Научно-практический рецензируемый журнал
"Современные проблемы здравоохранения и медицинской статистики" 2024 г., № 1
Scientific journal "Current problems of health care and medical statistics" 2024 г., № 1
ISSN 2312-2935

УДК 378.147:614.253.5:801.7
DOI 10.24412/2312-2935-2024-1-589-607

ИСПОЛЬЗОВАНИЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ТЕКСТА ДЛЯ КЛАССИФИКАЦИИ СТАТЕЙ В СИСТЕМАТИЧЕСКОМ ОБЗОРЕ ПО МЕДИЦИНЕ

Д.Н. Бегун, Е.В. Гаврилова, Н.В. Мирзаева, О.В. Головко, Н.В. Заришняк

ФГБОУ ВО "Оренбургский государственный медицинский университет" Министерства здравоохранения Российской Федерации, Оренбург, Российская Федерация

Введение. Систематические обзоры являются относительно новой областью применения технологий анализа текста.
Цель исследования – использование методов интеллектуального анализа текста с помощью программы RapidMiner для предварительной обработки и отбора статей в систематическом обзоре.
Материалы и методы. Был осуществлен поиск публикаций по ключевым словам «nursing education» AND «teaching models» OR «teaching methods» в базах данных Medline через интерфейс системы PubMed NLM (www.pubmed.com) и ScienceDirect (sciencedirect.com). Критерии включения: статьи на английском языке опубликованные по теме обзора с 2014 г. по 2023 г. Парсер WebHarvy (компания SysNucleus) позволил собрать данные и импортировать их в M.Excel с выделением следующих столбцов – «Title», «Author», «Journal», «Year», «Output», «Abstract», всего 8451 публикаций. С помощью программы RapidMiner (компания Altair) была проведена предварительная обработка текста с последующей кластеризацией (метод K-Meams) и классификацией - метод тематического моделирования при помощи скрытого распределения Dirichlet (LDA).
Результаты. Было удалено 305 документа (11 – пропущенные значения, 295 - дубликаты) и набор составил – 8146 документов. Количество публикаций по теме обзора с 2014 по 2023 гг. выросло в 7 раз (с 219 до 1601 статей). Статьи были опубликованы в 1171 журнале, но большая часть из них в 9 журналах, общее количество авторов – 6884. Метод кластеризации позволил выделить 4 кластера публикаций, которые позволяют судить только об основной теме каждого кластера. С помощью метода тематического моделирования было выделено 10 тем статей, которые позволяют судить не только о темах статей, но и о типах статей. Так Topic_6 включает обзорные статьи по обучению сестринскому уходу в образовании.
Обсуждение. Самый важный этап классификации текста — выбор лучшего классификатора. Скрытое распределение Дирихле (LDA) — рассматривает каждый документ как смесь тем, а каждую тему — как смесь слов и выявляет скрытые темы в корпусе данных.
Выводы (заключение). В нашем случаем, применение тематического моделирования с помощью скрытого распределения Dirichlet (LDA) дало очень хорошие результаты, но может оказаться, что при другом наборе данных, эта модель будет неэффективна.

Ключевые слова: обзорная статья, интеллектуальный анализ текста, RapidMiner, кластеризация, тематическое моделирование

Научно-практический рецензируемый журнал
"Современные проблемы здравоохранения и медицинской статистики" 2024 г., № 1
Scientific journal "Current problems of health care and medical statistics" 2024 г., № 1
ISSN 2312-2935

USING TEXT MINING TO CLASSIFY ARTICLES IN A SYSTEMATIC REVIEW OF MEDICINE

D.N. Begun, E.V. Gavrilova, N.V. Mirzaeva, O.V. Golovko, N.V. Zarishnyak

Orenburg State Medical University (OrSMU), Orenburg, Russian Federation

Introduction. Systematic reviews are a relatively new application of text mining technologies.
The purpose of the study is to use text mining methods using the RapidMiner program to pre-process and select articles in a systematic review.
Materials and methods. A search for publications was carried out using the keywords “nursing education” AND “teaching models” OR “teaching methods” in the Medline databases through the PubMed NLM (www.pubmed.com) and ScienceDirect (sciencedirect.com) system interface. Inclusion criteria: articles in English published on the topic of review from 2014 to 2023. The WebHarvy parser (SysNucleus company) made it possible to collect data and import it into M.Excel with the following columns highlighted - “Title”, “Author”, “Journal” ", "Year", "Output", "Abstract", a total of 8451 publications. Using the RapidMiner program (Altair company), pre-processing of the text was carried out, followed by clustering (K-Meams method) and classification - topic modeling method using latent Dirichlet distribution (LDA).
Results. 305 documents were removed (11 were missing values, 295 were duplicates) and the set amounted to 8146 documents. Number of publications on the review topic from 2014 to 2023. increased 7 times (from 219 to 1601 articles). Articles were published in 1171 journals, but most of them in 9 journals, the total number of authors is 6884. The clustering method made it possible to identify 4 clusters of publications, which allow us to judge only the main topic of each cluster. Using the topic modeling method, 10 topics of articles were identified, which make it possible to judge not only the topics of the articles, but also the types of articles. Thus, Topic_6 includes review articles on nursing education in education.
Discussion. The most important step in text classification is choosing the best classifier. Latent Dirichlet Allocation (LDA) - Treats each document as a mixture of topics and each topic as a mixture of words and identifies hidden topics in the data corpus.
Conclusions (conclusion). In our case, applying topic modeling using Latent Dirichlet Allocation (LDA) gave very good results, but it may turn out that this model will not be effective on a different data set.

Keywords: review article, text mining, RapidMiner, clustering, topic modeling

Введение. Применение интеллектуального анализа текста на этапе отбора статей для систематических обзоров привлекательна, но не так много публикаций об анализе текста для систематических обзоров. Подавляющее большинство статей на эту тему публикуется учеными-компьютерщиками в журналах и материалах конференций в области медицинской информатики или искусственного интеллекта.

ISSN 2312-2935

Во-вторых, чтобы эти технологии получили широкое распространение, они должны быть доступны, точны и проверены IT-специалистами, специалистами по статистике. Систематические обзоры являются относительно новой областью применения технологий анализа текста. Некоторые предположения о технологиях интеллектуального анализа текста в других приложениях не выполняются при переносе в контекст обзора. Методы интеллектуального анализа текста и их эффективность использования на предварительном этапе отбора статей для обзора еще не систематизирована и окончательно не доказана [1].

Подготовка квалифицированных медицинских сестер имеет важное значение для практического здравоохранения, поскольку помогает улучшить качество обслуживания пациентов в медицинских учреждениях. В сестринском образовании в настоящее время основное внимание уделяется обучению, как основной технологии получения компетенций необходимых для практической работы в организациях здравоохранения. Новые квалификационные требования к медицинским сестрам требуют применения инновационных моделей и методов обучения [2].

Обзор литературы по данной теме — отличный способ синтеза результатов исследований, чтобы показать доказательства и выявить области, в которых необходимы дополнительные исследования, что является важнейшим компонентом создания теоретических основ и построения концептуальных моделей обучения. Систематический обзор моделей и методов обучения в сестринском деле поможет понять проблемы в обучении и наметить пути их решения.

Цель исследования – использование методов интеллектуального анализа текста с помощью программы RapidMiner для предварительной обработки и отбора статей в систематическом обзоре.

Материалы и методы. На первом этапе мы осуществили поиск публикаций по ключевым словам «nursing education» AND «teaching models» OR «teaching methods» в базах данных Medline через интерфейс системы PubMed NLM (www.pubmed. com) и ScienceDirect (англоязычная база научных статей по широкому спектру областей: от компьютерных технологий до психологии издательского дома Elsevier, sciencedirect.com).

Критериями включения являлись: статьи на английском языке, опубликованные по теме обзора с 2014 г. по 2023 г.

На втором этапе проводился сбор публикаций с помощью парсера WebHarvy (компания SysNucleus). Программа позволяет собирать данные с любого веб-сайта, обрабатывать вход в систему; поддерживает отправку форм, навигацию, нумерацию страниц, категории и ключевые слова. Все публикации были импортированы в M. Excel с выделением следующих столбцов – «Title», «Author», «Journal», «Year», «Output», «Abstract». Всего было собрано 8451 публикации.

На третьем этапе для проведения интеллектуального анализа текстовой информации мы применили программный пакет RapidMiner (компания Altair) как среду для глубинного анализа текста. В RapidMiner были импортированы модули для интеллектуального анализа текста – «Text Processing», «Operator Toolbox». Для статистической обработки набора данных был загружен модуль «Statistics Extension». Была проведена предварительная обработка текста с последующей кластеризацией (метод K-Meams) и классификацией - метод тематического моделирования при помощи скрытого распределения Dirichlet (LDA).

Результаты.

Статистическая обработка набора данных. Для статистической обработки данных мы также воспользовались программой RapidMiner и оператором «Statistics». Кроме того, мы использовали оператор «Remove Duplicates» для удаления дубликатов, так как наш набор документов был составлен на основе поиска в двух базах данных. Было удалено 305 документа (11 – пропущенные значения, 294 - дубликаты) и наш набор составил – 8146 документов.

Количество публикаций в зарубежных изданиях по теме «Модели и методы обучения в сестринском образовании» за 10 лет (2014-2023 гг.) выросло в 7 раз (рис.1), что говорит о их важности и значимости в обучении медицинских сестер.

Статьи были опубликованы в 1171 журнале, но большая часть из них в 9 журналах (таб.1).

Общее количество авторов – 6884, но при этом следует учитывать количество публикаций у одного автора разное (от 19 до 1) – самое большое количество публикаций у авторов Russell Yancei – 17 и Zenobia Chan – 9. В среднем на одного автора приходится – 1,2 публикации.

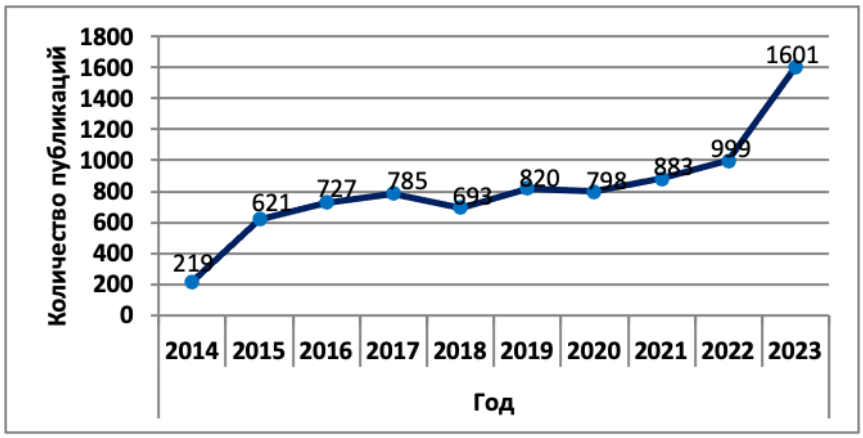

Рисунок 1. Количество публикаций за период с 2014 г. по 2023 г. в корпусе текстовых данных

Таблица 1

Журналы с наибольшим количеством публикаций по теме обзора

№	*Журнал*	*Страна*	*Количество публикаций*
1	Nurse Education Today	Великобритания	971
2	Journal of Nursing Education	США	682
3	Nurse Education in Practice	Великобритания	538
4	Nursing Education Perspectives (NEP)	Великобритания	264
5	Journal of Professional Nursing	США	195
6	BMC Medical Education	Великобритания	163
8	Journal of Interprofessional Care	США и Великобритания	115
9	International Journal of Environmental Research and Public Health	Швейцария	113

Предварительная обработка текста. В Rapidminer для чтения и загрузки данных из файлов Microsoft Excel имеется оператор – «Read Excel», который загружает данный файл в программу (рис.2). Для преобразования содержания файла в текстовые данные был использован оператор – «Nominal to Text». Оператор «Process Document from Data» - это вложенный оператор, он содержит подпроцесс, состоящий из множества операторов, которые соединены последовательно: «Tokenize», «Filter Stopwords (English)», «Filter Tokens

(by Length)», «Stem (Porter)», «Transform Cases», Generate n-Grams (Terms) (рис.2). И у каждого вложенного оператора имеются параметры, которые можно выбрать в зависимости от цели и задач исследования.

Оператор «Tokenize» разбивает последовательность строк на слова и убирает все знаки препинания (образуются токены); оператор «Filter Stopwords (English)» отфильтровывает стоп-слова - the, is,and, has, of, are и т. д.; «Filter Tokens (by Length)» - фильтрует токены по их длине; «Stem (Porter)» - этот оператор формирует английские слова с помощью алгоритма формирования корней Портера, применяя итеративную замену суффиксов слов на основе правил с целью уменьшения длины слов до достижения минимальной длины. Оператор «Transform Cases» - преобразует все символы документа в нижний или верхний регистр. «Generate n-Grams (Terms)» - создает терм n-грамм токенов в документе. N-граммы представляют собой представление на основе строк без какой-либо лингвистической обработки. Мы использовали фразеологический подход с образованием 2-х N-грамм, который основан на лингвистически сформированных фразах.

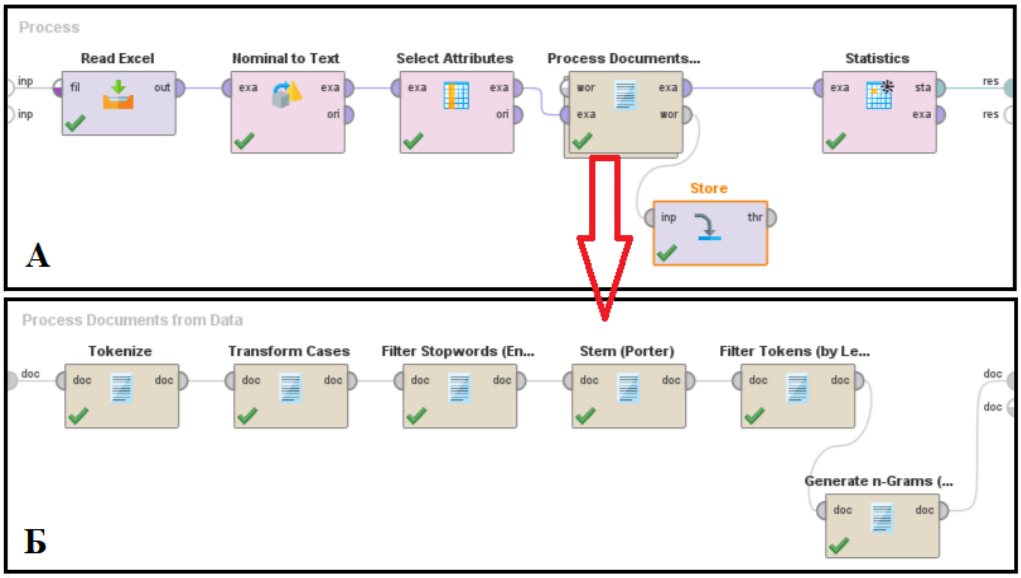

Рисунок 2. Предварительная обработка корпуса текстовых данных в программе Rapidminer. А – основной процесс в программе Rapidminer; Б – подпроцесс оператора «Process Document from Data»

Мы определили частоту встречаемости слов в каждом примере с помощью TF-IDE (term frequency-inverse document frequency, частота термина), метод взвешивания термина с которой конкретное слово (токен) встречается в коллекции документов (рис.3А). Были определены вектор слов и составлено «облако слов» (рис.3Б), ассоциация слов и схожесть (подобие) документов (рис. 3В).

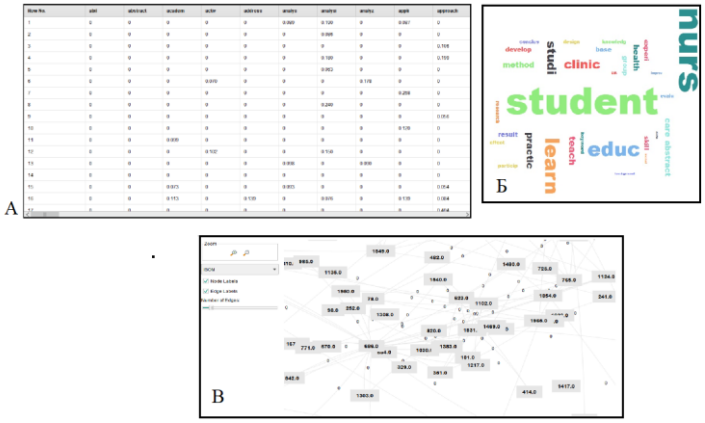

Рисунок 3. Интеллектуальный анализ текста. А – частота встречаемости термина в различных документах в наборе данных; Б – «облако слов»; В – графическое представление схожести (подобия) документов

Кластеризация и классификация документов. Для кластеризации документов на первом этапе мы применили метод K-Meams – алгоритм кластеризации, обладающий незначительной сложностью реализации, совместно с высокими показателями производительности. Основная идея данного алгоритма заключается в декомпозиции множества на k кластеров, обладающих индивидуальными центроидами, при этом центроид кластера формируется таким образом, чтобы обладать тесной взаимосвязанностью с точки зрения заданной меры сходства со всеми объектами, принадлежащими данному кластеру. Атрибут «Abstract» был выбран с помощью оператора «Select Attribut». Оператор «K-Meams» - были установлены параметры «NumericalMeasures» с «CosineSimilarity» и k=4 Добавлен оператор «Cluster Model Visualizer».

Было выделено 4 кластера, построена таблица центроидов токенов, произведена визуализация кластеров (рис.4). На рисунке 4, мы видим количество документов в каждом кластере, основные токены и частоту их встречаемости. Основной темой Cluster_0 являются

– исследование, результат, заключение; cluster_1 - учиться, учить, медсестра; cluster_2 - медсестры, воспитывать, учиться; cluster_3 - учить, учиться, исследовать.

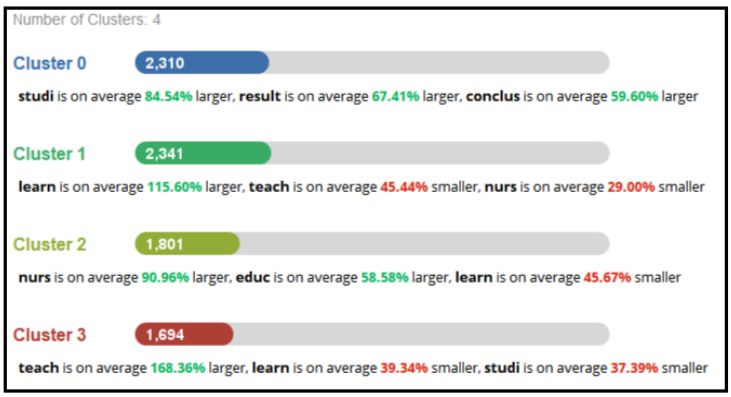

Рисунок 4. Визуализация кластеров (K-Meams)

Полученные данные не позволяют сделать вывод, какие типы статей (обзорные, оригинальные исследования и т.д) входят в каждый кластер, мы можем только судить только об основной теме.

На следующем этапе было проведено тематическое моделирование при помощи скрытого распределения Dirichlet (LDA). Это метод машинного обучения без учителя, который позволяет выявлять скрытые темы в наборе (корпусе) документов. Для реализации LDA мы использовали оператор RapidMiner «Extract Topics from Data». Реализация LDA использует ParallelTopicModel библиотеки Mallet и выборку Гиббса для применения модели [3]. Были установлены параметры: «iteration» (количество повторений) -1000, «number of topics» (количество тем) – «true», т.е. количество тем определяется автоматически.

При таком параметре «number of topics» (количество тем) – «true»: α (Dir(α) - распределение Дирихле, используемая эвристика: 50/количество тем) и β (Dir(β) - двухпараметрическое семейство абсолютно непрерывных распределений, используемая эвристика: 50/Количество слов) определяются автоматически и были оптимизированы при каждой итерации. Параметр Dir(α) еще называют параметром концентрации, который определяет тенденцию распределения - равномерную (α=1), концентрированную (α>1) или разреженную (α < 1) [4].

Было выделено 10 тем с основными темами (таб.2): Topic_0 – «Симуляционное обучение студентов медсестер клиническим навыкам»; Topic_1 – «Обучение студентов уходу за пациентами с COVID»; Topic_2 – «Уход за пациентами для сохранения психического здоровья»; Topic_3 – «Обучение студентов на онлайн курсах»; Topic_4 – «Межпрофессиональное обучение студентов»; Topic_5 – «Клиническое обучение студентов»; Topic_6 включает обзорные статьи по обучению сестринскому уходу в образовании; Topic_7 – «Групповое обучение студентов медсестер»; Topic_8 – «Значение образования медсестер в здравоохранении»; Topic_9 - «Преподавание в обучении сестринскому уходу». Выделены не только темы корпуса данных, но и типы статей. Вектор производительности процесса представлен в таблице 3.

Таблица 2

Результаты тематического моделирования при помощи скрытого распределения Дирихле (LDA)

Index	*Nominal value*	*Absolute count*	*Fraction*	*Basic terms*
5	Topic_0	839	0.102	Simulation, students, clinical, nursing, patient
7	*Topic_1*	*578*	*0.070*	Students, nursing, study, COVID, learning
10	Topic_2	353	0.043	Care, health, patients, mental, knowledge
6	Topic_3	649	0.079	Learning, students, teaching, online, course
8	Topic_4	551	0.067	Medical, interprofessional, students, education, training
2	Topic_5	1337	0.164	Students, clinical, learning, nursing, study
9	Topic_6	506	0.062	Review, studies, education, nursing, learning
4	Topic_7	994	0.122	Students, nursing, group, study, learning
1	Topic_8	1341	0.164	Nursing, health, education, faculty, students
3	Topic_9	998	0.122	Nursing, learning, students, education, teaching

AlphaSum, Beta, BetaSum составили соответственно 1,5; 0,04; 1484,4. Если $\alpha<1$ и $\beta<1$, то есть вероятность распределения ближе к реальному распределению тем в корпусе данных. Среди показателей так же имеют высокое значение это «Avg(coherence)», «Avg(rank_1_docs)», «Avg(word-length)», «Avg(exclusivity)». «Avg(document_entropy)», «Corpus_dist» [4].

Таблица 3

Вектор производительности процесса тематического моделирования при помощи скрытого распределения Dirichlet (LDA)

	Topic_0	*Topic_1*	*Topic_2*	*Topic_3*	*Topic_4*	*Topic_5*	*Topic_6*	*Topic_7*	*Topic_8*	*Topic_9*
Coherence	-8,3	-13,2	-16,4	-12,3	-9,9	-7,4	-7,4	-8,0	-10,4	-10,3
Rank_1_docs	0,2	0,1	0,1	0,1	0,1	0,2	0,2	0,2	0,2	0,1
Word-length	8,0	6,6	6,6	7,2	9,8	7,2	7,4	6,6	7,4	8,0
exclusivity	0,4	0,2	0,5	0,3	0,4	0,1	0,4	0,2	0,2	0,1
Document_entropy	7,8	7,5	7,2	7,7	7,3	8,1	6,8	7,8	8,1	8,0
Eff_num_words	146,5	236,7	222,5	193,6	222,9	212,4	237,5	198,4	288,0	279,9
Uniform_dist	3,8	3,4	3,2	3,5	3,5	3,6	3,7	3,9	3,4	3,3
Corpus_dist	0,9	0,9	1,6	0,9	1,1	0,5	1,2	0,8	0,7	0,7
Token-doc-diff	0,01	0,004	0,003	0,004	0,01	0,004	0,001	0,004	0,002	0,002
Allocation_count	0,2	0,2	0,1	0,2	0,2	0,3	0,3	0,2	0,3	0,2

«Coherence» (когерентность)) - этот показатель измеряет, имеют ли тенденцию слова в теме встречаться вместе. Баллы суммируются за каждую отдельную пару слов с самым высоким рейтингом. Оценка представляет собой логарифм вероятности того, что документ, содержащий хотя бы один экземпляр слова с более высоким рейтингом, также содержит хотя бы один экземпляр слова с более низким рейтингом. Большие отрицательные значения указывают на слова, которые встречаются нечасто; значения ближе к нулю указывают на то, что слова имеют тенденцию чаще встречаться одновременно.

«Rank_1_docs» (ранг_1_документ) - некоторые темы в документах специфичны, в то время как другие на самом деле не «темы», а язык, который возникает, потому что мы пишем в определенном контексте. Содержательная тема встречается в относительно небольшом количестве документов, но если это произойдет, то будет создано много токенов. «Фоновая» тема встречается во многих документах и имеет большое общее количество токенов, но никогда не создает много токенов в одном документе. Этот показатель подсчитывает частоту, с которой данная тема является самой часто встречающейся темой в документе. Темы с большим количеством токенов часто имеют мало документов ранга 1.

«Word-length» (длина слова) - средняя длина слов в символах. Более длинные слова часто несут более конкретное значение, поэтому что, если тема объединяет много коротких слов, вероятно, это не очень конкретная тема.

«Exclusivity» (эксклюзивность) – этот показатель измеряет степень, в которой самые популярные слова в этой теме не появляются в качестве самых популярных слов в других темах, т.е. степень, в которой самые популярные слова являются «эксклюзивными». Значение представляет собой среднюю по каждому верхнему слову вероятность появления этого слова в теме, деленную на сумму вероятностей этого слова во всех темах. Как часто из самых популярных слов в теме они встречаются в других темах? Эксклюзивность коррелирует (отрицательно) с количеством токенов, но также указывает на более расплывчатые и общие темы.

«Document_entropy» - для этой метрики рассчитывается вероятность появления документов по заданной теме. Подсчитывается частота появления темы по всем документам и нормализуется, чтобы получить распределение, а затем вычисляется энтропия этого распределения. Низкая энтропия это не обязательно хорошо: она может указывать на необычные документы (вы случайно импортировали файл) или на наличие документов на других языках.

«Corpus_dist» (корпус_дистанция) - этот показатель измеряет, насколько далека тема от общего распределения слов в корпусе — по сути, то, что вы получили бы, если бы «обучили» модель одной теме, рассчитывается расстояние с использованием расхождения Kullback-Leibler. Большее расстояние означает, что тема более различима; меньшее расстояние означает, что тема больше похожа на распределение корпуса.

Обсуждение. За 10 лет сестринское образование за рубежом быстро развивалось и менялось, точно также как и само здравоохранение. Результаты обучения медицинских сестер зависят от технологии обучения и ее составляющих - моделей и методов обучения. Модели преподавания в сестринском образовании имеют прочную теоретическую основу и описывают среду обучения, которая включает общую цель; методы обучения; поведение, оценку знаний и предыдущие знания учащихся. В отличие от моделей, методы обучения имеют более узкое понятие и представляют собой способ взаимодействия между преподавателем и учащимися, в результате которого происходит передача и усвоение знаний, умений и навыков, предусмотренных содержанием обучения. В данном обзоре мы

бы хотели проследить эволюцию моделей и методов обучения в сестринском деле и их эффективность, а также использовать text mining для предварительного отбора статей в корпусе данных.

За последние несколько десятилетий проблемы классификации текста широко изучались и решались во многих приложениях [5-7]. Исследователи заинтересованы в разработке приложений, использующих методы классификации текста в связи с быстрым ростом публикаций по различным научным темам и возникающей проблемой их классификации. Самый важный этап классификации документов — выбор лучшего классификатора. Без полного концептуального понимания каждого алгоритма, невозможно определить наиболее эффективную модель для классификации нашего корпуса текстовых данных. Поэтому были рассмотрены все возможные варианты в программе Rapidminer - логистическая регрессия, наивный Байес, дерево решений, случайный лес, k-ближайший сосед (KNN), машина опорных векторов (SVM) [8-10], некоторые из этих алгоритмов показали высокую валидность и надежность, но их использование часто зависит от знаний пользователя в области информатики и применяемых алгоритмов, возможностей персонального компьютера (оперативная память), используемых программ или платформ для интеллектуального анализа текста. Было выбрано тематическое моделирование при помощи скрытого распределения Dirichlet (LDA), эта модель неконтролируемого машинного обучения выводит скрытые темы из текстовых документов, корпусов или электронных архивов с помощью вероятностного подхода и наиболее часто используется для классификации документов при подготовке систематических обзоров [11].

Быстро развивающаяся область машинного обучения разрабатывает программные инструменты, которые помогают в составлении систематических обзоров [8]. Машинное обучение предлагает подходы, позволяющие избежать ручного и трудоемкого отбора большого количества исследований путем определения приоритетности соответствующих исследований посредством активного обучения. Однако существующие инструменты имеют существенные недостатки. Многие из них представляют собой приложения с закрытым исходным кодом и алгоритмами «черного ящика», что проблематично, поскольку прозрачность и владение данными необходимы в эпоху открытой науки [12]. Существующим программам и платформам не хватает необходимой гибкости для работы с широким спектром возможных наборов данных для систематического обзора. В

систематических обзорах оптимальный тип классификатора будет зависеть от переменных параметров, таких как доля релевантных публикаций в первоначальном поиске и сложность критериев включения, используемых исследователем [13]. По этой причине любая успешная система должна учитывать широкий спектр типов классификаторов. Сравнительное тестирование имеет решающее значение для понимания реальной производительности любой системы машинного обучения, но такие варианты тестирования в настоящее время по большей части отсутствуют [14].

Автоматизированные методы контент-анализа продемонстрировали эффективность в решение целого ряда существенных проблем. Однако эти методы не устраняют необходимость чтения текстов. Действительно, глубокое понимание текстов является одним из ключевых преимуществ человека при применении автоматизированных методов. Необходимо применение нескольких методов для классификации текстовых данных, так как применяемые модели могут вводить в заблуждение или просто ошибаться [15].

Заключение. Исследования по автоматизации систематических обзоров с помощью text mining смещены в сторону этапа проведения, т.е. этапа отбора статей из набора данных. Ряд авторов предлагает использовать text mining на этапе формулирования поисковых запросов, определения критериев включения и исключения, представления полученных данных [14,16]. Важнейшими аспектами использования технологий и инструментов интеллектуального анализа текста для подготовки систематических обзоров являются надежные, масштабируемые, эффективные и доступные сервисы для классификаций очень больших коллекций документов. Не менее важным является вопрос о том, каков правильный баланс между автоматизацией процесса и вмешательством пользователя, контролем [16-18].

Эффективность методов автоматизации систематических обзоров на основе искусственного интеллекта изучалась в таких областях, как медицина или вычислительная техника. Применение одного метода для решения одной и той же задачи в другой области может быть неприемлемым из-за специфической терминологии или других типов исследовательских работ. На данный момент не разработаны алгоритмы применения интеллектуального анализа текста полностью понятные и надежные для значительной автоматизации подготовки систематических обзоров в различных областях.

В нашем случаем применение тематического моделирования при помощи скрытого распределения Dirichlet (LDA) дало очень хорошие результаты, но может оказаться, что при другом наборе данных, эта модель будет неэффективна.

Список литературы

1. Li D., Wang Z., Wang L., Sohn S. et all. A Text-Mining Framework for Supporting Systematic Reviews. Am. J. Inf. Manag. 2016;1(1):1-9. PMID: 29071308; PMCID: PMC5653323.

2. Малошонок Н.Г., Щеглова И.А. Модели организации обучения студентов в университете: основные представления, преимущества и ограничения. Университетское управление: практика и анализ. 2020;24(2):107-120. DOI: 10.15826/umpa.2020.02.017

3. Newman D., Asuncion A., Smyth P., Welling M. Distributed algorithms for topic models. J. Mach. Learn. Res. 2009;10:1801–1828.

4. McCallum A.K. MALLET: A machine learning for language toolkit. 2002. URL (last checked 26 June 2012). http://mallet.cs.umass.edu

5. Jiang M., Liang Y., Feng X. et all. Text classification based on deep belief network and softmax regression. Neural. Comput. Appl. 2018;29:61–70. DOI: 10.1007/s00521-016-2401-x

6. Kowsari K., Brown D.E., Heidarysafa M., Jafari Meimandi K., Gerber M.S., Barnes L.E. HDLTex: Hierarchical Deep Learning for Text Classification. Machine Learning and Applications (ICMLA). In Proceedings of the 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 18–21 December 2017. DOI: 10.1109/ICMLA.2017.0-134

7. Audebert N., Herold C., Slimani K., Vidal C. (). Multimodal Deep Networks for Text and Image-Based Document Classification. In: Cellier, P., Driessens, K. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Communications in Computer and Information Science, 2020, Vol. 1167. Springer, Cham. DOI:10.1007/978-3-030-43823-4_35

8. Li D, Wang Z, Wang L, Sohn S, Shen F, Murad MH, Liu H. A Text-Mining Framework for Supporting Systematic Reviews. Am J Inf Manag. 2016 Nov;1(1):1-9. PMID: 29071308; PMCID: PMC5653323.

9. Bannach-Brown A., Przybyła P., Thomas J. et al. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. Syst. Rev. 2019;8:2-12. DOI: 10.1186/s13643-019-0942-7.

10. Khalil H., Ameen D., Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. Journal of Clinical Epidemiology. 2022;144:22-42. DOI:10.1016/j.jclinepi.2021.12.005},

11. Iparraguirre-Villanueva O., Sierra-Liñan F., Salazar J.L. et all. Search and classify topics in a corpus of text using the latent dirichlet allocation model. Indonesian Journal of Electrical Engineering and Computer Science. 2023;30(1):246~256. DOI: 10.11591/ijeecs.v30.i1.pp246-256

12. Hassija V., Chamola V., Mahapatra A. et al. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cogn. Comput. 2023. https://arxiv.org/ftp/arxiv/papers/2304/2304.04780.pdf. DOI: 10.1007/s12559-023-10179-8

13. Schmidt L., Finnerty Mutlu A.N., Elmore R. et all. Data extraction methods for systematic review (semi)automation: Update of a living systematic review. F1000Res. 2021;19;10:401-420. DOI: 10.12688/f1000research.51117.2.

14. MacFarlane A., Russell-Rose T., Shokraneh F. Search strategy formulation for systematic reviews: Issues, challenges and opportunities. Intelligent Systems with Applications. 2022;15:2667-3053. DOI:10.1016/j.iswa.2022.200091

15. Caballero-Julia D., Campillo P. Epistemological Considerations of Text Mining: Implications for Systematic Literature Review. Mathematics. 2021;9(16):1-26. DOI:10.3390/math9161865

16. L. Feng Y., Chiam and S. Lo, Text-Mining Techniques and Tools for Systematic Literature Reviews: A Systematic Literature Review. 24-th Asia-Pacific Software Engineering Conference (APSEC), Nanjing, China. 2017:41-50. DOI: 10.1109/APSEC.2017.10

17. Voskanyan Y., Shikina I., Kidalov F., Davidov D.Medical Care Safety - Problems and Perspectives. In: Antipova T. (eds) Integrated Science in Digital Age. ICIS 2019. Lecture Notes in Networks and Systems, vol 78. Springer, Cham. DOI: 10.1007/978-3-030-22493-6_26

18. Voskanyan Y., Shikina I., Kidalov F., Andreeva O., Makhovskaya T.Impact of Macro Factors on Effectiveness of Implementation of Medical Care Safety Management System. (2021) Impact of Macro Factors on Effectiveness of Implementation of Medical Care Safety Management System. In: Antipova T. (eds) Integrated Science in Digital Age 2020. ICIS 2020. Lecture Notes in Networks and Systems, vol 136. Springer, Cham. https://doi.org/10.1007/978-3-030-49264-9_31

References

1. Li D., Wang Z., Wang L., Sohn S. et all. A Text-Mining Framework for Supporting Systematic Reviews. Am. J. Inf. Manag. 2016;1(1):1-9. PMID: 29071308; PMCID: PMC5653323.

2. Maloshonok N.G., Shcheglova I.A. Models of organization of teaching students at the university: basic assumptions, advantages and limitations. Universitetskoe upravlenie: praktika i analiz. 2020;24(2):107-120. DOI: 10.15826/umpa.2020.02.017

3. Newman D., Asuncion A., Smyth P., Welling M. Distributed algorithms for topic models. J. Mach. Learn. Res. 2009;10:1801–1828.

4. McCallum A.K. MALLET: A machine learning for language toolkit. 2002. URL (last checked 26 June 2012). http://mallet.cs.umass.edu

5. Jiang M., Liang Y., Feng X. et all. Text classification based on deep belief network and softmax regression. Neural. Comput. Appl. 2018;29:61–70. DOI: 10.1007/s00521-016-2401-x

6. Kowsari K., Brown D.E., Heidarysafa M., Jafari Meimandi K., Gerber M.S., Barnes L.E. HDLTex: Hierarchical Deep Learning for Text Classification. Machine Learning and Applications (ICMLA). In Proceedings of the 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 18–21 December 2017. DOI: 10.1109/ICMLA.2017.0-134

7. Audebert N., Herold C., Slimani K., Vidal C. (). Multimodal Deep Networks for Text and Image-Based Document Classification. In: Cellier, P., Driessens, K. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Communications in Computer and Information Science, 2020, Vol. 1167. Springer, Cham. DOI:10.1007/978-3-030-43823-4_35

8. Li D, Wang Z, Wang L, Sohn S, Shen F, Murad MH, Liu H. A Text-Mining Framework for Supporting Systematic Reviews. Am J Inf Manag. 2016 Nov;1(1):1-9. PMID: 29071308; PMCID: PMC5653323.

9. Bannach-Brown A., Przybyła P., Thomas J. et al. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. Syst. Rev. 2019;8:2-12. DOI: 10.1186/s13643-019-0942-7.

10. Khalil H., Ameen D., Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. Journal of Clinical Epidemiology. 2022;144:22-42. DOI:10.1016/j.jclinepi.2021.12.005},

11. Iparraguirre-Villanueva O., Sierra-Liñan F., Salazar J.L. et all. Search and classify topics in a corpus of text using the latent dirichlet allocation model. Indonesian Journal of Electrical Engineering and Computer Science. 2023;30(1):246~256. DOI: 10.11591/ijeecs.v30.i1.pp246-256

12. Hassija V., Chamola V., Mahapatra A. et al. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cogn. Comput. 2023. https://arxiv.org/ftp/arxiv/papers/2304/2304.04780.pdf. DOI: 10.1007/s12559-023-10179-8

13. Schmidt L., Finnerty Mutlu A.N., Elmore R. et all. Data extraction methods for systematic review (semi)automation: Update of a living systematic review. F1000Res. 2021;19;10:401-420. DOI: 10.12688/f1000research.51117.2.

14. MacFarlane A., Russell-Rose T., Shokraneh F. Search strategy formulation for systematic reviews: Issues, challenges and opportunities. Intelligent Systems with Applications. 2022;15:2667-3053. DOI:10.1016/j.iswa.2022.200091

15. Caballero-Julia D., Campillo P. Epistemological Considerations of Text Mining: Implications for Systematic Literature Review. Mathematics. 2021;9(16):1-26. DOI:10.3390/math9161865

16. L. Feng Y., Chiam and S. Lo, Text-Mining Techniques and Tools for Systematic Literature Reviews: A Systematic Literature Review. 24-th Asia-Pacific Software Engineering Conference (APSEC), Nanjing, China. 2017:41-50. DOI: 10.1109/APSEC.2017.10

17. Voskanyan Y., Shikina I., Kidalov F., Davidov D.Medical Care Safety - Problems and Perspectives. In: Antipova T. (eds) Integrated Science in Digital Age. ICIS 2019. Lecture Notes in Networks and Systems, vol 78. Springer, Cham. DOI: 10.1007/978-3-030-22493-6_26

18. Voskanyan Y., Shikina I., Kidalov F., Andreeva O., Makhovskaya T.Impact of Macro Factors on Effectiveness of Implementation of Medical Care Safety Management System. (2021) Impact of Macro Factors on Effectiveness of Implementation of Medical Care Safety Management System. In: Antipova T. (eds) Integrated Science in Digital Age 2020. ICIS 2020. Lecture Notes in Networks and Systems, vol 136. Springer, Cham. https://doi.org/10.1007/978-3-030-49264-9_31

Финансирование. Исследование не имело спонсорской поддержки.
Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.
Acknowledgments. The study did not have sponsorship.
Conflict of interests. The authors declare no conflict of interest.

Сведения об авторах

Бегун Дмитрий Николаевич – доктор медицинских наук, доцент, заведующий кафедрой сестринского дела, ФГБОУ ВО «Оренбургский государственный медицинский университет» Минздрава России, ул. Советская, д. 6, г. Оренбург, 460000, Российская Федерация, e-mail: doctorbegun@yandex.ru, ORCID 0000-0002-8920-6675, SPIN-код: 8443-4400

Гаврилова Екатерина Владиславовна – старший преподаватель кафедры сестринского дела ФГБОУ ВО «Оренбургский государственный медицинский университет» Минздрава России, ул. Советская, д. 6, г. Оренбург, 460000, Российская Федерация, e-mail: ekaterina2474@mail.ru, ORCID 0000-0001-9580-9045, SPIN-код: 3177-1114

Мирзаева Нелли Владимировна – ассистент кафедры сестринского дела, ФГБОУ ВО «Оренбургский государственный медицинский университет» Министерства здравоохранения Российской Федерации, г. Оренбург, Россия, г. Оренбург, ул. Советская, 6, e-mail: arhipova.nelli@mail.ru, ORCID 0009-0000-0832-2192, SPIN-код 1399-1429

Головко Ольга Валентиновна – старший преподаватель кафедры сестринского дела, ФГБОУ ВО «Оренбургский государственный медицинский университет» Минздрава России, ул. Советская, д. 6, г. Оренбург, 460000, Российская Федерация, email: golovko.040371@mail.ru, ORCID 0000-0001-6515-8683, SPIN-код: 3672-2138

Заришняк Наталья Владимировна – кандидат медицинских наук, ассистент кафедры сестринского дела, ФГБОУ ВО «Оренбургский государственный медицинский университет» Минздрава России, ул. Советская, д. 6, г. Оренбург, 460000, Российская Федерация, email: wengerenko@mail.ru, ORCID 0000-0003-2742-3161, SPIN-код: 1307-1759

Information about the authors

Begun Dmitry Nikolaevich – doctor of Medical Sciences, Associate Professor, Head of the Department of Nursing, Federal State Budgetary Educational Institution of Higher Education "Orenburg State Medical University" of the Ministry of Health of the Russian Federation, Russia, Orenburg, st. Sovetskaya, 6, e-mail: doctorbegun@yandex.ru, ORCID 0000-0002-8920-6675, SPIN-код 8443-4400

Gavrilova Ekaterina Vladislavovna – senior lecturer of the Department of Nursing, Federal State Budgetary Educational Institution of Higher Education "Orenburg State Medical University" of the Ministry of Health of the Russian Federation, st. Sovetskaya, 6, Orenburg, 460000, Russian Federation, email: ekaterina2474@mail.ru, ORCID 0000-0001-9580-9045, SPIN: 3177-1114

Mirzaeva Nelli Vladimirovna – assistant of the Department of Nursing, Federal State Budgetary Educational Institution of Higher Education "Orenburg State Medical University" of the Ministry of Health of the Russian Federation, Orenburg, Russia, Orenburg, st. Sovetskaya, 6, e-mail: arhipova.nelli@mail.ru, ORCID 0009-0000-0832-2192, SPIN-код 1399-1429

Golovko Olga Valentinovna – senior lecturer of the Department of Nursing, Federal State Budgetary Educational Institution of Higher Education “Orenburg State Medical University” of the Ministry of Health of Russia, st. Sovetskaya, 6, Orenburg, 460000, Russian Federation, email: golovko.040371@mail.ru, ORCID 0000-0001-6515-8683, SPIN: 3672-2138

Zarishnyak Natalya Vladimirovna – candidate of medical sciences, assistant of the department of nursing, Federal State Budgetary Educational Institution of Higher Education "Orenburg State Medical University" of the Ministry of Health of the Russian Federation, st. Sovetskaya, 6, Orenburg, 460000, Russian Federation, email: wengerenko@mail.ru, ORCID 0000-0003-2742-3161, SPIN: 1307-1759

Статья получена: 06.12.2023 г.
Принята к публикации: 25.03.2024 г.