.В. ПОЗДНЯКОВ (dvpozdnyakov@hse.ru)
(Национальный исследовательский университет
«Высшая школа экономики», Москва),
А.А. СОЛОВЬЕВ (andrey.a.soloviev@gmail.com),
Д.В. ЖУКОВ (dimas.zhukov@gmail.com)
(ООО «Бэбиблог», Москва),
М.Ш.И. МАЛИК, д-р философии (Ph. D.) (mumalik@hse.ru),
Д.И. ИГНАТОВ, канд. техн. наук (dignatov@hse.ru)
(Национальный исследовательский университет
«Высшая школа экономики», Москва)

ПРИМЕНЕНИЕ ТРАНСФОРМЕРОВ ДЛЯ ОПРЕДЕЛЕНИЯ ПРОФИЛЬНОГО ВРАЧА НА ОСНОВЕ ЗАПРОСОВ ПОЛЬЗОВАТЕЛЕЙ[1]

Представлен новый подход, использующий модель RuBERT для классификации пользовательских запросов в области медицинских консультаций с учетом специализации эксперта. В ходе исследования был собран обширный набор данных, который использовался для дообучения модели RuBERT. Метрика качества полученной модели F1-score составила более 91,8% как при использовании блоковой кросс-валидации, так и при разделении набора данных на обучающую и тестовую выборки. Подход демонстрирует высокую обобщающую способность для различных медицинских подобластей, таких как кардиология, неврология и дерматология. Предложенный подход позволяет сократить время на определение наиболее подходящего специалиста и тем самым повышает качество консультации и медицинской помощи.

Ключевые слова: трансформер, медицинский текст, многоклассовая классификация.

DOI: 10.31857/S0005231024030076, **EDN:** TQAELK

1. Введение

Спрос на квалифицированную медицинскую помощь растет, особенно вместе с ростом доступности телемедицины в цифровую эпоху. Поскольку онлайн-платформы выступают в качестве источников медицинской информации [1], все большую важность приобретает обеспечение точности и достоверности консультаций. Одной из таких платформ является Babyblog.ru [2],

[1] Исследование выполнено с использованием суперкомпьютерного комплекса НИУ ВШЭ. Разработано при финансовой поддержке Фонда содействия развитию малых форм предприятий в научно-технической сфере fasie.ru. Два первых соавтора внесли равнозначный вклад в исследование и подготовку статьи к публикации. Исследование второго автора осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ. Онлайн-демо доступно по ссылке: https://www.babyblog.ru/classifier

где пользователи могут задавать вопросы и получать ответы, в том числе и от медицинских специалистов.

Однако большое количество пользовательского контента[2] может ставить под сомнение научную достоверность распространяемой информации [3]. Следовательно, существует необходимость в создании механизмов, обеспечивающих проверку и обогащение пользовательского контента за счет участия экспертов. Такой совместный подход позволяет специалистам проверять сообщения, комментарии и обсуждения пользователей и предоставлять экспертные мнения, корректируя ответы непрофессионалов и обеспечивая предоставление точной и достоверной медицинской информации.

С учетом значительного объема пользовательского контента на различных платформах, охватывающего широкий спектр тем, включая медицинские, псевдомедицинские и немедицинские области, задача идентификации контента, требующего медицинской или профессиональной проверки, становится все более актуальной. Кроме того, классификация на основе тематики и специализации позволит направить пользователя к соответствующему специалисту для консультации.

Для решения этой задачи был разработан автоматический классификатор медицинских текстов, который определяет вероятность принадлежности текста к материалам по той или иной медицинской специальности. Реализация включает в себя интеграцию классификатора на платформе, в которой он идентифицирует медицинский контент и присваивает ему соответствующие медицинские специальности. Впоследствии специалисты каждой из соответствующих специальностей получают уведомления для проверки контента и предоставления своего ответа.

Система классификации позволяет оптимизировать процесс проверки за счет сокращения количества нерелевантной информации, получаемой медицинским специалистом, уменьшения нагрузки, связанной с проверкой контента, и ускорения получения профессиональных ответов пользователем.

Цель исследования — разработать и изучить эффективность системы на основе трансформеров для классификации пользовательского медицинского контента в контексте сайта Babyblog.ru. Используя различные методы обработки естественного языка (Natural Language Processing, NLP), данное исследование улучшает доступность медицинской информации для пользователей с сохранением при этом ее научной строгости и надежности.

2. Обзор релевантной литературы

В сфере классификации медицинских текстов стоит отметить статью [4], в которой предлагается использовать гибридный подход (Hybrid Model, HyM), сочетающий сразу несколько технологий глубокого обучения: LSTM, TEXT-CNN, BERT, TF-IDF, а также механизм внимания (attention mechanism).

[2] Различное информационно-значимое содержимое цифровых носителей, которое создается пользователями.

Предложенный подход позволяет определить, к какому специалисту направить пациента на основе описания симптомов.

В [5] описываются важные аспекты обработки текста, в частности обработка больших текстовых данных, объем которых растет ежедневно. Авторы отмечают необходимость автоматизации обработки текста и приходят к выводу о том, что современные подходы, такие как трансформеры и механизм внимания, могут быть крайне эффективными в этой задаче.

В своем исследовании авторы представили модель двунаправленного трансформера (bidirectional transformer, BiTransformer), построенную на основе двух блоков двунаправленного позиционного кодирования. Такой подход позволяет поместить в один контекст информацию, находящуюся как перед, так и после каждого токена, что улучшает способность модели находить связи в тексте и повышает возможности модели в обработке сложных текстовых данных.

Чтобы оценить эффективность механизмов внимания в процессе классификации, авторы сравнивают четыре модели: с использованием долгой краткосрочной памяти (Long Short Term Memory, LSTM), с механизмом внимания, трансформер и предложенный BiTransformer. Эксперименты проводятся на большом наборе текстов на турецком языке, включающем 30 классов.

В ходе экспериментов было показано, что модели классификации, использующие трансформер и механизм внимания, превосходят классические методы глубокого обучения. Это демонстрирует способность трансформеров выявлять полезные закономерности и учитывать контекст в текстовых данных.

Также авторы изучили влияние использования предобученных векторных представлений (“эмбеддингов”, от англ. embedding) на качество модели. Эмбеддинги хранят семантические представления слов и предобучаются на большом корпусе текстов. Они являются популярным способом улучшить качество модели в задачах обработки текстов на естественном языке (ОТЕЯ). Авторы показывают, как предобученные эмбеддинги могут еще больше увеличить эффективность и точность моделей классификации текстов.

В ходе экспериментов было показано, что среди всех рассмотренных подходов к классификации текстов наилучшие результаты показал предложенный авторами BiTransformer.

Работа [5] дает представление о потенциале механизма внимания и трансформеров в обработке текстов. Появление BiTransformer и его показатели в задаче классификации текстов открывают новые возможности для будущих исследований и применения моделей на основе трансформеров в задачах ОТЕЯ. Результаты исследования имеют важное значение для автоматизации обработки текстовых данных, анализа тональности текста (sentiment analysis), информационного поиска и других сфер, связанных с использованием текстов. Поскольку спрос на эффективные и точные методы обработки текстов продолжает расти, данное исследование вносит значительный вклад в развитие этой области и служит ценным пособием для исследователей и практиков в области обработки естественного языка.

3. Сбор данных: составление исчерпывающего набора данных для классификации медицинских текстов

В этом разделе описывается процесс сбора данных, включающий разработку парсера и применение нормализаторов для дальнейшего использования в эксперименте.

3.1. Парсинг данных

Для получения данных были рассмотрены русскоязычные сайты, которые предоставляют доступ к вопросам пользователей и ответам экспертов на эти вопросы. При выборе источников учитывались следующие критерии:

1) открытость данных по вопросам медицинской тематики,

2) наличие аннотации о медицинской специализации вопроса,

3) наличие верифицированного ответа от эксперта, обладающего соответствующей специализацией.

Таким образом были выбраны следующие источники: sprosivracha.com [6], doctu.ru [7], 03online.com [8] и health.mail.ru [9]. Был разработан парсер, позволяющий параллельно и асинхронно собирать информацию из открытых источников. В ходе сбора с помощью парсера каждый вопрос сохраняется в виде HTML (от англ. HyperText Markup Language — «язык гипертекстовой разметки») документа для дальнейшей обработки. В табл. 1 представлено количество данных, полученных из каждого источника данных.

Таблица 1. Сравнение платформ с медицинскими вопросами

Сайт	Количество вопросов	Процент от общего числа
sprosivracha.com	550 000	23,2
doctu.ru	83 000	3,5
03online.com	1 148 000	48,4
health.mail.ru	590 000	24,9

На следующем шаге алгоритм асинхронно обрабатывает каждый элемент полученных данных и извлекает части, содержащие текст вопроса к врачу и специальность врача. Полученные данные (текст вопроса и специальность врача) заносятся в таблицу вместе с URL (от англ. Uniform Resource Locator — «единообразный указатель местонахождения ресурса») исходного документа, который используется как идентификатор источника. После завершения процесса все данные экспортируются в CSV (от англ. Comma-Separated Values — значения, разделенные запятыми) файл для дальнейшего использования.

3.2. Аугментация данных

Анализ полученных данных показал, что распределение вопросов по специальностям имеет вид, похожий на распределение Парето. Это можно объяснить тем, что некоторые медицинские специальности гораздо более востребованы, чем другие, что приводит к дисбалансу классов [10].

Для минимизации дисбаланса и улучшения способности модели к обобщению были применены некоторые методы аугментации данных, предоставляе-

Рис. 1. Распределение классов после аугментации (первые 70 классов).

мые библиотекой nlpaug [11]. С ее помощью были созданы дополнительные синтетические записи путем перестановки слов в предложениях с сохранением прежнего контекста и смысла предложения, а также перестановки самих предложений.

Благодаря процессу аугментации были созданы дополнительные данные для непопулярных классов, что уменьшило дисбаланс и привело вид распределения к более равномерному по всем медицинским специальностям (см. рис. 1). Кроме того, это позволило улучшить способность модели работать с новыми данными в процессе тестирования.

В результате был получен набор данных с близким к равномерному распределением классов, размер которого составил 5 миллионов вопросов, в соотношении приблизительно 50 000 вопросов на класс для 97 классов.

Несмотря на то, что уже существуют аналогичные наборы данных, полученный набор данных превосходит их по количеству учитываемых заболеваний и медицинских специальностей. Кроме того, большинство существующих наборов данных составлены с использованием профессиональной лексики, которая отличается от описания состояния здоровья обычным человеком. Полученный набор данных устраняет это упущение, что будет способствовать более целостному пониманию опыта людей в области здоровья.

4. Предлагаемый метод

В этом разделе описывается предлагаемый метод, исследуются различные трансформеры и подходы к их обучению. Блок-схема метода представлена на рис. 2 (Start – Начало; Define tokenizer – Определить токенизатор; Define model for Sequence Classification – Определить модель для классификации последовательностей; Data – Данные; Separation into training and validation samples – Разделение на обучающие и проверочные выборки; length train and valid – длина обучающих и проверочных; Creating dataset – Создание набора данных; batch size – размер пакета; Train_dataloader – Загрузчик данных для обучения; Valid_dataloader – Загрузчик данных для проверки; Calculate loss and metrics – Рассчитать потери и метрики; Sending metrics to wandb – Отправка метрик в wandb; Metrics are better than in the previous iteration? – Метрики лучше, чем в предыдущей итерации?; Saved optimizer parameters, weights on file of pth.tar – Сохранены параметры оптимизатора,

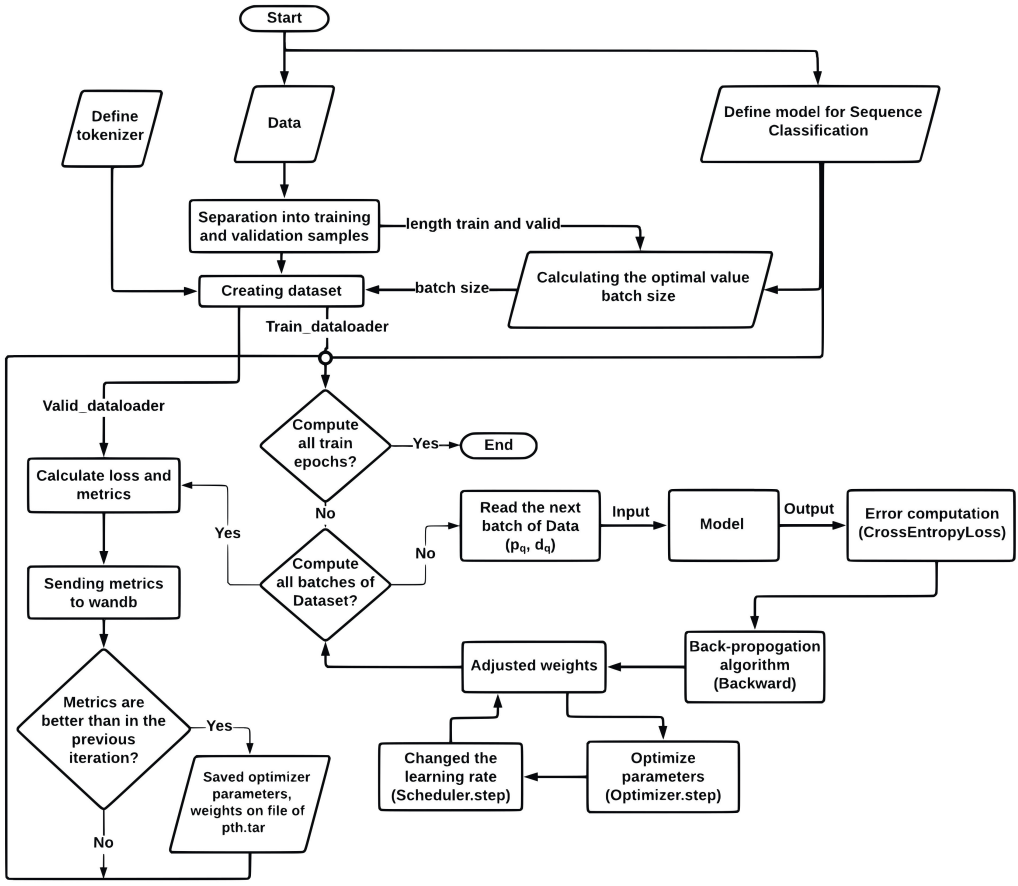

Рис. 2. Блок-схема предлагаемого метода.

веса в файле pth.tar; Compute all train epochs? – Вычислены все эпохи обучения?; Read the next batch of Data (pq, dq) – Считать следующий пакет данных (pq, dq); Adjusted weights – Настроенные веса; Changed the learning rate (Scheduler.step) – Изменена скорость обучения (Scheduler.step); Input – Вход; Model – Модель; Output – Выход; Error computation (CrossEntropyLoss) – Вычисление ошибки (Перекрестная энтропия); Back-propagation algorithm (Backward) – Алгоритм обратного распространения (Backward); Optimize parameters (Optimizer.step) – Оптимизировать параметры (Optimizer.step); End – Конец.).

4.1. Модели-трансформеры

Как правило, нейронные сети обучаются с использованием алгоритма обратного распространения ошибки [12], который оптимизирует параметры модели исходя из минимизации ошибки и улучшения целевой меры качества на тестовой выборке. Однако этот метод сильно зависит от выбора самого алгоритма оптимизации [13], поскольку алгоритм может сойтись к локальному минимуму в процессе вычисления градиентов, что делает модель неспособной

обобщать данные и ухудшает качество предсказания (затухание градиентов). Для решения данной задачи был задействован алгоритм AdamW [14], один из передовых методов оптимизации, который использует информацию об изменении коэффициента скорости обучения для аппроксимации направления антиградиента, а также имеет внутренний "импульс", что улучшает сходимость функции. Этот алгоритм значительно повышает качество обучения, однако он чувствителен к выбору изначального коэффициента скорости обучения. Для регулировки коэффициента скорости обучения был применен косинусный планировщик (cosine scheduler) [15]. Этот планировщик регулирует скорость обучения для каждого пакета (batch) данных. В качестве функции потерь была выбрана кросс-энтропия, поскольку ее величина показывает, насколько хорошо модель справляется с задачей классификации. Применение алгоритма AdamW и планировщика для обучения классификатора на основе BERT было выбрано по следующим причинам.

AdamW Optimizer: AdamW – это версия алгоритма Adam, которая показала свою эффективность в дообучении моделей трансформеров [16]. Данный алгоритм лучше реализует затухание весов, тем самым предотвращает переобучение [17].

Cosine Scheduler: Косинусный планировщик изменяет скорость обучения, начиная с низкой скорости обучения и постепенно увеличивая ее в течение обучения. Период разогрева с низкой скоростью позволяет модели сходиться быстрее, уменьшая нестабильность или колебания функции потерь во время обучения [18].

Трансформеры являются предпочтительным вариантом для классификации медицинских текстов по сравнению с классическими методами, поскольку верно следующее [19].

Предобучение. Трансформеры предобучены на больших корпусах данных, что с самого начала позволяет им получать лучшее представление о языке и закономерностях в нем. Это, в свою очередь, позволяет применять трансформеры в различных задачах после совсем небольшого дообучения.

Контекстное представление. Трансформеры используют двунаправленный механизм внимания, что позволяет им создавать контекстное представление слова в рамках целого предложения. Это важный аспект в задаче классификации текста, которая зависит от качества понимания смысла и контекста предложения.

Перенос обучения (Transfer Learning). Возможно предобучение модели на больших неразмеченных данных из смежной сферы с последующим дообучением на малых размеченных данных из основной сферы. Это позволяет упростить процесс обучения в условиях, когда количество размеченных данных невелико.

Высокая эффективность. Трансформеры показали, что превосходят традиционные методы машинного обучения в различных задачах ОТЕЯ, включая классификацию текстов. Это можно объяснить их способностью улавли-

вать контекстуальные представления и связи между словами, которые имеют большое значение для понимания семантики предложения.

Предобучение на русскоязычных текстах. Модели, предварительно обученные на больших русскоязычных корпусах, показывают более высокое качество в задачах обработки русских текстов по сравнению с обучением с нуля. Предварительное обучение позволяет модели выстроить представление об устройстве языка и упрощает построение контекстных связей.

Важно отметить, что традиционные методы машинного обучения попрежнему широко используются и могут давать хорошие результаты при решении конкретных задач NLP. Однако такие возможности трансформеров, как предобучение, контекстное представление и др., делают это семейство моделей мощным инструментом для классификации медицинских текстов.

4.2. Обучение модели

Используются архитектуры и предобученные веса моделей, полученные с помощью пакета transformers [20]. В самом начале модель инициализируется, а также подготавливается токенизатор с помощью модуля AutoTokenizer пакета transformers. Выходной слой модели изменен вручную для задачи классификации.

Затем подбирается оптимальный размер пакета данных (batch size). Для этого генерируется небольшой синтетический набор данных и затем на основе этого набора данных выполняется поиск по сетке значений (grid search), чтобы определить такой размер пакета, при котором скорость обучения и количество утилизируемых вычислительных ресурсов максимальны. Этот шаг важен для обеспечения максимальной эффективности модели при запуске на удаленном сервере.

В процессе обучения важным моментом является агрегация энергий после применения функции активации Softmax [21]. Итоговая энергия, представленная в виде вероятностных оценок, служит индикатором уверенности модели в отнесении входных данных к определенным классам. Такая мера уверенности играет важную роль в окончательных предсказаниях модели. Важно отметить, что здесь целевые метки — это числовые идентификаторы или номера классов, предварительно закодированные с помощью метода LabelEncoder, а входные данные представляют собой вопросы на естественном языке с описанием жалоб пользователя.

5. Постановка эксперимента

В данном разделе описывается порядок проведения эксперимента, длительность обучения каждой модели и используемое оборудование.

5.1. Оборудование

Для обучения моделей было задействовано два графических ускорителя NVIDIA V100 с 32GB памяти в каждом. Также системе было предоставлено

250GB оперативной памяти для более эффективного хранения и обработки большого набора данных. Обучение моделей выполнялось на суперкомпьютере cHARISMa [22].

5.2. Длительность обучения

Длительность обучения каждой модели зависит от ее архитектуры. Ниже представлены результаты измерений для каждой модели.

- **SBERT [23]**: Модель SBERT потребовала приблизительно 54 ч для обучения. Большая длительность обучения может быть обусловлена глубокой архитектурой и сложным механизмом внимания.
- **LaBSE [24]**: Модель LaBSE потребовала приблизительно 12 ч на обучение. Можно предположить, что на это оказала влияние эффективная архитектура модели и достаточный уровень предобучения.
- **RuBERT [25]**: Обучение модели RuBERT заняло приблизительно 13 ч. Архитектура модели, разработанная специально для применения с текстами на русском языке, потребовала дополнительное время для завершения дообучения.
- **BERT [26]**: Как и LaBSE, модель BERT потребовала приблизительно 12 ч на обучение.
- **BART [27]**: Модель BART потребовала больше времени на обучение – приблизительно 55 ч. Это может быть обусловлено сложностью модели и необходимостью дополнительного обучения структуры кодировщиков–декодировщиков.

Полный цикл оптимизации гиперпараметров, обучения и тестирования, включая кросс-валидацию по k-блокам со значением $k = 3$, занимает от 3 до 12 сут в зависимости от модели.

Для обучения некоторых моделей нужны значительные временные и вычислительные ресурсы, однако степень улучшения показателей моделей и получение сравнительных характеристик нескольких моделей оправдывает усилия, затраченные на дообучение моделей.

6. Результаты экспериментов

В этом разделе представлены результаты экспериментов и их анализ.

На рис. 3 изображены графики кривых обучения моделей LaBSE, SBERT, BERT и BART. Целевая метрика представлена F1-мерой, больше — лучше. Очевидно, что LaBSE показывает лучшую точность по сравнению со многими другими моделями. Кривая обучения LaBSE демонстрирует быструю сходимость и высокое качество. Однако именно для русскоязычного текста модель RuBERT достигает наивысшего качества благодаря предварительному обучению на русскоязычном корпусе текстов.

Модели SBERT, BART и BERT показывают более низкую точность и менее быструю сходимость в рассматриваемой задаче. Модели LaBSE и RuBERT лучше всего подходят для рассматриваемой задачи классификации текстов.

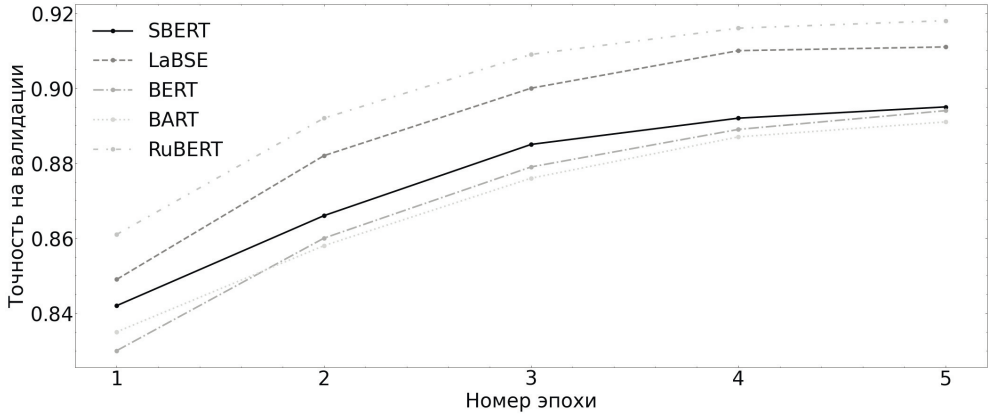

Рис. 3. Кривая обучения различных моделей.

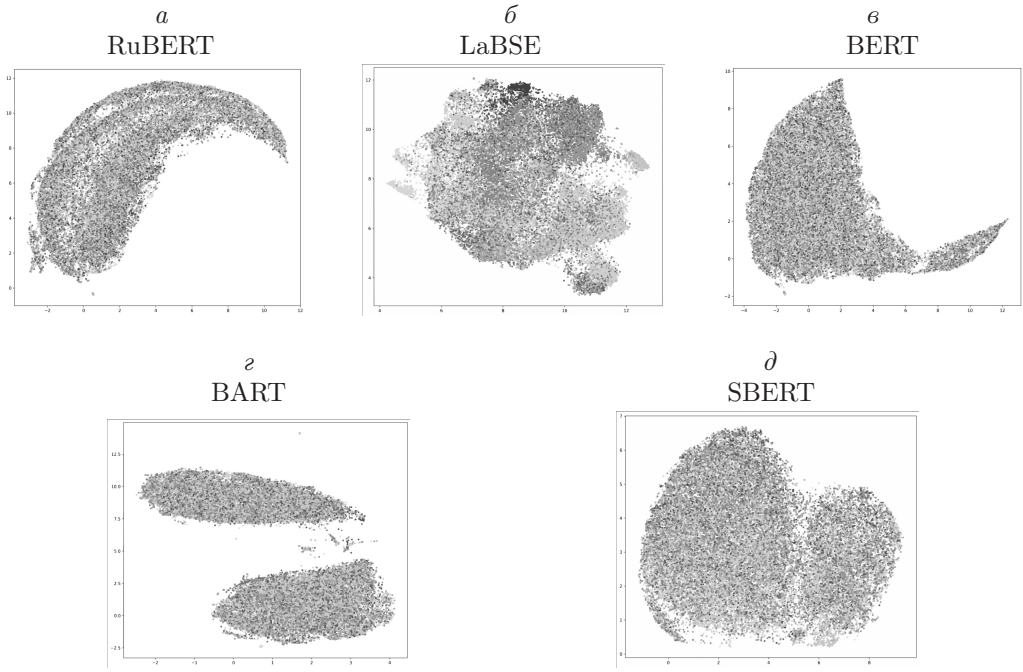

Рис. 4. Визуализация результатов UMAP для различных моделей с использованием собранного набора данных.

Это можно объяснить тем, что LaBSE хорошо различает сущности в тексте, что видно на изображении, полученном с помощью метода UMAP (Uniform Manifold Approximation and Projection[3]), который преобразует многомерные эмбеддинги в двумерное представление. Также видно, что изображение, полученное с помощью UMAP для RuBERT, похоже на другие для методов, которые работают хуже. Однако несмотря на это, благодаря адап-

[3] https://umap-learn.readthedocs.io/

Таблица 2. Сравнение моделей

Модель	K-fold (F1-score, $k = 3$)		Split (F1-score, train — 90%)	
	–	аугментация	–	аугментация
BART	0,798	0,891	0,794	0,896
BERT	0,743	0,894	0,760	0,903
LaBSE	0,824	0,911	0,833	0,913
LogRegression	0,457	0,552	0,531	0,564
Random Forest	0,521	0,579	0,596	0,603
RuBERT	**0,839**	**0,918**	**0,852**	**0,918**
SBERT	0,782	0,905	0,761	0,895
SVM	0,525	0,565	0,534	0,598

Таблица 3. Метрики модели RuBERT для классификации медицинской специальности

Специальность	Точность	Полнота	F1-Оценка	Поддержка
Венеролог	0,7763	0,8112	0,7934	15 140
Гастроэнтеролог	0,7574	0,7339	0,7455	14 839
Гинеколог	0,7834	0,7459	0,7642	14 844
Дерматолог	0,7111	0,6569	0,6829	14 941
Детский хирург	0,8405	0,8782	0,8589	14 847
Инфекционист	0,8409	0,7986	0,8192	14 924
Кардиолог	0,8646	0,8567	0,8606	14 836
ЛОР	0,7555	0,7432	0,7493	15 276
Невропатолог	0,6633	0,5834	0,6206	15 058
Нейрохирург	0,8797	0,9025	0,8910	14 898
Онколог	0,8796	0,8742	0,8769	14 957
Офтальмолог	0,9403	0,9210	0,9305	14 936
Педиатр	0,6482	0,5712	0,6073	15 087
Психолог	0,7759	0,7215	0,7477	15 020
Сексолог-андролог	0,7904	0,6955	0,7399	15 148
Стоматолог	0,8815	0,8893	0,8854	14 861
Терапевт	0,5066	0,3738	0,4302	15 080
Травматолог-ортопед	0,7981	0,7683	0,7829	15 081
Уролог	0,6445	0,6240	0,6341	15 110
Хирург	0,6705	0,5818	0,6230	14 929
Эндокринолог	0,8478	0,8072	0,8270	15 011
точность	0,9111	0,9111	0,9111	0,9031
среднее значение	0,9177	0,9205	0,9189	1 470 000
взвешенное среднее	0,9178	0,9201	0,9189	1 470 000

тации модели для русского языка, RuBERT после дообучения начинает показывать высокую точность, см. рис. 4.

В ходе обучения моделей были измерены результаты для каждой модели в различных режимах обучения.

Таблица 2 дает представление о качестве каждой модели в условиях эксперимента при использовании полученного набора данных. Высокий показа-

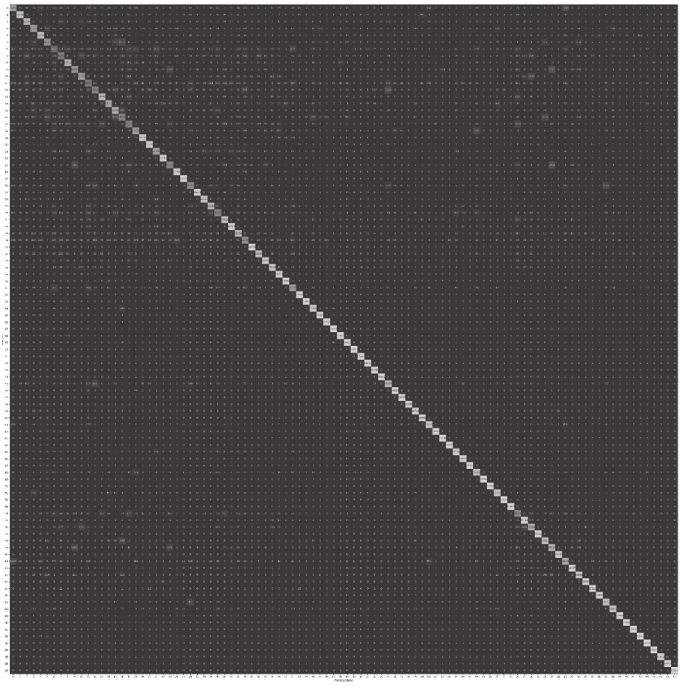

Рис. 5. Матрица ошибок классификатора RuBERT на тестовом наборе данных.

тель метрики у моделей RuBERT и LaBSE говорит о том, что эти модели эффективнее других отражают смысл и контекст текстовых данных в условиях эксперимента. SBERT и BERT также продемонстрировали достаточную производительность, хотя и несколько ниже, чем RuBERT и LaBSE. BART продемонстрировал еще более низкую F1-меру, что может быть объяснено влиянием задачи и конкретного набора данных.

В табл. 3 представлено качество предсказаний модели RuBERT в рамках отдельных классов (21 класс был выбран случайным образом и отсортирован по названию специальности). Представлены точность (precision), полнота (recall), F1-мера (F1-score) и количество текстов в конкретном классе (support). Вместе эти показатели позволяют судить о способности модели корректно различать и предсказывать медицинские специальности по входному тексту.

Матрица ошибок (см. рис. 5) позволила детально изучить результаты классификации, выделив истинно положительные, истинно отрицательные, ложно положительные и ложно отрицательные результаты. Этот анализ выявил заметную тенденцию: большинство ошибок связано со сложностью структуры и семантики настоящих медицинских текстов, где нюансы языка и контекста могут привести к проблемам классификации. Например, обращения, рассматриваемые терапевтом, имеют невысокую полноту (0,3738) в силу возможного общего характера практики (поскольку в реальных больницах посе-

щение терапевта обычно необходимо для обращения к специалистам других профилей).

Стоит отметить, что в ходе оценки модели с использованием только синтетических данных матрица ошибок показывает, что такие данные классифицировались лучше. Это резкое различие показывает, что модель хуже работает с неоднородным, но настоящим текстом пользователя по сравнению с более предсказуемыми и однородными синтетическими данными.

Также были обнаружены некоторые ограничения модели.

Во-первых, длина входного текста ограничена 128 словами и в некоторых случаях ее недостаточно для понимания тонких нюансов, доступных в более длинных данных. При превышении этого порога данные представляются в виде разреженных векторов, что потенциально приводит к потере информации и снижению точности.

Во-вторых, возможное различие в стиле письма тестовых данных и обучающих данных может повлиять на точность предсказаний. Обучение модели только одному определенному стилю ограничивает ее способность адаптироваться к новым, ранее не встречавшимся стилям письма.

Кроме этого, еще одно ограничение создает наличие вопросов, затрагивающих темы, слабо представленные в обучающих данных. Модель хуже предсказывает, когда сталкиваются с вопросами, затрагивающими незнакомые ситуации, поскольку ей не хватает информации о контексте и значении предложения.

7. Заключение

В данной работе был собран исчерпывающий набор данных из открытых источников для классификации медицинских текстов жалоб пользователей среди 97 классов медицинских специальностей по 50 000 экземпляров на класс. Набор данных сбалансирован различными методами аугментации. С использованием полученного набора данных были обучены современные модели трансформеров: SBERT, BERT, LaBSE, BART и RuBERT. Модель RuBERT показала лучшее качество с F1-мерой 91,8%. Полученные результаты позволяют сделать вывод о том, что модели-трансформеры, в частности RuBERT, крайне эффективны в задаче классификации текстов. Способность трансформеров “захватывать” контекстное представление и обнаруживать сложные закономерности в естественном языке делает их значительно точнее по сравнению с классическими методами.

Дальнейшие исследования могут быть направлены на изучение применимости этих моделей трансформеров для дообучения на небольших наборах данных или на узкоспециализированных наборах данных. Кроме того, существует потенциал для разработки новой архитектуры трансформеров, специально предназначенных для задач классификации текстов. Эти архитектуры могут включать в себя знания, специфичные для конкретной области, что, в свою очередь, еще больше повысит точность классификации.

Также перспективным направлением будущих работ может стать исследование новых методов переноса обучения, стратегий дообучения и оптимизации гиперпараметров для моделей-трансформеров. Существуют широкие возможности для развития области классификации медицинских текстов с использованием моделей-трансформеров, и эти будущие работы могут способствовать разработке более точных и эффективных моделей.

СПИСОК ЛИТЕРАТУРЫ

1. Trusting Social Media as a Source of Health Information: Online Surveys Comparing the United States, Korea, and Hong Kong / H. Song // J. Medic. Internet Res. 2016. V. 18. No. 3. P. 25. URL: https://www.jmir.org/2016/3/e25. https://doi.org/10.2196/jmir.4193
2. БэбиБлог — Ответы на любые вопросы о беременности, детях и семейной жизни. Accessed: December 19 , 2022. https://www.babyblog.ru/
3. *Keshavarz H.* Evaluating credibility of social media information: current challenges, research directions and practical criteria // Inform. Discover. Deliver. 2021. V. 49. No. 4. P. 269–279. https://doi.org/10.1108/IDD-03-2020-0033
4. Automatic medical specialty classification based on patients' description of their symptoms / C. Mao / BMC Medical Informatics and Decision Making. 2023. V. 23. https://doi.org/10.1186/s12911-023-02105-7
5. *Tezgider M., Yildiz B., Aydin G.* Text classification using improved bidirectional transformer // Concurrency and Computation: Practice and Experience. 2022. V. 34. No. 9. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.6486. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.6486. https://doi.org/https://doi.org/10.1002/cpe.6486
6. СпросиВрача: Задай вопрос врачу онлайн и получи ответ мгновенно. Accessed: February 17, 2023. https://sprosivracha.com/
7. ДОКТУ — поиск лучших врачей и клиник в России. Accessed: February 17, 2023. https://doctu.ru/
8. 03 Онлайн — медицинские консультации в режиме онлайн. Accessed: February 17, 2023. https://03online.com/
9. health.mail.ru — Поиск по болезням, лекарствам и ответам врачей. Accessed: February 17, 2023. https://health.mail.ru/
10. *Johnson J.M., Khoshgoftaar T.M.* Survey on deep learning with class imbalance // Journal of Big Data. 2019. V. 6. No. 1. P. 27. https://doi.org/10.1186/s40537-019-0192-5
11. *Ma E.* NLP Augmentation. 2019. Accessed: February 17, 2023. https://github.com/makcedward/nlpaug
12. *Hecht-Nielsen R.* III.3 – Theory of the Backpropagation Neural Network (Based on "nonindent" by Robert Hecht-Nielsen, which appeared in Proceedings of the International Joint Conference on Neural Networks 1, 593–611, June 1989). © 1989 IEEE / Neural Networks for Perception / H. Wechsler (Ed.). Academic Press, 1992. P. 65–93. ISBN 978-0-12-741252-8. https://doi.org/10.1016/B978-0-12-741252-8.50010-8. URL: https://www.sciencedirect.com/science/article/pii/B9780127412528500108

13. *Shaheen Z., Wohlgenannt G., Filtz E.* Large Scale Legal Text Classification Using Transformer Models. 2020. arXiv: 2010.12871 [cs.CL]
14. Understanding AdamW through Proximal Methods and Scale-Freeness / Z. Zhuang. 2022. arXiv: 2202.00089 [cs.LG]
15. Automated Learning Rate Scheduler for Large-batch Training / C. Kim. 2021. arXiv: 2107.05855 [cs.LG]
16. Attention Is All You Need / A. Vaswani. 2017. arXiv: 1706.03762 [cs.CL]
17. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes / Y. You. 2020. arXiv: 1904.00962 [cs.LG]
18. Are Transformers more robust than CNNs? / Y. Bai // Advances in Neural Information Processing Systems. 2021. P. 34. Curran Associates, Inc. P. 26831–26843. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/e19347e1c3ca0c0b97de5fb3b690855a
19. A Survey on Text Classification: From Shallow to Deep Learning / Q. Li. 2021. arXiv: 2008.00364 [cs.CL]
20. Transformers: State-of-the-Art Natural Language Processing / T. Wolf [et al.] // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online : Association for Computational Linguistics. 2020. P. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6
21. *Maida A.* Cognitive Computing and Neural Networks: Reverse Engineering the Brain / Handbook of Statistics. V. 35. Elsevier. 2016. P. 39–78. https://doi.org/10.1016/bs.host.2016.07.011 URL: https://doi.org/10.1016/bs.host.2016.07.011
22. *Kostenetskiy P.S., Chulkevich R.A., Kozyrev V.I.* HPC Resources of the Higher School of Economics / J. Physics: Conf. 2021. P. 1740. No. 1. P. 012050. https://doi.org/10.1088/1742-6596/1740/1/012050 URL: https://dx.doi.org/10.1088/1742-6596/1740/1/012050
23. *Reimers N., Gurevych I.* Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 2019. arXiv: 1908.10084 [cs.CL]
24. Language-agnostic BERT Sentence Embedding / F. Feng. 2022. arXiv: 2007.01852 [cs.CL]
25. *Kuratov Y., Arkhipov M.* Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. 2019. arXiv: 1905.07213 [cs.CL]
26. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin. 2019. arXiv: 1810.04805 [cs.CL]
27. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension / M. Lewis. 2019. arXiv: 1910.13461 [cs.CL]

Статья представлена к публикации членом редколлегии А.А. Галяевым.

Поступила в редакцию 08.07.2023

После доработки 07.11.2023

Принята к публикации 20.01.2024