Анализ тональности текстов на основе словарей: текущие и будущие тенденции по данным PubMed

Sentiment Analysis Based on Dictionaries: Current and Future Trends from PubMed collection

Шарнин М.М.
Канд. техн. наук, старший научный сотрудник, ФГУ «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва
e-mail: mc@keywen.com

Charnine M.M.
Candidate of Technical Sciences, Senior Researcher, Federal Research Center Informatics and Control of the Russian Academy of Sciences, Moscow
e-mail: mc@keywen.com

Калинин С.С.
Канд. филол. наук, преподаватель кафедры Иностранные языки, лингвистика и перевод, ФГАОУ ВО «Пермский национальный исследовательский политехнический университет», г. Пермь
e-mail: rage_of_gods@inbox.ru

Kalinin S.S.
Candidate of Philological Sciences, Lecturer, Department of Foreign Languages, Linguistics and Translation, Perm National Research Polytechnic University, Perm
e-mail: rage_of_gods@inbox.ru

Аннотация
В работе представлено прогнозное библиометрическое исследование трендовых тем в коллекции «PubMed» в области анализа тональности текстов на основе словарей. Исследование выполнено с использованием коллекции научных статей, индексирующихся в библиографической базе данных «PubMed», из которой были отобраны 147 статей, имеющие в заголовках и аннотациях ключевые слова «sentiment analysis» (анализ тональности) и «dictionary» (словарь). Выявлен значительный рост (в 21 раз за 6 лет) ежегодно публикуемых подобных статей. Рассчитан и представлен рейтинг релевантных ключевых слов в отобранных статьях. Среди релевантных ключевых слов выявлены трендовые ключевые слова с прогнозируемым долгосрочным ростом трендов. Представлена семантическая карта трендовых ключевых слов, содержащая информацию о новизне и долгосрочности трендов. В результате визуального анализа семантической карты выявлены две трендовые темы: (1) вопросы благополучия, (2) удовлетворение пациентов и заказчиков. К вопросам благополучия относятся трендовые подтемы: одиночество, депрессия, устойчивость к подобным проблемам, стратегии их преодоления.
Ключевые слова: обработка естественного языка, формальные языковые модели, библиометрический анализ, анализ тональности, семантическая карта, долгосрочный прогноз развития научных трендов, библиографическая база данных PUBMED.

Abstract
The paper presents a predictive bibliometric study of trending topics in the PubMed collection in the field of sentiment analysis of texts based on dictionaries. The design of the present study uses a collection of scientific articles indexed in the PubMed bibliographic database. 147 articles that

had the keywords “sentiment analysis” and “dictionary” in the titles and abstracts were retrieved from the mentioned base. A significant increase (21 times over 6 years) in the number of annually published similar articles was revealed. The rating of relevant keywords in the selected articles was calculated and presented. Among the relevant keywords, trending keywords with predicted long-term trend growth were identified. A semantic map of trending keywords is drawn, containing information on the novelty and longevity of trends. As a result of visual analysis of the semantic map, two trending topics were determined: (1) well-being issues, (2) patient and customer satisfaction. Well-being issues include trending subtopics: loneliness, depression, resilience to such problems, and strategies of their coping.
Keywords: natural language processing, formal linguistic models, bibliometrical analysis, sentiment analysis, semantic map, long-term prediction of transformation of scientific trending topics, PubMed database collection.

Введение

В компьютерной лингвистике под анализом тональности или сентимент-анализом (**sentiment analysis**) понимают методы анализа текста, предназначенные для автоматизированного выявления в текстах эмоционально окрашенной лексики и эмоциональной оценки авторами тех или предметов либо явлений (либо выражения авторского мнения). Оценивание того, являются ли данные тексты позитивными или негативными, имеет широкое применение во многих областях, таких как социология, маркетинг и реклама, психология, экономика и политология. Автоматизированные подходы позволяют быстро, воспроизводимо и с высокой точностью определять тональность у практически неограниченного количества текстов.

Различают методы анализа тональности, основанные на лексике и словарях, на машинном обучении и больших языковых моделях, а также гибридные методы [1]. Методы, основанные на правилах и словарях, осуществляют поиск эмотивной лексики (репрезентирующей лексическую тональность) в тексте по заранее составленным тональным словарям и правилам с применением лингвистического анализа. По совокупности найденной эмотивной лексики текст может быть оценен по шкале, содержащей количество негативно окрашенных и позитивно окрашенных лексических единиц.

Одним из наиболее популярных и широко используемых тональных словарей является VADER (Valence Aware Dictionary and sEntiment Reasoner) [2]. VADER – это инструмент анализа тональности на основе лексикона и грамматических правил, который специально ориентирован на анализ настроений, выражаемых в социальных сетях. VADER также хорошо работает с текстами из других областей. Словарь VADER содержит более 7500 лексем, для которых указана как полярность настроения (положительное/отрицательное), так и интенсивность выраженности настроения по шкале от –4 до +4.

В работе Л.М. Ганди и соавторов сравнивались популярные инструменты анализа настроений, основанные на обработке естественного языка (VADER, TEXT2DATA и Linguistic Inquiry and Word Count), а также на использовании большой языковой модели (а именно – ChatGPT4.0) [3]. В результате авторы рекомендовали VADER, как единственный бесплатный инструмент, который они оценили, из-за его превосходного качества (дискриминации), допускающее также и дополнительное улучшение, если длина комментариев составляет не менее 100 символов.

VADER приспособлен для анализа англоязычных текстов. В работе И. Мохаммеда и Р. Прасада предложена аналогичная методика для анализа других языков [14]. Отмечается, что анализ тональности с помощью методов обработки естественного языка потенциально может обладать более высоким качеством чем методы на основе словарей/лексики.

Но применение таких методов для каждого языка требуют набора лингвоспецифичных данных: только около 20 языков имеют текстовые корпусы, достаточные для использования методов обработки естественного языка.

Результаты показывают, что предложенная модель на основе словарей/лексики является перспективным инструментом для анализа настроений в различных приложениях для языков с низкими лингвистическими ресурсами.

Методы анализа тональности текстов на основе словарей стремительно развиваются. За последние 6 лет количество ежегодно публикуемых статей в этой области выросло в 21 раз. Однако до сих пор выявление трендов в данной предметной области выполняется на основе экспертных оценок, причем не приводятся данные о долгосрочности и точности прогноза для выявленных будущих трендов. Настоящая работа призвана исправить этот недостаток.

В ней авторы стремились к достижению трех основных целей исследования:

1) из имеющихся в литературе работ по анализу тональности текстов на основе словарей выявить трендовые ключевые слова, имеющие наиболее долговременные растущие тренды по количеству статей и цитирований;
2) выявить и визуализировать трендовые темы (тенденции) из близких по семантике трендовых ключевых слов;
3) с помощью предлагаемого в работе метода выявить наиболее перспективные трендовые темы и получить для них оценки долгосрочности и точности прогноза.

Обзор существующих исследований

В научной литературе имеется ряд работ с анализом трендов в области анализа тональности текстов на основе словарей. Методы анализа тональности чаще всего применяются для анализа текстов социальных сетей, таких как «Twitter», «Reddit» и «Weibo», поскольку социальные сети содержат огромное количество эмоционально окрашенных текстов, касающихся самых разных явлений из области медицины, психологии, социологии, маркетинга, рекламы, экономики и политологии. Подобные явления, описанные в коллекции научных статей «PubMed», обычно касаются вопросов благополучия (well-being), одиночества (loneliness), депрессии (depression), устойчивости (resilience) к подобным проблемам, стратегии их преодоления (coping strategy), вовлеченности (engagement) в эти вопросы, степень удовлетворенности пациентов и заказчиков (patients' satisfaction, customer satisfaction) оказываемыми им медицинскими услугами. Приведем примеры статей, изучающих подобные явления на материале социальных сетей.

В статье Х. Шаха и М. Хусех проводится сравнительный анализ одиночества в США и Индии с использованием данных из социальной сети «Twitter» и анализа тональности с помощью словаря VADER [4]. Авторы отмечали, что одиночество, как широко распространенная проблема общественного здравоохранения во всем мире, имеет далеко идущие последствия для психического и физического благополучия, а также экономической производительности труда.

В исследовании изучались различия в динамике одиночества в разных городах, выявлялись географические различия в коррелируемых темах. Твиты с негативной окраской были дополнительно проанализированы на предмет психосоциальных лингвистических особенностей для поиска значимых корреляций между одиночеством и социально-экономическими и эмоциональными темами и факторами.

В работе Д. Вальдеса и соавторов анализируются тенденции эмоционального благополучия населения по данным из «Twitter» с помощью моделирования тем скрытого распределения Дирихле и тонального анализа по словарю VADER [5]. Авторы использовали эмоционально обогащенные тексты в социальных сетях для создания «Словаря общественного мнения». Затем, объединив этот словарь с векторными моделями представления слов и анализом тенденций поиска, они разработали композитный индекс тревожности и депрессии, который может отражать уровень психического здоровья региона в течение определенного периода времени.

В исследовании Ю. Хсвен и соавторов данные из анонимного веб-сообщества «Inspire», посвященного вопросам здоровья, были обработаны с помощью моделирования

тем латентного распределения Дирихле и тонального анализа по словарю VADER [6]. Анализировались сообщения больных с интерстициальным циститом/болевым синдромом. Эти симптомы могут существенно повлиять на качество жизни людей, влияя на их психическое, физическое, сексуальное и финансовое благополучие.

Целью анализа было выявить мнение и настроение лиц, страдающих вышеупомянутыми заболеваниями, по отношению к эффективности специальных лекарств. Результаты показали, что поставщики медицинских услуг могут извлечь пользу из рассмотрения идей, которыми делятся на форумах, чтобы лучше понять индивидуальные предпочтения, проблемы и ожидания. С.А. Рао с соавторами проанализировали текстовые комментарии, оставленные пользователями социальной сети «Reddit», с помощью словаря VADER [7]. Целью анализа было изучение реакции на отмену плановых хирургических операций из-за пандемии COVID-19. Авторами было выявлено, что доступность плановой хирургии оказывается также важной и для эмоционально-психического здоровья пациентов.

А Одентан вместе с соавторами в исследовании «I Let Depression and Anxiety Drown Me…» обработали информация из приложения «Q-Life», которое направлено на повышение устойчивости людей к стрессу и неблагоприятным жизненным событиям с помощью различных механизмов преодоления трудностей, включая ведение журнала [8]. Использовались методы тематического анализа и тонального анализа, включая два метода на основе лексикона и восемь методов машинного обучения для классификации записей журнала на положительную или отрицательную полярность настроений.

Был проведен тематический анализ отрицательных и положительных записей журнала, чтобы определить темы, представляющие факторы, связанные с устойчивостью либо отрицательно, либо положительно, и определить различные механизмы преодоления трудностей. Результаты выявили 14 отрицательных тем, таких как стресс, беспокойство, одиночество, отсутствие мотивации, болезнь, проблемы в отношениях, депрессия и тревога. Также было выявлено 13 положительных тем, включая самоэффективность, благодарность, социализацию, прогресс, релаксацию и физическую активность. Определены 7 механизмов преодоления трудностей, включая управление временем, качественный сон и психологическую внимательность.

В работе В. Вогт и соавторов для оценки качества рефракционной хирургии анализировались письменные отзывы с «Healthgrades» (популярного сайта, на котором приводятся рейтинги врачей и отзывы пациентов о них) [9]. Отзывы были разделены по региону и стажу работы. Использовался анализ частотности слов и анализ настроений с применением словаря VADER. Анализ частотности лексических единиц показал, что пациенты ценят неклинические аспекты лечения, включая взаимодействие с персоналом, страховое покрытие их медицинских расходов и время ожидания, что позволяет предположить, что улучшение неклинических факторов может повысить степень удовлетворенности пациентов рефракционно-хирургическим лечением.

В работе тайских исследователей на основании отзывов с сайта «TripAdvisor» был проведен анализ настроений и удовлетворенности клиентов ресторанов во время пандемии COVID-19 в Паттайе (Таиланд) [10]. Анализ тональности/настроений проводился с помощью словаря VADER. Были выявлены две проблемные области, а именно обслуживание и персонал, а также еда и ее вкус, которые требуют срочного вмешательства. Результаты этого исследования предлагают ценную информацию о поведении и потребностях клиентов, тем самым давая возможность ресторанному бизнесу улучшить качество обслуживания.

Из приведенного обзора можно сделать следующие выводы. В рассмотренных статьях анализ тональности часто используется вместе с анализом тем и анализом частоты слов и фраз. Для тематического анализа чаще всего используется латентное распределение Дирихле. Анализируемые данные делятся на части как с помощью известных характеристик (регион, стаж работы и т.д.), так и по скрытым темам, выявленным с помощью тематического анализа. Тональный анализ скрытых тем позволяет выявить положительные и отрицательные факторы, а также определить различные механизмы преодоления трудностей.

Анализируемые в приведенных статьях вопросы (благополучие, одиночество, депрессия и т.д.) имеют различную динамику и прогнозируемые тренды в области анализа тональности текстов на основе словарей. Данная статья предоставляет информацию о долгосрочности прогнозируемых будущих трендов и точности этих прогнозов.

Постановка задачи. Исходный текстовый материал

Предметная область, исследуемая в нашей работе, – это публикации, касающиеся анализа тональности текстов на основе словарей. Настоящая статья посвящена прогнозу и визуализации трендовых тем в исследуемой области. Если говорить более точно, то нами ставится следующая задача: дать прогноз долгосрочности роста трендов с известной точностью на 3 и более лет вперед для трендовых тем в области исследования на базе коллекции PubMed, а также предложить средства визуализации прогноза трендовых тем.

Для решения поставленной задачи была проанализирована известная текстовая коллекция по медицинской тематике «PubMed», которая по состоянию на начало 2025 г. содержала более 38 миллионов статей по медицине, биологии и связанным наукам. Из этой коллекции было выделено 147 статей, содержащие в заголовках и аннотациях слова sentiment analysis (анализ тональности) и dictionary (словарь). Данные этих 147 статей мы называем в дальнейшем Локальной коллекцией.

Таким образом, в рамках данной работы мы работаем с двумя коллекциями: с коллекцией «PubMed», содержащей более 38 миллионов записей/статей, и с Локальной коллекцией, содержащей 147 статей. Локальная коллекция является частью коллекции «PubMed», причем частью, наиболее сильно связанной с исследуемой областью. Свойства ключевого слова из Локальной коллекции (частота, вероятность, тренды, контекст и т.д.) в дальнейшем называются локальными, а соответствующие свойства этого же ключевого слова в рамках коллекции «PubMed» называются глобальными. Как локальные, так и глобальные свойства необходимы для выявления характерных/релевантных и трендовых ключевых слов.

По данным коллекции «PubMed» был рассчитан следующий график роста по годам: количества статей в сфере анализа тональности текстов на основе словарей (см. рис.1)

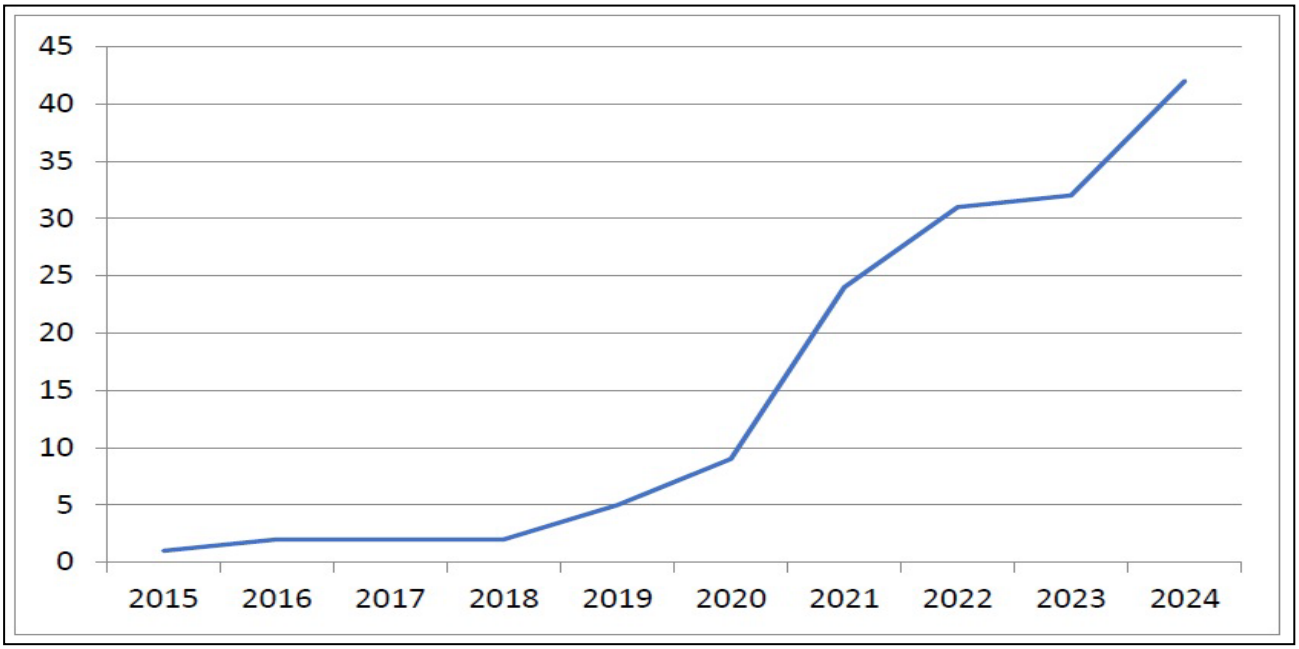

Рис. 1. График количества научных статей в сфере анализа тональности текстов на основе словарей за различные годы по данным «PubMed».
Запрос: (sentiment analysis) AND (dictionary)

Из рис. 1 видно, что наблюдается резкий рост количества ежегодно публикуемых научных статей, индексируемых в «PubMed», в области анализа тональности текстов на основе словарей. Точнее, с 2018 по 2024 количество статей в этой области выросло в 21 раз и в настоящее время их более 147.

Этапы анализа

Для выявления и визуализации трендовых тем используется библиометрический анализ, в котором находят применение ряд математических и статистических методов к изучению библиографической коллекции «PubMed». Библиометрический анализ исследуемой области содержит 4этапа:

1) Определение поисковых запросов, с помощью которых можно найти статьи из PubMed, относящиеся к исследуемой области.
2) Построение Локальной коллекции, в которую включены заголовки статей из PubMed, удовлетворяющих поисковым запросам, и анализ динамики их количества за последние несколько лет.
3) Выявление в полученном материале релевантных, а затем и трендовых ключевых слов с упорядочиванием их по степени перспективности.
4) Построение семантической карты и выявление трендовых тем (тенденций) с помощью методов и программных средств визуальной аналитики.

Трендовые ключевые слова определяются с учетом долгосрочных прогнозов их трендов в «PubMed», рассчитанных с помощью пакета машинного обучения CatBoost [11]. Авторская методика долгосрочного прогноза трендов описана в работах [12; 13]. Точность прогноза на 3 года вперед составляет более 60%. При этом 59% трендовых слов, выявленных в 2020 г., остались трендовыми в 2023 г.

Результаты прогноза трендов визуализируются с помощью нейросети Word2Vec и алгоритма t-SNE на семантической карте. С помощью визуальной аналитики выявляются трендовые темы, входящие в кластеры трендовых слов на семантической карте. Трендовая тема состоит из одного или нескольких близких по смыслу трендовых ключевых слов. Выявленные будущие тренды сравниваются с уже опубликованными в научной литературе.

Результаты библиометрического анализа. Рейтинг релевантных ключевых слов

Характерными/релевантными в исследуемой области являются те ключевые слова, у которых локальная вероятность вхождения в заголовки статей выше, чем аналогичная глобальная вероятность. Локальная вероятность определяется по Локальной коллекции, а глобальная – по коллекции «PubMed».

Верхние позиции в рейтинге релевантных ключевых слов занимают следующие ключевые слова: sentiment analysis (анализ настроений), lexicon (лексикон), Twitter, Reddit, sentiment classification (классификация настроений), public sentiment (общественные настроения), social media (социальные сети), Dirichlet allocation (распределение Дирихле), processing NLP (обработка естественного языка), topic (тема), Weibo и т.д.

Данный рейтинг представляет наиболее релевантные ключевые слова из 3527 ключевых слов без предлогов, упоминающихся в Локальной коллекции три или более раз.

Рейтинг трендовых ключевых слов

Трендовые ключевые слова были отобраны среди релевантных ключевых слов с учетом их локальных параметров в Локальной коллекции и глобальных параметров в коллекции «PubMed». Следует заметить, что при составлении рейтинга анализируются показатели/вероятности для слов, пар слов, и троек из соседних слов без предлогов.

С помощью алгоритма машинного обучения CatBoost по методике, описанной в предыдущем разделе, были рассчитаны прогнозы долгосрочности роста трендов для релевантных слов и выявлены трендовые ключевых слова. Отобранные ключевые слова были упорядочены по совокупности показателей и в результате получился рейтинг

трендовых ключевых слов. Рейтинг составлялся с учетом следующих показателей: положение в рейтинге характерных слов, данные прогноза, частота в коллекции «PubMed» и в Локальной коллекции, а также по тенденциям роста в Локальной коллекции.

Верхние позиции в рейтинге трендовых ключевых слов занимают следующие термины (**на рис. 2 они выделены красным цветом**): unexplored (неисследованный), indicates (указывает), satisfaction (удовлетворение), differentiate (дифференцировать), learning methods (методы обучения), demonstrates (демонстрирует), well-being (благополучие), algorithm (алгоритм), sentiment analysis (анализ тональности), highlighting (выделение), coping (преодоление), evaluate (оценивать), pre-trained (предварительно обученный), machine-learning (машинное обучение).

Вторую группу терминов составляют те, которые выделены **на рис. 2 выделены синим цветом**: suggests (предполагает), promoted (продвигаемый), evolving (развивающийся), future research (будущие исследования), autoencoder (автоэнкодер), challenging (бросающий вызов/проблематичный), contributes (вносит вклад), identify (идентифицирует), learning model (модель обучения), machine learning (машинное обучение), engagement (вовлеченность), dataset (набор данных), depressive symptoms (симптомы депрессии), effectively (эффективно), resilience (устойчивость), social media (социальные медиа), loneliness (одиночество), natural language (естественный язык), learning models (модели обучения), deep learning (глубокое обучение), fuzzy (нечеткий), twitter («Twitter») и т.д. У лексем из первой группы (от unexplored до machine-learning) прогноз роста в «PubMed» составляет более 3 лет (т.е. после 2027 г.).

Построение семантической карты

Для целей визуального анализа с помощью нейросети Word2Vec была рассчитана мера семантического подобия (**semantic similarity**) характерных и трендовых ключевых слов. По этим данным с помощью алгоритма t-SNE построена семантическая карта (см. рис. 2). Чем выше мера подобия – тем меньше расстояние между ключевыми словами на семантической карте.

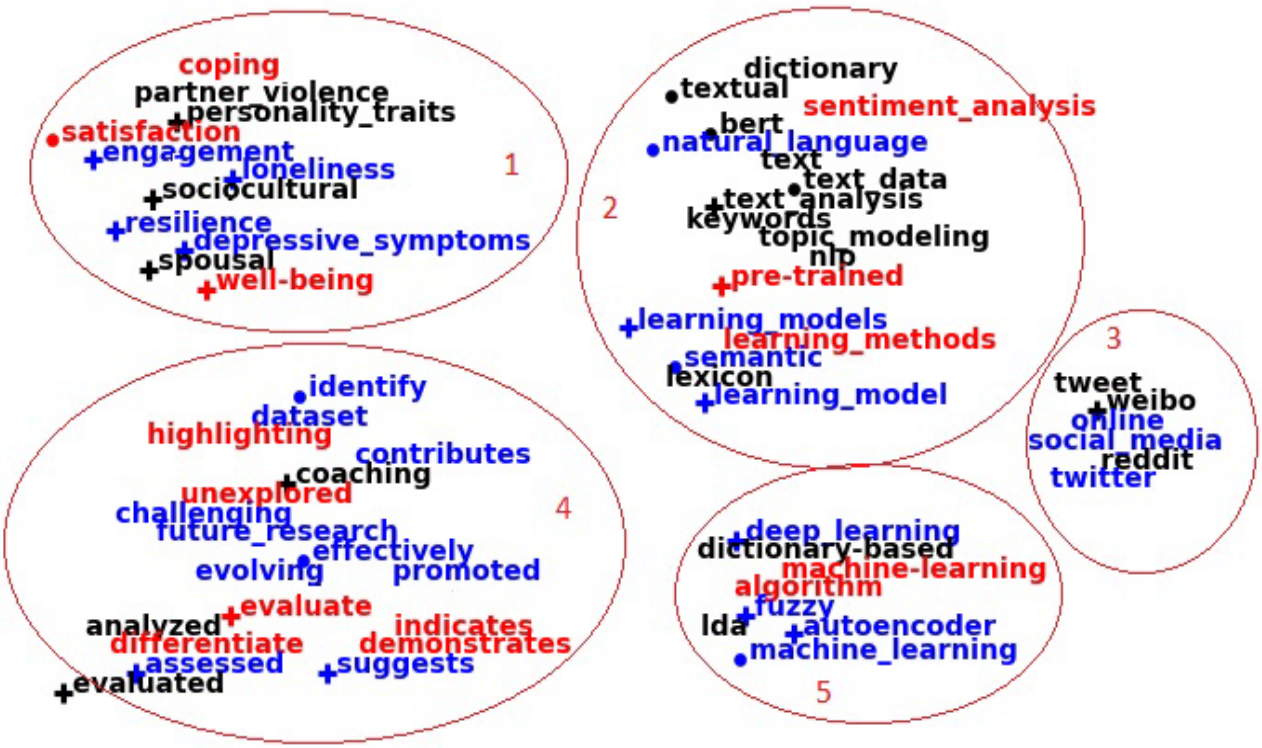

Рис. 2. Семантическая карта характерных и трендовых ключевых слов в области анализа тональности текстов на основе словарей

На рис. 2 представлены наиболее характерные и трендовые термины в области анализа тональности текстов на основе словарей. Красным цветом выделены наиболее перспективные термины, имеющие самые долгосрочные тренды в коллекции «PubMed», синим – среднесрочные тренды, черным – минимальные тренды. Плюсом (+) выделены ключевые слова, имеющие самые новые тренды, а точкой – средние по времени новизны тренды.

На рис. 2 также можно видеть пять кластеров, которые содержат близкие по семантике термины. Кластер 1 содержит термины, определяющие области применения исследуемых методов анализа тональности. В кластере 2 содержатся термины, соответствующие лингвистическим моделям и ресурсам. Здесь наиболее перспективными для тонального анализа являются предварительно обученные большие языковые модели.

В настоящее время более популярными являются более простые модели, основанные на словарях/лексиконе. Кластер 3 содержит названия социальных сетей («Twitter», «Reddit» и «Weibo»), к данным из которых чаще всего применяется тональный анализ. Кластер 4 содержит глаголы (evaluate, indicates и т.д.), которые часто встречаются в анализируемой предметной области и указывают на возможность точного измерения настроений авторов и связанных с ними параметров и характеристик. В кластере 5 содержатся термины, определяющие различные методы анализа тональности и тематического анализа.

Наибольший интерес представляет первый кластер с трендовыми терминами, определяющими области применения исследуемых методов анализа тональности. Красным цветом здесь выделены термины well-being (благополучие), satisfaction (удовлетворение), coping (преодоление), которые имеют наибольшие прогнозы долгосрочного роста трендов. Синим цветом в первом кластере выделены термины +loneliness (одиночество), +partner (партнер), +resilience (устойчивость), +depressive symptoms (симптомы депрессии), +engagement (вовлеченность), имеющие средние тренды.

Эти трендовые термины входят в две относительно несвязанные трендовые темы: (1) вопросы благополучия, (2) степень удовлетворенности пациентов и заказчиков. К вопросам благополучия относятся трендовые подтемы: одиночество, депрессия, устойчивость к подобным проблемам, стратегии их преодоления.

Заключение

Таким образом, визуальный анализ семантической карты помогает выделить кластеры трендовых терминов, определить входящие в кластеры трендовые темы, оценить их динамику и увидеть картину в целом, включая трендовые направления из наиболее перспективных тем: в нашем материале к таковым можно отнести лексические единицы, связанные с тематикой психологического и психического благосостояния и благополучия, а также тематике, связанной с проблемами одиночества, депрессии и стратегиями преодоления этих состояний.

Еще одной важной трендовой темой является степень удовлетворенности пациентов предоставляемой им медицинской помощью. Это подтверждается как данными визуализации полученных нами результатов, так и расчетом тех трендов в определенных научных областях, которые, по нашим прогнозам, и представлениям, должны будут получить существенную популярность в ближайшие годы.

Кроме того, следует заметить, что с помощью описанных методов анализа нами было выделено 5 кластеров, в которых группируются семантически близкие лексические единицы и термины. Таким образом, нашу методику можно использовать и для выявления основных лексико-семантических и лексико-тематических групп в какой-либо предметной области.

Помимо прочего, при помощи выделения кластеров основных терминов (либо ключевых слов) в анализируемой предметной области, а также при помощи визуализации семантической карты можно выявлять отношения между этими лексическими единицами, а также между лексическими единицами из разных лексико-семантических

и лексико-тематических групп, что позволяет прояснить лингвистическую структуру субъязыка той или иной предметной области (в частности, то, что касается ее лексико-семантической структуры).

Литература

1. Qi Y. Sentiment analysis using Twitter data: a comparative application of lexicon-and machine-learning-based approach [Text] / Y. Qi, Z. Shabrina. – Social network analysis and mining. – 2023. – 13(1). – 31.
2. Hutto C. Vader: A parsimonious rule-based model for sentiment analysis of social media text [Text] / C. Hutto, E. Gilbert. – Proceedings of the international AAAI conference on web and social media. – 2014. – Vol. 8. – No. 1. – Pp. 216–225.
3. Gandy L.M. Public Health Discussions on Social Media: Evaluating Automated Sentiment Analysis Methods [Text] / L.M. Gandy [et al.]. – JMIR Formative Research. – 2025. – 9(1). – e57395.
4. Shah H.A. Mapping loneliness through comparative analysis of USA and India using social intelligence analysis [Text] / H.A. Shah, M. Househ. – BMC public health. – 2024. – 24(1). – 253.
5. Valdez D. Social media insights into US mental health during the COVID-19 pandemic: longitudinal analysis of Twitter data [Text] / D. Valdez [et al.]. – Journal of medical Internet research. – 2020. – 22(12). – e21418.
6. Hswen Y. Sentiments of Individuals with Interstitial Cystitis/Bladder Pain Syndrome Toward Pentosan Polysulfate Sodium: Infodemiology Study [Text] / Y. Hswen [et al.]. – JMIR Formative Research. – 2025. – 9(1). – e54209.
7. Rao S.A. Social media responses to elective surgery cancellations in the wake of COVID-19 [Text] / S. A. Rao [et al.]. – Annals of surgery. – 2020. – 272(3). – e246–e248.
8. Oduntan A. “I Let Depression and Anxiety Drown Me…”: Identifying Factors Associated With Resilience Based on Journaling Using Machine Learning and Thematic Analysis [Text] / A. Oduntan [et al.]. – IEEE journal of biomedical and health informatics. – 2022. – 26(7), 3397-3408.
9. Vought V. Application of sentiment and word frequency analysis of physician review sites to evaluate refractive surgery care [Text] / V. Vought [et al.]. – Advances in Ophthalmology Practice and Research. – 2024. – 4(2). – Pp. 78–83.
10. Pleerux N. Sentiment analysis of restaurant customer satisfaction during COVID-19 pandemic in Pattaya, Thailand [Text] / N. Pleerux, A. Nardkulpat. – Heliyon. – 2023. – Vol. 9. – Issue 11. – e22193.
11. Prokhorenkova L. CatBoost: unbiased boosting with categorical features [Text] / L. Prokhorenkova [et al.]. – Advances in neural information processing systems. – Vol. 32. – Montreal, QC, 2018. – P. 6638–6648.
12. Charnine M. Research trending topic prediction as cognitive enhancement [Text] / M. Charnine [et al.]. – 2021 international conference on cyberworlds (CW). – IEEE. – 2021. – Pp. 217–220. – DOI: 10.1109/CW52790.2021.00044.
13. Charnine M. Visualization of Research Trending Topic Prediction: Intelligent Method for Data Analysis [Text] / M. Charnine, A. Tishchenko, L. Kochiev. – Proceedings of the 31th International Conference on Computer Graphics and Vision. – 2021. – Vol. 2. – P. 1028–1037.
14. Mohammed I. Building lexicon-based sentiment analysis model for low-resource languages [Text] / I. Mohammed, R. Prasad. – MethodsX. – 2023. – 11. – 102460.