Прикладная лингвистика
Applied linguistics

Научная статья
УДК 81’33 + 004.8
DOI: 10.14529/ling240409

МОДЕЛЬ РЕГРЕССИОННОГО АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТА ДЛЯ ОЦЕНКИ УРОВНЯ УДОВЛЕТВОРЕННОСТИ КЛИЕНТА

О.И. Бабина, *babinaoi@susu.ru*
Южно-Уральский государственный университет, Челябинск, Россия

Аннотация. Статья посвящена моделированию автоматизированного извлечения тональной информации об удовлетворенности клиента из текста отзыва. В условиях стремительного роста объемов текстовой информации анализ тональности текста становится ключевым инструментом принятия решений в сфере маркетинга, социологии, политологии и в других областях. Это обусловливает интерес к разработке точных и масштабируемых методов анализа тональности как одного из ключевых направлений обработки естественного языка сегодня. Целью данного исследования является разработка модели анализа тональности текста для решения задачи предсказания степени удовлетворенности пользователя медицинским учреждением по тексту отзыва, с применением гибридного подхода на основе лексиконов и машинного обучения. Работа выполнена на материале корпуса отзывов о частных клиниках Челябинска с портала 2GIS объемом 100 тыс. словоупотреблений. Единицы корпуса с помощью предметно-ориентированного тонального лексикона отнесены к четырем тональным классам (сильно отрицательная, умеренно отрицательная, умеренно положительная и сильно положительная оценка). В данной работе предложена модель множественной линейной регрессии для предсказания степени удовлетворенности пользователя на основе параметров, в качестве которых выступают доли тонально размеченных единиц в тексте. Модель построена и обучена методом гребневой регрессии с настройкой параметра регуляризации через кросс-валидацию. Построенная модель показала высокую точность предсказаний пользовательских рейтингов медцентров со среднеквадратической ошибкой 0,0226 и коэффициентом детерминации 0,8182. Таким образом, предложенная модель на основе гибридного подхода подтвердила свою эффективность в предсказании оценок удовлетворенности по текстам.

Ключевые слова: анализ тональности текста, оценка удовлетворенности, гибридный подход, предметно-ориентированный тональный лексикон, множественная линейная регрессия, гребневая регрессия, медицинский центр, отзыв

Для цитирования: Бабина О.И. Модель регрессионного анализа тональности текста для оценки уровня удовлетворенности клиента // Вестник ЮУрГУ. Серия «Лингвистика». 2024. Т. 21, № 4. С. 63–70. DOI: 10.14529/ling240409

Original article
DOI: 10.14529/ling240409

REGRESSION-BASED SENTIMENT ANALYSIS MODEL FOR PREDICTING CUSTOMER SATISFACTION

O.I. Babina, *babinaoi@susu.ru*
South Ural State University, Chelyabinsk, Russia

Abstract. The paper focuses on the development of a model for the automated extraction of customer satisfaction information from textual inputs. As sentiment analysis has emerged as a pivotal tool for decision-making in the fields such as marketing, sociology, political science and others, it becomes particularly important in the context of the rapid expansion of textual information. Consequently, this promotes the growing interest in developing precise and scalable sentiment analysis methods, positioning it as a critical area in up-to-date natural language processing. The objective of this study, therefore, is to develop a sentiment analysis model to tackle the challenge of predicting customer satisfaction with medical institutions based on review texts. Specifically, this is achieved through a hybrid approach that integrates lexicon-based techniques and a machine learning methodology. The research material of the study is a corpus of reviews on private medical centers in Chelyabinsk, sourced from the 2GIS portal, and encompassing 100,000 word usages. Evaluative lexical units within this corpus have been labeled by sentiment tags – strongly negative, moderately negative,

© Бабина О.И., 2024.

moderately positive, and strongly positive – using a domain-specific sentiment lexicon. In this paper, we proposea multiple linear regression model for predicting customer satisfaction, leveraging parameters defined as the proportions of units labeled by each sentiment within the text. The model has been developed and trained as a ridge regression with L2-regularization, employing cross-validation techniques. The model demonstrated high accuracy in forecasting user ratings of medical centers, achieving a mean squared error of 0.0226 and the coefficient of determination of 0.8182.

Keywords: sentiment analysis, customer satisfaction, hybrid approach, domain-specific sentiment lexicon, multiple linear regression, ridge regression, medical center, review

For citation: Babina O.I. Regression-based sentiment analysis model for predicting customer satisfaction. *Bulletin of the South Ural State University. Ser. Linguistics.* 2024;21(4):63–70. (in Russ.). DOI: 10.14529/ling240409

Введение

Анализ тональности текста (сентимент-анализ) представляет собой сегодня одну из ключевых задач в области обработки естественного языка. Решение задачи предполагает автоматизированное выявление оценочности в текстах, эмоциональной окраски текста, что позволяет классифицировать тексты как выражающие положительную или отрицательную оценку описываемого объекта (продукта, услуги, социальной ситуации и прочего) или не содержащие оценочности (нейтральные). В условиях стремительного увеличения объемов текстовой информации, возникающей в результате повседневного общения, социальных сетей и онлайн-отзывов, анализ тональности становится незаменимым инструментом в маркетинге и рекламе для мониторинга мнений о продуктах и услугах компании с целью своевременного реагирования на негативные отзывы и адаптации маркетинговых стратегий [14], в сфере обслуживания клиентов для персонализации взаимодействия с клиентами [8], в политических исследованиях для выявления предпочтений избирателей и предсказания результатов выборов [23], в социологии и психологии для мониторинга общественного мнения, настроений и эмоционального состояния населения [10, 22], в финансовой сфере, где сентимент-анализ применяется для анализа настроений инвесторов и иных лиц, вовлеченных в биржевые операции, что впоследствии является основой для прогнозирования изменений на фондовом рынке [6], и в других сферах.

Подходы к моделированию анализа тональности сегодня включают:

1) подход на основе лексиконов: такой подход основан на нахождении в корпусе тональных лексических единиц, зафиксированных в словаре тональной лексики. В качестве такого словаря могут использоваться существующие лексиконы и тезаурусы (например, SentiWordNet [4]), RuSentiLex [13], а также составляться специальные предметно-ориентированные лексиконы на основе корпусов текстов [9, 20];

2) подход на основе машинного обучения: в рамках реализации этого подхода обучаются модели на текстовых данных для автоматического определения сентимента. В исследованиях предпринимаются попытки построения моделей анализа тональности как с использованием классических алгоритмов машинного обучения (k-ближайших соседей, наивный байесовский классификатор, машины опорных векторов [11, 18, 19, 21] и другие алгоритмы – см. обзор в [2]), а также моделей глубокого обучения на основе нейросетей, включая CNN, LSTM, BiLSTM [7; 17], и предобученных больших языковых моделей [5, 12, 16];

3) гибридный подход: эти методы сочетают подходы на основе лексиконов и машинного обучения [15, 24].

В данной работе мы предлагаем метод предсказания по тексту степени удовлетворенности клиента на основе гибридного подхода к анализу тональности текста. Метод включает построение модели линейной регрессии на основе корпуса текстов отзывов на медицинские учреждения, при этом в качестве независимых объясняющих переменных используются данные о присутствии в тексте оценочной лексики, найденной с помощью поиска по предметно-ориентированному тональному словарю.

Материал и методы

Материалом исследования послужил корпус русскоязычных отзывов, собранный с портала 2GIS и включающий отзывы о частных клиниках города Челябинска, оставленные на портале в период от начала функционирования платформы до 15 февраля 2023 года. В дополнение к текстам отзывов корпус содержит метаданные текстов, включая уникальный идентификатор текста, название медицинского центра, источник публикации, пользовательский рейтинг – оценку, выставленную клинике автором отзыва по шкале от одного до пяти.

Случайная выборка из собранного корпуса, включающая 1587 текстов объемом 100 тыс. словоупотреблений, была на предыдущих этапах исследования размечена лингвистами НОЦ «Лингвоинновационные технологии» ЮУрГУ с применением платформы концептуального аннотирования [3]. Базу знаний для аннотирования составил предметно-ориентированный словарь, включающий 1) тональную одно- и многокомпонентную лексику (оценочные предикаты), 2) слова-модификаторы оценочных предикатов, 3) лексические единицы, репрезентирующие аспекты оценки медицинского центра (например, Качество обслуживания, Персонал, Запись на прием и т. п.), 4) единицы, при-

надлежащие другим концептуальным классам, которые характеризуют предметную область, но не подвергаются оценке со стороны клиентов (например, Пациент, Метод лечения и т. п.). Некоторые концептуальные классы и соответствующие им примеры лексических единиц базы знаний, используемой для аннотирования корпуса, приведены в таблице.

При разметке платформа заключает единицу, найденную в словаре, в фигурные скобки и через знак «тильда» указывает метки концептуальных классов, обозначенные для единицы в словаре. Таким образом, размеченные с помощью платформы тексты корпуса имеют вид:

{Были}~UNK {на}~O {приеме}~D {у}~O {лора}~DR {Шафикова}~UNK {АЗ}~UNK – {назначено}~UNK {адекватное лечение}~Q~POS, {промывание}~D {дало результат}~Q~POS – {уже}~UNK {на следующий день}~T {ребенок}~P {начала дышать носом}~Q~POS, {после}~UNK {повторного промывания}~D – {надеюсь}~UNK {болезнь}~D {отошла}~UNK {от}~O {моей}~UNK {девочки}~P {надолго}~O. {Благодарна}~POS {клинике}~C – {время посещения}~T {подобрали}~UNK {быстро}~POS ,

где метка *O* указывает на общую, неспецифичную для предметной области лексику, имеющуюся в словаре, метка *UNK* обозначает, что лексическая единица не представлена в словаре, значение остальных меток можно увидеть в табл. 1.

В данной работе мы далее рассматривали только разметку тонального типа. Схема разметки оценочной лексики включала четыре тональные метки: POSX – сильно положительная оценка, POS – умеренно положительная оценка, NEG – умеренно отрицательная оценка, NEGX – сильно отрицательная оценка. Метками POS и NEG помечались лексические единицы, манифестирующие, как правило, рациональную, не маркированную эмоционально оценку. Например,

Обслужили {***без очередей***}~POS, {***быстро***}~POS, {***качественно***}~POS, {***современно***}~POS .

{***Неделю ждала***}~NEG *результата. Помощи я* {***так и не получила***}~NEG.

Метки POSX, NEGX использовались для обозначения: а) (преувеличенно) эмоциональных оценочных лексических единиц, б) оценочных предикатов (как правило, выраженных прилагательным), обозначающих максимальную интенсивность / степень выраженности признака, в) многокомпонентных лексических единиц, содержащих модификаторы-усилители. Например,

{***Ужасно***}~NEGX ! *Обслуживание* {***отвратительное***}~NEGX ! (эмотивная лексика)

{***Отличная***}~POSX *клиника и* {***классный***}~POSX *стационар* (максимальная интенсивность признака)

Выражаю {***огромную*** *благодарность*}~POSX *доктору* <имя>, *я ему* {***очень*** *признательна*}~POSX (модификатор-усилитель)

Результатом аннотирования стал «золотой» корпус, где концептуальная неоднозначность была разрешена вручную [20]. В частности, ручная коррекция тональной разметки требовалось в случаях, когда оценочные единицы использовались в ироничных высказываниях, где встречались лексические единицы, отмеченные в словаре как положительные, однако использующиеся в конкретном контексте как окказионально негативные. Например:

На просьбу хотя бы сфотографировать малыша мне сказали «что там фотографировать?

Лексико-концептуальная база знаний для разметки корпуса

Концептуальный класс	Метка	Лексические примеры	Тип
Умеренно положительная оценка	POS	высокий профессионализм, отзывчивый, помог	Тональность
Сильно положительная оценка	POSX	суперотличный доктор, огромнейшей души человек	Тональность
Умеренно отрицательная оценка	NEG	не работала, унижать, не смогли поставить диагноз	Тональность
Сильно отрицательная оценка	NEGX	хамский, содрать деньги, беспредел, шок	Тональность
Усиление признака	MAG	очень, чрезвычайно, от души	Модификатор
Ослабление признака	ANT	в меру, вроде	Модификатор
Врач	DR	доктор, врач, педиатр	Аспект
Клиника	C	бесплатная поликлиника, роддом	Аспект
Персонал	S	девочки на ресепшн, персонал	Аспект
Качество медобслуживания	Q	качество работы, подробно объяснил	Аспект
Время приема	T	время посещения, ожидание, простоял в очереди	Аспект
Метод лечения	D	рентген, операция, анализы	Другие
Пациент	P	постоянный клиент, ребенок, пожилой пациент	Другие

index	Text	NEGX	NEG	POS	POSX	Count	Rating
0	Очень хороший медицинский центр. Впервые мне понравилось платить за услуги. Цены приемлемы. Регистратура потрясающая. Врачи безупречные. Рекомендую.	0	0	4	2	6	5
1	Рекомендую крутого специалиста отоларинголога Карпенко, максимально грамотный подход к работе. Спасибо за вашу работу!	0	0	3	1	4	5
2	Лежала на послеоперационном восстановлении. повязки всегда вовремя меняли, обход с утра был, постоянно спрашивали, как я себя чувствую. По соотношению цена-качество хороший вариант.	0	0	2	0	2	5
3	Ужасно! Отвратительно! Отдала за приём нейрохирурга 2 тыс рублей, а получила 20 минут молчания и бумажку со словами: квчн Я все написал квчк . Общие фразы, много воды, никакой конкретики, закончился приём фразой: квчн Идите к другому врачу квчк . Я требую возврата денежных средств, мне не оказана нужная помощь.	2	7	0	0	9	1
4	Отвратительная работа операторов, каждый раз при звонке дают разную информацию. Один оператор говорит одно , другой другое. Точной информации нет… только когда придёшь в клинику узнаешь достоверно.	1	1	0	0	2	1
5	12.07.2022 Были на приёме у онколога, ул. Труда 187б. Я была удивлена поведением сотрудниц в регистратуре, как они громко разговаривали и простите, просто квчн ржали квчк на всю поликлинику, вы вообще-то с онкобольными работаете, и так себя ведёте, было ощущение что я на базаре нахожусь. Администрация, театр начинается с вешалки, проведите беседу.	0	6	0	0	6	3
6	Была на приеме 22 июня. Приходила не первый раз и всегда все было отлично: Отдельные входы для детей и взрослых, удобно записываться, приветливые администраторы, хорошие специалисты, большой выбор услуг, адекватные цены. Но в последний раз с парковкой беда, т.к. рядом парк развлечений и зоопарк(надо бы отдельную парковку для клиентов), придя на прием мне сказали что описание будет только через 2 часа после проведения процедуры, и удивились, что при звонке мне это не сказали. Было очень неприятно, особенно, что мне нужны были результаты именно сейчас для другого врача, а не через часа.	1	3	8	1	13	4
7	Были у травматолога, хороший специалист, на этом все. Космические цены просто. Наложили гипс ровно как 40 лет назад, претензий бы не было, если бы стоило все это в 2 раза дешевле. А то стоит как немецкая медицина а по итогу копеечная, нудобная, большая накладка.	0	7	2	0	9	2

Рис. 1. Данные по количеству оценочных единиц в текстах отзывов

Кружочек с палочкой?». {***Очень «приятно»***}~NEGX , {***спасибо***}~NEG !

Для каждого текста в «золотом» корпусе мы подсчитали количество лексических единиц, используемых в текущем тексте. Таким образом, для каждого текста были получены количественные данные по четырем параметрам, соответствующим тональным меткам. Примеры текстов с информацией о количестве оценочных единиц различных тональных классов приведены на рис. 1. В таблице на рис. 1 в колонке index расположен уникальный идентификатор текста отзыва в корпусе, колонка Text хранит оригинальный текст отзыва, в колонках NEGX, NEG, POS, POSX указано количество лексических единиц в тексте отзыва, которые при разметке получили соответствующую метку, Count содержит общее количество тонально размеченных единиц в тексте, Rating содержит числовую оценку от 1 до 5, которую оставил пользователь вместе с текстом отзыва.

С учетом задачи последующего моделирования регрессии для полученных числовых данных была проведена нормализация с целью устранения влияния различий в длине отзывов, улучшения сходимости метода и повышения надежности результатов.

Значения пользовательских рейтингов были нормализованы посредством деления на 5, таким образом значения пользовательского рейтинга клиники r_i в наборе данных принимали значения из множества: $r_i \in \{0{,}2; 0{,}4; 0{,}6; 0{,}8; 1.0\}$.

Нормализация значений в колонках POS, POSX, NEG и NEGX включала деление количеств тональных меток в тексте на общее количество тонально размеченных единиц в тексте (значение в колонке Count для i-го текста):

$$p_i^j = \frac{n_i^j}{Count_i}, \quad j \in \{NEGX, NEG, POS, POSX\}, \quad i \in [1..N],$$

где p_i^j – нормализованное значение j-й тональной метки для i-го отзыва корпуса, n_i^j – количество единиц, помеченных j-й тональной меткой в i-м отзыве корпуса , $Count_i$ – общее количество тонально размеченных лексических единиц в i-м отзыве корпуса, N – общее количество отзывов в корпусе.

Полученные таким образом значения p_i^j мы рассматривали как независимые объясняющие переменные в задаче множественной регрессии, r_i – как целевую зависимую переменную моделирования. Тогда классическую линейную модель множественной регрессии можно представить в виде

$$r_i = \beta_0 + \beta_{NEGX} \cdot p_i^{NEGX} + \beta_{NEG} \cdot p_i^{NEG} + \beta_{POS} \cdot p_i^{POS} + \beta_{POSX} \cdot p_i^{POSX} + \varepsilon_i,$$

где $i \in [1..N]$, β_0, β_{NEGX}, $\beta_{NEG}, \beta_{POS}, \beta_{POSX}$, – параметры модели, ε_i – случайная ошибка.

Для выбора метода нахождения параметров регрессии мы оценили коэффициенты корреляции между объясняющими переменными построенного набора данных. Матрица корреляций представлена на рис. 2. Как можно видеть, между переменными

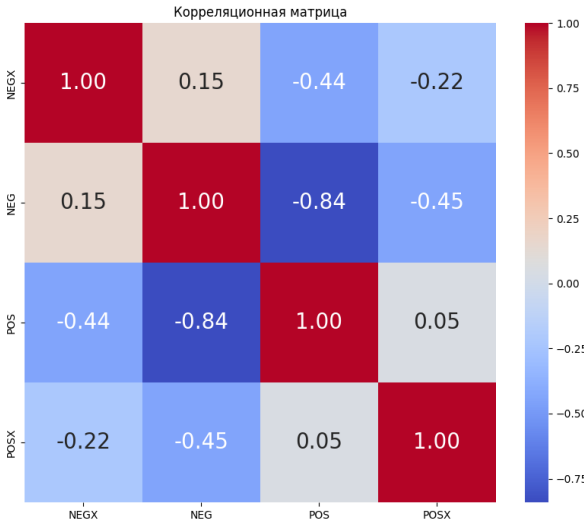

Рис. 2. Корреляционная матрица объясняющих переменных модели

POS и NEG наблюдается достаточно сильная отрицательная корреляция (коэффициент корреляции равен –0,84), что указывает на мультиколлинеарность независимых переменных. Этого можно было ожидать, так как наличие оценочных лексических единиц в силу специфики дискурса отзывов является практически всегда обязательным компонентом текста, а методика определения значений p_i^j, предполагающая нахождение долей каждой из переменных в тексте, предполагает тесную взаимосвязь между полученными значениями. Однако приведение к единой шкале со значениями POS и NEG на концах не представляется целесообразным, так как один и тот же отзыв может содержать одновременно позитивные и негативные элементы, каждый из которых должен быть учтен в модели. Кроме того, рассмотрение POS и NEG как отдельных параметров обеспечит возможность в будущем дифференцированно извлекать мнения по отдельным аспектам.

В связи с этим для преодоления влияния мультиколлинеарности мы проводили моделирование с помощью метода гребневой регрессии. Вычисление параметров модели проводилось путем минимизации суммы квадратов отклонений между наблюдаемыми и предсказанными значениями с L2-регуляризацией [1, с. 60]:

$$Ridge = \sum_{i=1}^{N}(r_i - \hat{r}_i)^2 + \lambda \sum_j {\beta_j}^2 =$$

$$= \sum_{i=1}^{N}\left(r_i - \sum_j \beta_j p_i^j\right)^2 + \lambda \sum_j {\beta_j}^2 \to min,$$

$$j \in \{NEGX, NEG, POS, POSX\},$$

где r_i – наблюдаемые значения нормализованного рейтинга i-го отзыва, $\hat{r}_i$ – предсказанные значения рейтинга i-го отзыва, вычисленные согласно построенной модели регрессии, $Ridge$ – значение суммы квадратов отклонений наблюдаемого и предсказанного рейтингов с L2-регуляризацией, β_j – параметры модели регрессии.

Поиск коэффициентов реализовывался на Python. Поиск оптимального значения коэффициента регуляризации λ выполнялся с применением кросс-валидации с делением набора данных на 5 подвыборок, для чего был создан ряд из 100 кандидатов значения λ в диапазоне от 10^{-4} до 10^4 с равномерным распределением в логарифмическом масштабе. Построение модели гребневой регрессии, включая нахождение оптимального коэффициента λ и параметров регрессии β_j, осуществлялось на базе класса RidgeCV из библиотеки scikit-learn. Для построения регрессионной модели выборка отзывов была разделена на обучающую и тестовую в соотношении 9:1. Таким образом, в обучающую выборку вошли 1428 отзывов, в тестовую – 159 отзывов.

Результаты и обсуждение

Изменение усредненного по 5 подвыборкам значения среднеквадратической ошибки (MSE) для оценки коэффициента регуляризации λ в ходе кросс-валидации на обучающей выборке показано на рис. 3. Оптимальное значение коэффициента регуляризации, минимизирующее среднюю ошибку по подвыборкам обучающего набора данных, было получено при $\lambda = 0{,}2675$.

Полученное в результате настройки модели на обучающей выборке уравнение регрессии для определения степени удовлетворенности клиента медицинским центром приняло вид:

$$r_i = 0{,}6108 - 0{,}4574 \cdot p_i^{NEGX} - 0{,}3301 \cdot p_i^{NEG} + 0{,}3939 \cdot p_i^{POS} + 0{,}3937 \cdot p_i^{POSX}$$

Полученная модель хорошо интерпретируется. Свободный член полученной модели устано-

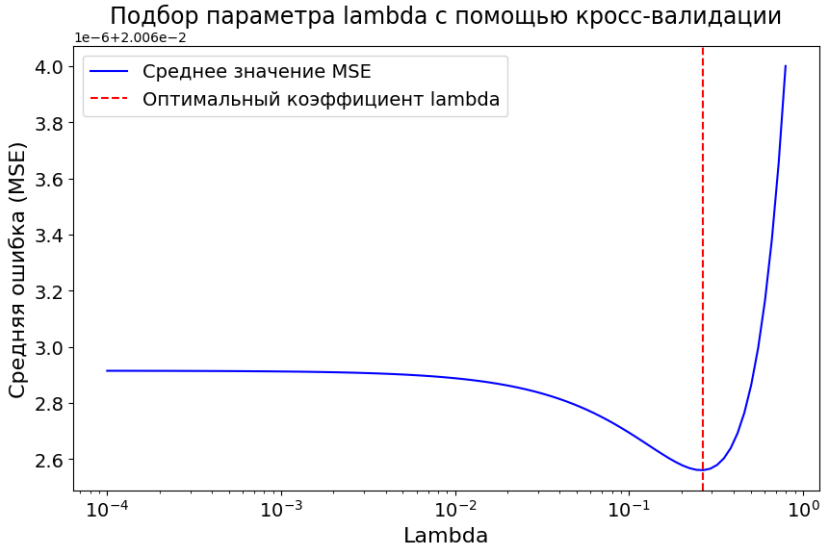

Рис. 3. График изменения средней ошибки для различных коэффициентов λ модели гребневой регрессии

вился на уровне, близком к середине диапазона значений пользовательских рейтингов клиники, заданных пользователями (середина диапазона наблюдаемых значений равна 0,6). Преобладание в тексте лексических единиц, размеченных негативными тональными метками, приводит к уменьшению значения рейтинга клиники (для соответствующих объясняющих переменных получены отрицательные коэффициенты), а наличие положительных тональных меток, наоборот, увеличивает значение рейтинга.

Среднеквадратическая ошибка предсказаний модели оказалась равной 0,0226, средняя абсолютная ошибка предсказаний составила 0,0884. Таким образом, в среднем отклонения от истинного значения являются незначительными. Коэффициент детерминации модели равен 0,8182, что указывает на достаточно высокую объясняющую способность построенной модели.

Примеры предсказаний, полученных с помощью построенной модели линейной регрессии на тестовой выборке, приведены на рис. 4.

Данные на рис. 4 демонстрируют довольно точное соответствие предсказанных значений истинным значениям рейтингов, выставленных пользователями (ср. значения в колонках *Предсказанные значения* и *Истинные значения*).

Таким образом, построенная модель линейной регрессии позволяет на основании значений параметров, показывающих доли лексических единиц, размеченных разнополярными тональными метками, предсказывать значение пользовательского рейтинга клиники. Очевидно, такая модель требует наличия процедуры, позволяющей на основе входного текста определять доли тонально размеченных лексических единиц в тексте. В настоящем исследовании для обучения модели использовался «золотой» корпус текстов, где тональные метки текстовым единицам сообщались на основе словаря и далее корректировались вручную. Однако автоматизация процедуры предсказания рейтинга оцениваемого объекта на основе регрессионного анализа тональности текстов требует построения процедуры определения однозначных тональных меток в тексте автоматически. Полагаем, в качестве такой процедуры может рассматриваться применение тонального словаря в совокупности с правилами разрешения неоднозначности концептуальных меток, а также экспериментирование с применением методов машинного обучения для определения тональной принадлежности единиц текста.

Заключение

В данной работе представлена модель предсказания пользовательского рейтинга медицинских учреждений на основе регрессионного анализа тонально маркированной лексики в тексте отзыва. В качестве объясняющих переменных в модели регрессии рассматриваются доли лексических единиц в тексте, имеющих одну из множества тональных меток {NEGX, NEG, POS, POSX}. Целевая переменная представляет собой числовую оценку в диапазоне [0, 1], показывающую степень удовлетворенности пользователя клиникой, которая может варьировать от крайнего недовольства от посещения клиники (0) до полной удовлетворенности качеством оказанных услуг соответственно (1). Тональные метки назначаются текстовым единицам на основе предметно-ориентированного тонального словаря. Таким образом, построена гибридная модель, опирающаяся на использование лексикона для подготовки данных и применяющая техники машинного обучения для нахождения коэффициентов регрессии.

Полученные результаты демонстрируют эффективность гибридного подхода к предсказанию пользовательских рейтингов медицинских учреж-

index	Тексты	Предсказанные значения (y_pred)	Истинные значения (y_test)
1345	Хорошая клиника, чисто, персонал в целом доброжелательный, крайне редко ухожу не довольной. Очень хороший подход к клиенту	1.005001677808914	1.0
385	Есть такое понятие, как медицинская тайна, не так ли? Впервые обратилась в отделение на Труда и искреннее удивилась комнате сбора анализов! Итак, комната номер 100, открытая для всех, которой все желающие, помимо прочего, пользуются, как туалетом. Именно туда необходимо отнести все свои анализы, подписать и оставить... На обозрение всех заходящих и выходящих! Это просто ужас и позор! Естественно никакой мед. сестры там нет, остаётся только надеяться, что ваше «добро» дойдёт до места назначения и деньги потрачены не зря.	0.23734368761719576	0.4
420	Клиника неплохая. Медицинский персонал квалифицированный. Цены правда не маленькие, но если нужно срочно получить результат анализов или консультацию врача, то вам туда.	0.8237496388810508	0.8
940	Очень редко пишу отзывы! Центр достаточно не плохой, но есть свои минусы. Регистратура- обслуживают долго, на звонки не отвечают! Не предупреждают врачей о задержке клиента! Вот к примеру сегодня была записана на УЗИ, во первых 20 мин не могла дозвониться, предупредить, что могу опоздать, потом дозвонилась, предупредила, опоздала на 4 мин, и на моё место приняли другого клиента, пришлось ждать 30 мин. А могла и не торопиться, и мчаться через весь город!!!! Не хочу наблюдаться в этом центре по беременности!!!! Был опыт в другом центре, где намного лучше обслуживание!	0.46178861741151395	0.4
623	Не сильно отличается от государственной поликлиники, отношение персонала оставляет желать лучшего. А вот специалист, которая делает рентген, ей спасибо и за отношение и за её работу.	0.6427691281462824	0.6
1018	Хорошее место 👍Всем рекомендую.	1.0047301496158192	1.0
356	В кол.центре за справку для бега сказали нужно отдать 750руб. + Для справки нужно ЭКГ 700руб. Итого:1450руб. По приезду на ресепшн мне сказали к этой справке ещё нужно заплатить за прием к терапевту 1300 руб.... И цена стала: 2750 руб. В 2 раза просто подняли ценник. Развернулся и ушел. Больше не ногой сюда.	0.2808081066767455	0.2
272	В лотосе мне лишь понравилась девушка,которая ставила укол, Екатерина, она самая вежливая. По номеру телефона отвечают холодно и коротко....."будете записываться" тоном, будто все уже надоело.... В детском отделении посещала Лора Сушарину., не сказать,что доктор великолепный, как мне описывали по телефону, по мне - самый обычный доктор, как и везде. Цены дорогие, за повторный приём не делают скидок. По телефону стоимость говорят сбивчиво, путаясь.....	0.5222663033192672	0.4
174	Очень дорого. Однажды долгое время принимался с травмой ноги. Обошел 5 травматологов-ортопедов, из них лишь один специалист оказался с прогрессивным мышлением и тот был молодой парень лет 26-28! Остальные же врачи максимально флегматичные товарищи и мало чем отличаются от муниципальных коллег. Приятен факт хорошего оборудования, но при таких доходах, как у "Лотоса" это должно быть нормой, поэтому не прибавляет баллов к оценке. В целом, все мои друзья и знакомые оценивают средне.	0.6944778454990733	0.6
902	Просто кошмар, в центр не возможно дозвониться. Оставляла сообщение на сайте, типо "мы вам перезвоним"	0.15041484949809625	0.2

Рис. 4. Предсказания модели линейной регрессии на тестовой выборке

дений на основе регрессионного анализа тональности текстов. Разработанная модель множественной линейной регрессии показала высокую точность предсказаний, что подтверждается низкими значениями среднеквадратической и средней абсолютной ошибок, а также высоким коэффициентом детерминации. Эти результаты подчеркивают потенциал использования анализа тональности для автоматизации извлечения мнений пользователей на основе текстовых данных. Однако для внедрения модели необходима разработка автоматизированных методов определения тональных меток в текстах, что позволит использовать построенную модель в практических приложениях для выполнения надежных предсказаний удовлетворенности клиента. Для решения задачи поиска в тексте тонально-маркированных единиц в будущем мы полагаем целесообразным рассмотреть применение предметно-ориентированных тональных словарей в сочетании с правилами разрешения неоднозначности, а также интеграцию методов машинного обучения.

Список литературы / References

1. Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. СПб.: Питер, 2022. 480 с. [Nikolenko S., Kadurin A., Arkhangelskaya E. *Glubokoe obuchenie* [Deep Learning]. Saint-Petersburg: Piter, 2022. 480 p.]

2. Самигулин Т.Р., Джурабаев А.Э.У. Анализ тональности текста методами машинного обучения // Научный результат. Информационные технологии. 2021. Т. 6, № 1. С. 55–62. [Samigulin T.R. Djurabaev A.E.U. [Sentiment Analysis of Text by Machine Learning Methods]. *Research Result. Information Technologies*. 2021, vol. 6, no. 1, pp. 55–62. (in Russ.)] DOI: 10.18413/2518-1092-2021-6-1-0-7.

3. Шереметьева С.О., Бабина О.И. Платформа для концептуального аннотирования многоязычных текстов. // Вестник Южно-Уральского государственного университета. Сер. Лингвистика. 2020. Т. 17, № 4. С. 53–60. [Sheremetyeva S.O., Babina O.I. [A platform for knowledge assisted conceptual annotation of multilingual texts]. *Bulletin of South Ural State University. Series Linguistics*. 2020, vol. 17, no. 4, pp. 53–60. (in Russ.)]

4. Baccianella S., Esuli A., Sebastiani F. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining . Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC'10, Valetta, Malta, 17–23 May 2010. Valettta: LREC, 2010. P. 2200–2204.

5. Burns J.C., Kelsey T. Comparison of Commercial Decoder-only Large Language Models for Multilingual Sentiment Analysis of Short Text // Research Square. 31 August 2024, preprint. URL: https://www.researchsquare.com/article/rs-4849789/v1 (accessed on 12 Oct 2024). DOI: 10.21203/rs.3.rs-4849789/v1.

6. Du K., Xing F., Mao R., Cambria E. Financial Sentiment Analysis: Techniques and Applications. *ACM Computing Surveys*. 2024, vol. 56, iss. 9, art. no. 220. DOI: 10.1145/3649451.

7. El-Affendi M.A., Alrajhi K., Hussain A. A Novel Deep Learning-Based Multilevel Parallel Attention Neural (MPAN) Model for Multidomain Arabic Sentiment Analysis. *IEEE Access*. 2021, vol. 9, pp. 7508–7518. DOI: 10.1109/ACCESS.2021.3049626.

8. Gooljar V., Issa T., Ramanan S.H., Abu-Salih B. Sentiment-based predictive models for online purchases in the era of marketing 5.0: a systematic review. *Journal of Big Data*. 2024, vol. 11, art. no. 107. DOI: 10.1186/s40537-024-00947-0.

9. Grljević O., Bošnjak Z., Kovačević A. Opinion mining in higher education: A corpus-based approach. *Enterrprise Information Systems*. 2020, vol. 16, iss. 5, art. no. 1773542. DOI: 10.1080/17517575.2020.1773542.

10. Gudankwar A. S., Mendhe P.M. Oghare L.N., Yemde A.R. Sentiments of Public Opinion. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*. 2024, vol. 10, no. 2, pp. 459–461. DOI: 10.32628/CSEIT2410239.

11. Jazuli A., Widowati, Kusumaningrum R. Auto Labeling to Increase Aspect-Based Sentiment Analysis Using K-Nearest Neighbors Method. *The 7th International Conference on Energy, Environment, Epidemiology and Information System (ICENIS 2022)*, Semarang, Indonesia, 9-10 August 2022 (E3S Web of Conference 359). EDP Sciences, 2022. Art. no. 05001. DOI: 10.1051/e3sconf/202235905001.

12. Kastrati Z., Ahmedi L., Kurti A., Kadriu F., Murtezaj D., Gashi F. A deep learning sentiment analyser for social media comments in low-resource languages. *Electronics*. 2021, vol. 10, iss. 10, art. no. 1133. DOI: 10.3390/electronics10101133.

13. Loukachevitch N., Levchik A. Creating a General Russian Sentiment Lexicon. *Proceedings of Language Resources and Evaluation Conference LREC-2016, Portorož, Slovenia, 23–28 May 2016*. ELRA, 2016. P. 1171–1176.

14. Loukili M., Fayçal Messaoudi, Mohammed El Ghazi. Machine Learning based Recommender System for E-Commerce. *International Journals of Artificial Intelligence*. 2023, vol. 12, no. 4, pp. 1803–1811. DOI: 10.11591/ijai.v12.i4.pp1803-1811.

15. Mohamed A., Zain Z.M., Shaiba H., Alturki N., Aldehim G. et al. LexDeep: hybrid lexicon and deep learning sentiment analysis using twitter for unemployment-related discussions during COVID-19. *Computers, Materials & Continua*. 2023, vol. 75, no. 1, pp. 1577–1601. DOI: 10.32604/cmc.2023.034746.

16. Rusnachenko N., Golubev A., Loukachevitch N. Lare Language Models in Targeted Sentiment Analysis. arXiv:2404.12342 [cs.CL], 2024. URL: https://arxiv.org/abs/2404.12342 (accessed on 12 Oct 2024).

17. Salur M.U., Aydin I. A novel hybrid deep learning model for sentiment classification. *IEEE Access*. 2020. Vol. 8. P. 58080-58093. DOI: 10.1109/ACCESS.2020.2982538.

18. Samsir, Irmayani D., Edi F., Harahap J.M., Jupriaman, Rangkuti R.K., Ulya B., Watrianthos R. Naives Bayes Algorithm for Twitter Sentiment Analysis. Virtual Conference on Engineering, Science and Technology (ViCEST'2020). (Journal of Physics: Conference Series 1933). *IOP Publishing*, 2021. Art. no. 012019. DOI: 10.1088/1742-6596/1933/1/012019.

19. Sari A.W., Hermanto T.I., Defriani M. Sentiment Analysis of Tourist Reviews Using K-Nearest Neighbors Algorithm And Support Vector Machine. *Sinkron: Jurnal dan Penelitian Teknik Informatika*. 2023, vol. 8, no. 3, pp. 1366–1378. DOI: 10.33395/sinkron.v8i3.12447.

20. Sheremetyeva S.O., Babina O.I. On automated creation of gold-standard corpus for multi-aspect sentiment analysis. *Proceedings of the international conference "Internet and Modern Society"*. Saint-Petersburg: ITMO, 2024. (*in print*)

21. Wijati D, Atika P.D., Setiawati S, Rasim. Sentiment Analysis of Application Reviews using the K-Nearest Neighbors (KNN) Algorithm. *Penelitian Ilmu Komputer, Sistem Embedded and Logic*. March 2024. Vol. 12, no. 1. P. 209–218. DOI: 10.33558/piksel.v12i1.9490.

22. Yan Ch., Liu J., Liu W., Liu X. Research on Public Opinion Sentiment Classification based on Attention Parallel Dual-Channel Deep Learning Hybrid Model. *Engineering Applications on Artificial Intelligence*. 2022, vol. 116, art. no. 105448. DOI: 10.1016/j.engappai.2022.105448.

23. Yavari A., Hassanpour H., Rahimpour Cami B., Mahdavi M. Election Prediction Based on Sentiment Analysis using Twitter Data. *International Journal of Engineering*. 2022, vol. 35, no. 2, pp. 372–379. DOI: 10.5829/ije.2022.35.02b.13.

24. Zainuddin N., Selamat A. Ibrahim R. Hybrid sentiment classification on twitter aspect-based sentiment analysis. *Applied Intell*igence. 2018, vol. 48, pp. 1218–1232. DOI: 10.1007/s10489-017-1098-6.

Информация об авторе

Бабина Ольга Ивановна, кандидат филологических наук, доцент, заведующий кафедрой лингвистики и перевода, Южно-Уральский государственный университет, Челябинск, Россия, babinaoi@susu.ru

Information about author

Olga I. Babina, Candidate of Science (Philology), Associate Professor, Head of the Department of Linguistics and Translation, South Ural State University, Chelyabinsk, Russia, babinaoi@susu.ru

Статья поступила в редакцию 16.10.2024.
The article was submitted 16.10.2024.