Алгоритм сбора текстов для анализа тональности и тематического моделирования отзывов пациентов поликлиник

А. Д. Белкин, М. С. Коган, М. В. Болсуновская

Санкт-Петербургский политехнический университет Петра Великого

redloin@mail.ru, m_kogan@inbox.ru, bolsun_mv@spbstu.ru

Аннотация

Пользовательский фидбек является ценным источником информации для оценки качества услуг, оказываемых медицинскими учреждениями, в частности поликлиниками, и выявления проблемных аспектов. Однако, процесс сбора пользовательских отзывов представляет достаточно сложную техническую задачу ввиду отсутствия единых требований сбора отзывов пациентов на сайтах поликлиник, которым бы следовали все учреждения здравоохранения. Собираемые традиционным способом данные такого типа, как правило, не являются открытыми. В данной работе рассматривается процесс сбора и анализа пользовательских отзывов о медицинских учреждениях на примере поликлиник города Санкт-Петербурга. Для сбора данных был разработан алгоритм на языке программирования Python, использующий веб-скрапинг. Собранный набор данных содержит более 64 тысяч отзывов о 350 поликлиниках. Собранные данные будут предобработаны и использованы для обучения моделей анализа тональности и тематического моделирования. Приводятся результаты предварительного анализа пользовательских отзывов, которые характеризуются большим лингвистическим и стилистическим разнообразием. Полученные результаты могут быть использованы медицинскими учреждениями для улучшения качества обслуживания и повышения удовлетворенности пациентов. Кроме того, модель может быть применена для сравнительного анализа различных учреждений и выявления областей, требующих внимания.

Ключевые слова: отзывы пациентов поликлиник, открытые данные, Яндекс Карты, веб-скрапинг, естественная обработка языка, анализ тональности, тематическое моделирование

Библиографическая ссылка: Белкин А. Д., Коган М. С., Болсуновская М. В., Алгоритм сбора текстов для анализа тональности и тематического моделирования отзывов пациентов поликлиник // Компьютерная лингвистика и вычислительные онтологии. Выпуск 8 (Труды XXVII Международной объединенной научной конференции «Интернет и современное общество», IMS-2024, Санкт-Петербург, 24–26 июня 2024 г. Сборник научных статей). – СПб: Университет ИТМО, 2024. С. 70–78. DOI: 10.17586/2541-9781-2024-8-70-78.

1. Введение

По мере стремительного развития Интернета новый вид приобретают и различные онлайн-платформы, социальные сети, форумы и т. д. Некоторые подобные площадки агрегируют отзывы интернет-пользователей, описывающие какой-либо товар или заведение.

Отзыв представляет собой текст произвольной длины, написанный пользователем и выражающий его личное отношение к тому или иному объекту. Часто платформы дают возможность оценивать степень удовлетворенности в баллах от 1 до 5.

Пользовательский фидбек может быть очень ценным источником информации для организаций, направленных на предоставление различных услуг людям [1]. В отзывах пользователи могут указывать на неочевидные проблемы, недостатки или несоответствие действительности информации, указанной на сайте организации. Это может помочь быстро выявить и своевременно устранить недочеты. Большое количество положительных отзывов влияет на приток новых клиентов, так как люди в первую очередь обращают внимание на рейтинг и оценки. Также, положительные отзывы могут быть использованы в рекламных целях, демонстрируя их на сайте или в социальных сетях. С другой стороны, отзывы могут помочь в оценке конкурентных заведений или товаров, например, понять сильные и слабые стороны, выявить новые тренды и возможности для роста.

Эффективными инструментами изучения пользовательских отзывов являются анализ тональности и тематическое моделирование [2]. Тематическое моделирование позволяет выделить ключевые темы и проблемы, которые обсуждаются в отзывах [3]. Вдобавок, можно сравнивать различные группы пользователей и выявлять их потребности и предпочтения.

В свою очередь, анализ тональности комментариев и отзывов помогает понять то, какие эмоции выражают пользователи. Обученный на достаточно большом объеме данных алгоритм анализа тональности поможет исследовать данные, не имеющие таких же оценок или меток, как на онлайн-платформах. Существует множество традиционных и современных способов сбора отзывов, которые используются в настоящее время. К традиционным можно отнести бумажные формы для отзывов, как, например, книги жалоб и предложений или заполнение опросных листов [4]. Среди современных способов можно выделить электронные опросы, мониторинг социальных медиа и телефонные интервью. Такие данные не содержат конкретной метрики, говорящей об эмоциональном отношении опрошенного. В таких случаях и может пригодиться модель анализа тональности.

Особенно важно понимать потребности и опыт пациентов в сфере здравоохранения, так как их мнение является ключевым показателем качества. Интернет дал людям возможность выражать свои мнения на различных платформах, что создало большой объем данных, исследование которых может выявить тенденции, не замеченные ранее, в медицинских услугах. Исследователи из Имперского колледжа Лондона собрали комментарии пациентов с сайта NHS Choices за 2008, 2009 и 2011 годы (13 802 текста) и применили методы анализа тональности [5]. Данные за 2010 год (6 412 комментариев) использовались для проверки точности предсказания. Целью было обучение модели машинного обучения для автоматического определения рекомендаций пациентов, чистоты медицинского учреждения и уважительного отношения. Алгоритм классифицировал комментарии на основе примеров и проверял точность предсказания, сравнивая результаты с оценками пациентов по шкале Ликерта. Исследование показало высокую точность предсказаний, подтверждая соответствие прогнозов результатам опросов.

В другом исследовании группа ученых собрала корпус данных с сайта Otzovik, состоящий из 1,4 миллиона пользовательских отзывов на русском языке о здоровье и лекарствах, а также 500 подробно аннотированных отзывов [6]. Корпус называется RuDReC и находится в публичном доступе. Ученые затем использовали эти тексты для обучения моделей Multi-BERT, RuBERT и RuDR-BERT, сравнив их способность извлекать из отзывов пользователей информацию о названии лекарств, побочных эффектах, заболеваниях и здоровье.

В еще одной работе исследователи так же собрали корпус из 2 800 комментариев пользователей на форуме Otzovik. В статье анализируется то, как выбор различных компонентов модели влияет на точность распознавания сущностей. Исследователи применяют тот же подход к аннотированию текстов, что и авторы [6], и обучают на них

алгоритм нейронной сети. Результатом работы стало то, что был установлен новый уровень точности извлечения биомедицинских сущностей для русского языка на полноразмерном размеченном корпусе [7]. Собранные для исследования данные могут быть получены по запросу.

Также стоит отметить статью, в которой описывается способ выявления связей между лекарствами на основе семантического анализа текстовых данных [8]. Авторы применили веб-краулинг для сбора текстов с различных медицинских порталов. Всего было собрано примерно 2,5 млн текстов, из которых были извлечены названия препаратов, описание их действия и другие сущности. Затем полученные данные были векторизованы и сравнены между собой. Эксперименты показали, что такой подход может успешно идентифицировать препараты с похожими терапевтическими эффектами.

Однако в машинном обучении существует проблема нехватки размеченных данных, необходимых для обучения моделей. В открытом доступе не всегда можно найти достаточно информации или уже готовых «датасетов» с метками. Особенно это касается такого домена, как медицинские учреждения. В данном случае нехватка частично обусловлена тем, что структура сайтов поликлиник и больниц не располагает отдельной секцией для отзывов и оценки работы учреждения. Кроме того, очень много данных находится в закрытом доступе, так как они содержатся на бумажных носителях и не оцифрованы.

В последнее время обсуждают важность открытости данных и кода для сотрудничества государств в борьбе с глобальными проблемами. Доклад ЮНЕСКО подчеркивает, что открытые данные обеспечивают доступ к необходимой информации для исследований, политики и мер против глобальных кризисов, таких как COVID-19 [9]. Искусственный интеллект помогает анализировать большие объемы данных, выявляя скрытые закономерности для принятия решений. Для публикации данных нужно выполнить три этапа: разработка политики управления и обмена данными, решение о публикации данных и привлечение пользователей. На первом этапе определяются обязательства, принципы, законодательные ссылки, причины недоступности данных, сбор и обучение пользователей. Затем принимается решение о том, какие наборы данных будут опубликованы и на условиях лицензирования. После этого данные публикуются на веб-сайте. Этап привлечения пользователей включает в себя информирование, консультирование, поддержку международного взаимодействия и предотвращение злоупотреблений. Также, необходимо регулярно обновлять данные для сохранения их актуальности. Можно заметить, что предлагаемая процедура является цикличной поскольку после каждой публикации новых данных, она должна заново проходить все описанные этапы.

Тем не менее, далеко не все страны придерживаются указанных выше рекомендаций и принципов. По этой причине нехватка качественных наборов данных, необходимых для обучения алгоритмов машинного обучения, является актуальной проблемой во многих областях науки.

В связи с этим было принято решение сформировать подобный корпус с нуля при помощи компьютерных средств [10].

2. Разработка алгоритма

Для сбора информации и создания корпуса был разработан алгоритм и реализован на языке программирования Python. В качестве источника данных, с которого программа собирает отзывы, был выбран вебсайт онлайн-карт «Яндекс Карты». Формирование набора данных опирается на оценки, выставленные пользователями учреждениям. На сайте они имеют вид «звездочек» в диапазоне от 1 до 5.

Принцип работы алгоритма сводится к тому, что сначала задаются параметры, по которым будет идти поиск организаций и учреждений на сайте. В результате поиска выпадает список карточек организаций, который затем прокручивается вниз, пока не

достигнет заданного числа искомых организаций. С каждой карточки собирается ссылка, ведущая на соответствующие отзывы. Каждая ссылка обрабатывается, а отзывы, находящиеся на странице, прокручиваются до конца. Затем из них, оценок и имен пользователей формируется набор данных при помощи библиотеки «pandas». Наконец, сформированный «датасет» передается внешней таблице, вбирающей в себя все результаты обработки.

В нашем исследовании для реализации алгоритма использовались следующие библиотеки Python: «Selenium», «Beautiful Soup», «pandas» и «pymorphy3». Первые две библиотеки из перечисленных, необходимы для взаимодействия с веб-браузером и извлечения данных с сайта, соответственно. «Selenium» автоматизирует работу веб-браузеров, а «Beautiful Soup» облегчает «веб-скраппинг», разбирая HTML- и XML-файлы [11].

Программная реализация алгоритма состоит из 6 функций, которые в процессе работы многократно вызываются. Прежде всего задаются параметры веб-драйвера, который будет моделировать действия пользователя в браузере, давая тем самым доступ к необходимой информации.

На первом этапе указывается, какой браузер из 5 доступных будет использоваться для работы, поскольку структура сайтов для разных браузеров отличается. Для данного исследования был выбран «Google Chrome». Для ускорения работы кода следует применить специальный параметр, запрещающий использовать режим с графическим интерфейсом браузера. Данный режим весьма полезен, но только при разработке и экспериментах, так как он показывает каждое действие, которое было запрограммированно. В случае обычного использования программы нет необходимости следить за процессом моделирования работы браузера, так как это может оказать дополнительную нагрузку на компьютер.

На втором этапе задаются функции, которые затем объединяются в главной, исполняющей и автоматизирующей весь процесс. Первая функция очень важна, так как она необходима для имитации пролистывания страницы при помощи ползунка, поскольку одной из главных сложностей сбора данных стала процедурная загрузка отзывов и учреждений на сайте. Без этого решения программа может собирать только первые 5 предзагруженных текстов и ссылок.

Третья функция отвечает за циклический поиск в коде сайта ссылок, ведущих с карточек найденных учреждений на отзывы к ним.

Следующая функция совершает парсинг отзывов, поочередно переходя по заданному списку ссылок. Поскольку сайт не всегда успевает загрузиться быстро, необходимо искусственно замедлить работу кода, давая возможность прогрузиться всем нужным элементам. В зависимости от количества отзывов программа подсчитывает приблизительное число прокруток ползунка, чтобы загрузить все тексты. Для этого используется библиотека «time», которая тормозит процесс выполнения кода на нужное количество секунд. Также на этом этапе вместе с отзывами собираются и соответствующие им оценки от 1 до 5.

Стоит отметить, что не все пользователи ставят оценки. Хотя по нашим наблюдениям это довольно редкое явление, его необходимо учитывать: в противном случае, сталкиваясь с отсутствием оценки, код будет аварийно завершаться. Такие отзывы включаются в набор данных, но со специальной меткой. На этапе предобработки данных подобные тексты будут удалены или размечены вручную, если их количество будет достаточно большим. Помимо оценок данная функция собирает имена пользователей, которые затем используются для определения пола комментатора.

На финальном этапе главная функция принимает на вход параметры поиска данных: тип организации, город для сбора отзывов, район и количество организаций, которое нужно обработать. Количество организаций указывается исследователем или иным пользователем, желающим сформировать набор данных в каком-либо домене. Эти параметры попадают в поисковую строку на сайте, после чего начинают работать другие функции. При этом

функция контролирует, чтобы параметры вводились правильно. Если какой-то параметр не был введен, то в работе будет использовано базовое значение этого параметра.

Если было найдено недостаточное число организаций, запускается цикл, дополняющий список ссылок до тех пор, пока не наберется нужное количество. В случае, если было обработано слишком много сайтов, то список будет обрезан до обозначенного числа. Во избежание бесконечного поиска количество итераций ограничено максимальным числом допустимых циклов, к которым прибавляется еще 5 дополнительных.

Когда процесс сбора завершен, функция, используя цикл, объединяет все полученные данные в единый датасет и возвращает его для дальнейшей работы.

В рамках исследования программа собрала 64 451 отзыв о 350 поликлиниках Санкт-Петербурга. Время работы программы составило примерно 3 часа. В таблице приведены 5 примеров из собранного набора отзывов (в отзывах сохранена авторская орфография; ФИО упомянутых в отзывах сотрудников медучреждений были закодированы путём сокращения имен, фамилий и отчеств до инициалов). Фамилии упомянутых врачей скрыты с целью сохранения безопасности личной информации.

Таблица. Примеры отзывов из полученного набора данных

Номер строки	Текст отзыва	Оценка
60250	Хорошая поликлиника, нареканий нет!	5
43027	Лучшая поликлиника, все врачи отличные специалисты. В ПК есть лекотека со своим логопедом, психологом, массажистом во благо развития здоровья детей. Отдельно отмечу невролого до года О. О. В. и педиатра Т. В. Ф. внимательное отношение к детям и бережное отношение ко мне как к маме. Отвечали на все мои вопросы	5
40054	Сложно попасть к специалистам	3
3586	Хорошие специалисты	5
7818	Нехватает специалистов. Электронная очередь и по записи. Все равно у кабинетов толпа без билетников. Лезут без очереди. Спрашивается зачем тогда предварительная запись, когда все и так проходят. Бардак. Хотя врачи попадаются толковые. Но некоторые спешат, так как торопятся поскорее убежать на платный прием	3

Как показано на рисунке, основные темы, выявленные в отзывах, включают обсуждение медицинского персонала, медицинских учреждений, процедур, записей и самих процессов на прием ко врачу, а также инфраструктуры.

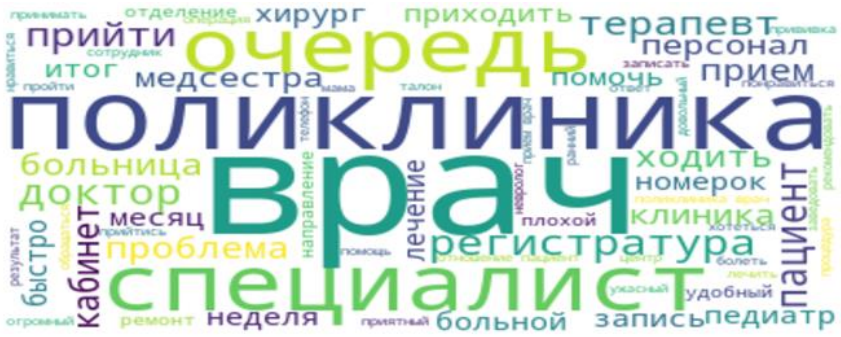

Рис. Облако слов, сформированное на основе отзывов

Поскольку нет строгих правил и ограничений, диктующих то, как надо писать отзывы, в них содержится множество индивидуальных лингвистических особенностей пользователей, которые будут кратко описаны в следующем разделе.

3. Предварительный лингвистический и стилистический анализ собранного датасета

Пациенты поликлиник, оставляющие отзывы, часто используют в своих сообщениях разнообразные фразеологизмы, сленг и эмодзи. Они также часто совершают грамматические и пунктуационные ошибки и выделяют отдельные слова заглавными буквами. По причине того, что отзывы посвящены медучреждениям, пользователи часто описывают свой опыт посещения врачей, упоминая конкретные даты, имена врачей и другую информацию. К примеру, некоторые пациенты выражают благодарность отдельным врачам, которые оказали им помощь. Другие, наоборот, критикуют кого-то, обращаются к администрации больницы или поликлиники, призывая принять меры.

Зачастую пользователи выражают свое отношение весьма эмоционально, используя в тексте скобочки для демонстрации эмоции. Их количество может варьироваться от одной до нескольких десятков. Они могут выражать как радость («))»), так и печаль («((»), а их количество указывает на силу испытанных эмоций. Более того, в некоторых случаях множественные скобочки сопровождают язвительные, саркастические и иронизирующие комментарии. Кроме того, пациенты могут ставить многоточие. В каких-то случаях многоточие используется при перечислении большого списка объектов, например специальностей врачей, либо для выражения расстройства, когда пациент не получил того, что ожидал. Также, для акцентирования внимания на каких-либо конкретных аспектах, пользователи выделяют слова заглавными буквами. Как правило, это встречается в негативных отзывах, например, при описании большого времени ожидания в очереди («ЧАС») или при неудовлетворительном качестве обслуживания в платной поликлинике («ПЛАТНАЯ»).

Вдобавок ко всему перечисленному, пациенты регулярно используют слэнг, неформальные слова и выражения. Они могут быть специфичны для определенных социальных групп, возрастных категорий или регионов и зачастую могут меняться и развиваться со временем (выявление этих закономерностей может стать предметом отдельного исследования в будущем). Например, часто можно встретить преобразованные для простоты употребления медицинские термины. Вместо слова «флюорография» вероятнее увидеть слово «флюшка». Очень много пользователей используют большое количество разговорных выражений в своих отзывах. Так, они подчеркивают ту или иную эмоцию, которую стремятся донести до других. К примеру, некоторые пользователи так описывают хорошую работу медперсонала: «Медсестра колет как богиня» или «просто крутой профессионал своего дела». Негативные отзывы так же очень богаты на примеры: «аж волосы дыбом», «фикалии в уши льют...» или «с горем пополам профосмотр был пройден».

Отдельно можно упомянуть эмодзи, которые встречаются в пользовательских отзывах. Эмодзи — это набор символов, представляющих широкий спектр эмоций, объектов, действий и идей. Они широко используются в современной коммуникации в социальных сетях, мессенджерах, форумах и других онлайн-платформах. Эмодзи позволяют выразить эмоции более точно и наглядно, чем простые текстовые сообщения, а в некоторых случаях, они могут заменять слова или целые фразы, экономя место и делая сообщение более лаконичным.

4. Дальнейшее исследование

Полученный набор данных станет основой для обучения ряда моделей анализа тональности, из которых будет выбрана наиболее эффективная. Однако перед этим данные следует предобработать. Для этого из данных будут извлечены в отдельный столбец эмодзи, которые были использованы пользователями для усиления эмоциональной составляющей отзыва. Столбец с эмодзи может быть использован для улучшения качества модели

тональности текста. Кроме того, данные нужно проверить на дубликаты, так как некоторые пользователи могут писать несколько полностью совпадающих отзывов. Они могут быть посвящены разным учреждениям, но одинаковые комментарии от одного пользователя снижают качество данных. В полученном наборе данных обнаружилось чуть более 2 тысяч дубликатов. Затем данные разделяются на токены, слова приводятся к основе и лемматизируются. Это необходимо для уменьшения размера словаря и улучшения обобщающей способности модели. После этого данные преобразуются в числовой формат, необходимый для обучения модели. Для этого используется один из существующих методов векторизации текстовых данных.

За обучением модели анализа тональности следует обучение алгоритма тематического моделирования. Выделение ключевых тем в отзывах с разделением на положительные и отрицательные, а также с разделением пользователей на группы позволит проанализировать наиболее важные потребности, недостатки и преимущества, которые находятся в текстах.

5. Заключение

Итогом данного исследования стал алгоритм, который собирает пользовательские отзывы на сайте «Яндекс Карты», используя настраиваемые параметры. Также было проанализировано влияние отзывов об различных организациях и составлен набор данных, содержащий отзывы пользователей о поликлиниках в городе Санкт-Петербург. Эти данные будут использованы для обучения модели анализа тональности. Сам алгоритм легко настраивается на сбор аналогичных данных о медучреждениях или других организациях, предоставляющих услуги населению, расположенных в любом регионе страны, при условии, что информация о них содержится на сайте «Яндекс Карты».

Обученная модель может помочь оценить удовлетворенность пациентов услугами медицинских учреждений, которые в свою очередь принять во внимание выявленные проблемы и предпринять меры по улучшению качества услуг. Более того, модель может использоваться для сравнительного анализа различных медицинских учреждений с целью выявления лучших практик и определения областей, в которых могут быть внесены улучшения.

Литература

[1] Dhahak K., Huseynov F. The Impact of Online Consumer Reviews (OCR) on Online Consumers' Purchase Intention // Journal of Business Research-Turk. 2020. Vol. 12 (2). P. 990–1005. URL: https://isarder.org/2020/vol.12_issue.2_article01.pdf (дата обращения 21.03.2024).

[2] Самигулин Т., Джурабаев А. Анализ тональности текста методами машинного обучения // Научный результат. Информационные технологии. 2021. № 1. URL: https://rrinformation.ru/journal/annotation/2376/ (дата обращения: 01.03.2024).

[3] Косарева Е., Давыдик Н. Применение тематического моделирования для интеллектуального анализа отзывов на русском языке // Дистанционные образовательные технологии: сб. трудов V Междунар.науч.-практ. конф., Симферополь, 22–25 сент. 2020 г. Симферополь: ИТ «АРИАЛ», 2020. С. 228–231. URL: https://elib.grsu.by/doc/65168 (дата обращения: 20.03.2024).

[4] Бурый М., Кравцова Т. Книга отзывов и предложений как элемент совершенствования качества услуг // E-Scio. 2023. №5 (80). URL: https://e-scio.ru/?p=21000 (дата обращения: 20.03.2024).

[5] Greaves F., Ramirez-Cano D., Millett C., Darzi A., Donaldson L. Use of sentiment analysis for capturing patient experience from free-text comments posted online // Journal of Medical Internet Research. 2013. Т. 15 (11): e239. DOI: 10.2196/jmir.2721. URL: https://www.jmir.org/2013/11/e239/ (дата обращения: 24.03.2024).

[6] Tutubalina E., Alimova I., Miftakhutdinov Z., Sakhovskiy A., Malykh V., Nikolenko S. The Russian Drug Reaction Corpus and Neural Models for Drug Reactions and Effectiveness Detection in User Reviews // Bioinformatics. 2020. T. 37 (2). DOI: 10.1093/bioinformatics/btaa675.
[7] Sboev A., Sboeva S., Moloshnikov I., Gryaznov A., Rybka R., Naumov A., Selivanov A., Rylkov G., Ilyin V. Analysis of the Full-Size Russian Corpus of Internet Drug Reviews with Complex NER Labeling Using Deep Learning Neural Networks and Language Models // Applied Sciences. 2022. T. 12 (1). DOI: 10.3390/app12010491.
[8] Tutubalina E., Miftakhutdinov Z., Nugmanov R., Madzhidov T., Nikolenko S., Alimova I., Tropsha A. Using semantic analysis of texts for the identification of drugs with similar therapeutic effects // Russian Chemical Bulletin. 2017. Vol. 66 (11). P. 2180–2189.
[9] Ziesche S. Open data for AI. What now? Paris: UNESCO, 2023. URL: https://unesdoc.unesco.org/ark:/48223/pf0000385841. 64 p. (дата обращения: 24.03.2024).
[10] Чижик А., Мельникова С., Захаров В. Социальное картирование на основании анализа тональности комментариев в социальных сетях // International Journal of Open Information Technologies. 2022. Т. 10, № 11. С. 75–79. URL: http://injoit.org/index.php/j1/article/view/1434 (дата обращения: 04.03.2024).
[11] Бастрикина В. Проектирование веб-скрапера для получения данных с сайтов книжных издательств // Актуальные проблемы авиации и космонавтики. 2018. № 14. С. 125–127.

Text Collection Algorithm for Sentiment Analysis and Topic Modelling of Patient Reviews for Polyclinics

A. D. Belkin, M. S. Kogan, M. V. Bolsunovskaya

Peter the Great St. Petersburg Polytechnic University

User feedback is a valuable source of information for assessing the quality of services provided by health care facilities, particularly polyclinics, and identifying problematic aspects. However, the process of collecting user feedback is technically challenging, as there are no uniform requirements for collecting patient reviews on polyclinic websites that all health care institutions should follow. This type of data collected in the traditional way is usually not open. This paper examines the process of collecting and analysing user feedback on health care institutions on the example of polyclinics in the city of St. Petersburg. For data collection, an algorithm was developed in the Python programming language using web scraping. The collected data set contains more than 64 thousand reviews of 350 polyclinics. The collected data will be preprocessed and used to train tone analysis and topic modelling models. The results of a preliminary analysis of user reviews, which are characterised by a large linguistic and stylistic diversity, are presented. The results can be used by healthcare providers to improve the quality of care and patient satisfaction. In addition, the model can be applied to benchmark different institutions and identify areas that require attention.

Keywords: polyclinic patients' feedback, open data, Yandex Maps, web scraping, natural language processing, sentiment analysis, topic modelling

Reference for citation: Belkin A. D., Kogan M. S., Bolsunovskaya M. V. Text Collection Algorithm for Sentiment Analysis and Topic Modelling of Patient Reviews for Polyclinics // Computational Linguistics and Computational Ontologies. Vol. 8 (Proceedings of the XXVII International Joint Scientific Conference «Internet and Modern Society», IMS-2024, St. Petersburg, June 24–26, 2024). — St. Petersburg: ITMO University, 2024. P. 70–78. DOI: 10.17586/2541-9781-2024-8-70-78.

Reference

[1] Dhahak K., Huseynov F. The Impact of Online Consumer Reviews (OCR) on Online Consumers' Purchase Intention // Journal of Business Research-Turk. 2020. Vol. 12 (2). P. 990–1005. URL: https://isarder.org/2020/vol.12_issue.2_article01.pdf (access date: 21.03.2024).

[2] Samigulin T., Djurabaev A. Analiz tonalnosti teksta metodami mashinnogo obucheniya // Nauchnyy rezultat. Informatsionnye tekhnologii. 2021. № 1. URL: https://rrinformation.ru/journal/annotation/2376/ (access date: 01.03.2024). (In Russian)

[3] Kosareva E., Davydik N. Primenenie tematicheskogo modelirovaniya dlya intellektual'nogo analiza otzyvov na russkom yazyke // Distantcionnye obrazovatel'nye tekhnologii: sb. trudov V Mezhdunar. nauch.-prakt. konf., Simferopol', 22–25 sent. 2020 g. — Simferopol': IT "ARIAL", 2020. — S. 228–231. URL: https://elib.grsu.by/doc/65168 (access date: 20.03.2024). (In Russian)

[4] Buryy M., Kravtsova T. Kniga otzyvov i predlozheniy kak element sovershenstvovaniya kachestva uslug // E-Scio. 2023. № 5 (80). URL: https://e-scio.ru/?p=21000 (access date: 20.03.2024). (In Russian)

[5] Greaves F., Ramirez-Cano D., Millett C., Darzi A., Donaldson L. Use of sentiment analysis for capturing patient experience from free-text comments posted online // Journal of Medical Internet Research. 2013. T. 15 (11): e239. DOI: 10.2196/jmir.2721. URL: https://www.jmir.org/2013/11/e239/ (access date: 24.03.2024).

[6] Tutubalina E., Alimova I., Miftakhutdinov Z., Sakhovskiy A., Malykh V., Nikolenko S. The Russian Drug Reaction Corpus and Neural Models for Drug Reactions and Effectiveness Detection in User Reviews // Bioinformatics. 2020. T. 37 (2). DOI: 10.1093/bioinformatics/btaa675.

[7] Sboev A., Sboeva S., Moloshnikov I., Gryaznov A., Rybka R., Naumov A., Selivanov A., Rylkov G., Ilyin V. Analysis of the Full-Size Russian Corpus of Internet Drug Reviews with Complex NER Labeling Using Deep Learning Neural Networks and Language Models // Applied Sciences. 2022. T. 12 (1). DOI: 10.3390/app12010491

[8] Tutubalina E., Miftakhutdinov Z., Nugmanov R., Madzhidov T., Nikolenko S., Alimova I., Tropsha A. Using semantic analysis of texts for the identification of drugs with similar therapeutic effects // Russian Chemical Bulletin. 2017. Vol. 66 (11). P. 2180–2189.

[9] Ziesche S. Open data for AI. What now? Paris: UNESCO, 2023. URL: https://unesdoc.unesco.org/ark:/48223/pf0000385841. 64 p. (access date: 24.03.2024).

[10] Chizhik A., Mel'nikova S., Zaharov V. Social'noe kartirovanie na osnovanii analiza tonal'nosti kommentariev v social'nyh setyah // International Journal of Open Information Technologies. 2022. T. 10, № 11. S. 75–79.URL: http://injoit.org/index.php/j1/article/view/1434 (access date: 04.03.2024). (In Russian)

[11] Bastrykina V. Proektirovanie veb-skrapera dlya polucheniya dannykh s saytov knizhnykh izdatel'stv // Aktual'nye problemy aviatsii i kosmonavtiki. 2018. № 14. S. 125–127. (In Russian)