DOI:

ОБРАБОТКА ОТКРЫТЫХ ВОПРОСОВ ВЕБ-ОПРОСОВ В СИСТЕМЕ ОБРАЗОВАНИЯ НА ОСНОВЕ МЕТОДОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА[1]

Силаева А.Э.[1], Никульчев Е.В.[1], Ильин Д.Ю.[1], Малых С.Б.[2]

[1]МИРЭА – Российский технологический университет, Россия, г. Москва, пр. Вернадского, д. 78; Российская академия образования, Россия, г. Москва,, ул. Погодинская, д. 8

[2] Психологический институт РАО, Россия, г. Москва,, ул. Москва, д. 9 к.4

silaeva.a.e@edu.mirea.ru, nikulchev@mail.ru, i@dmitryilin.com, malykhsb@mail.ru

Аннотация: Предложена методика анализа результатов массовых веб-опросов с открытыми вопросами, разработанная на основе использования методов обработки текстовых данных и LDA (латентное размещение Дирихле). Приведены результаты применения разработанного подхода для анализа профессионального веб-опроса в системе образования.

Ключевые слова: не менее трех ключевых слов, отражающих тему работы.

Введение

При проведении опросов на каждый открытый вопрос может быть дан уникальный ответ, при этом ответы могут быть разными по форме, но одинаковыми по смыслу. В случае большого объема данных для обработки текстовых ответов применяют методы искусственного интеллекта [1]. Однако использование инструментов искусственного интеллекта является существенной проблемой – применяемые модели интеллектуального анализа текста различны для каждого случая, так как каждая область имеет набор определенных слов с различными семантическими значениями [2].

В представленном исследовании тематическое моделирование проведено с помощью метода LDA (Laten Dirichlet Allocation) [3]. Этот метод вычислительного анализа используется для исследования тематической структуры коллекции текстовых данных [4].

В работе используется результатов веб-опросов в системе образования. Веб-опрос проводился с использованием системы DigitalPsyTools.ru [5] в форме обезличенных анкет. В системе предусмотрен платформонезависимый интерфейс, позволяющий проходить опрос на любом доступном устройстве. Корректность данных и осмысленность ответов в значительной степени были обеспечены организационными мероприятиями, проведенными органами управления образования. В опросе учувствовало более 38 тыс. чел. из 86 регионов.

1 Методики исследования

Из 26 вопросов анкеты со свободным ответом было выбрано два, так как они имели максимальное количество ответов:

1) «Есть ли у Вас еще какие-либо обязанности в школе, прямо не относящиеся к Вашей психологической деятельности?» – 3694 ответов;

2) «Укажите Ваши пожелания по развитию психологической службы в системе образования Вашего субъекта?» – 16700 ответов.

Методика анализа текстовых данных открытых вопросов состоит из четырех этапов.

A. Подготовка ответов свободной формы.
B. Анализ текстовых данных и тематическое моделирование.
C. Интерпретация тематического моделирования.
D. Группировка и использование результатов.

На этапе A проводится предобработка и валидация данных; происходит сбор, фильтрация данных и предварительный экспертный анализ текста, выявление особенностей формата представленных ответов. На этапе B на основе полученных данных, переданных из этапа A, создается корпус и словарь, применяются интеллектуальные алгоритмы с целью выявить закономерности и ключевые слова, наиболее точно определяющие выявленные группы. Для тематического моделирования применялся алгоритм LDA. На этапе C производится интерпретация выходных данных алгоритма LDA. Для этапа D выбраны дополнительные вопросы, с помощью которых, можно было сгруппировать свободные ответы по темам из этапа C.

Для предобработки данных разработан следующий алгоритм.

[1] *Работа выполнена при финансовой поддержке РФФИ (грант 17-29-02198)*

Шаг 1. Требуется оценить степень корректности ответов с семантической и грамматической точек зрения, выявить части речи, которые являются наиболее и наименее значимыми для последующей группировки ответов по тематикам. Выделение «стоп-слова», которые наименее значимы и подлежат удалению из текста. В рассматриваемом опросе к наименее значимым отнесены глаголы, которые не отражали принадлежность ответа к темам. При этом удалены пустые строки, односложные слова. Производится очистка исходного файла данных от посторонней информации. Определяются знаки пунктуации ".", ";", "!", "?" и последующей заглавной буквы как разделителя. Каждое предложение получает такой же id, какой был для изначального ответа, т.е. по итогу возможна ситуация, когда два и более предложений имеют один идентификатор. Возможны ситуации, когда ответ введен без соблюдения правил орфографии, что помешает разделить ответ на несколько предложений. Однако ситуация, когда предложения не были разделены (таким образом, запись может содержать несколько смыслов), видится более предпочтительной, чем целостное предложение, разделенное на несколько частей, когда части не позволят определить их смысл.

Шаг 2. Разбиение предложений на токены, лемматизация каждого слова, преобразование слов к начальной форме с исправлением до 2 опечаток, очистка от пунктуации, стоп-слов и замена буквы "е" на "е". {Заметим, что выделение в текстовом потоке минимальных фрагментов для последующего анализа в корпусной лингвистике принято называть токены (англ. token); лемматизация (англ. lemmatization): определение для всех токенов их начальной формы — леммы (англ. lemma)}[6]). Лемматизация реализуется с помощью библиотек Az.js (https://github.com/deNULL/Az.js) и pymorphy2 (https://github.com/kmike/pymorphy2), а перечень, стоп-слов взят из пакетов (https://github.com/stopwords-iso/stopwords-ru) и (https://www.nltk.org/api/nltk.html). После применения алгоритма LDA в этапе С список стоп-слов будет пополняться из-за специфики данных, которая была определена на этапе А.

Шаг 3. Построение словаря допустимых замен слов на основе графа подобия слов. Для рассматриваемой задачи использован Russian Distributional Thesaurus (https://nlpub.mipt.ru/Russian_Distributional_Thesaurus). Файл представлен в формате DSV (delimiter separated values), где каждая строчка содержит 1 запись. Запись состоит из левой части (слово) и правой части (список подобных слов, разделенный запятыми). Каждое слово из правой части имеет оценку, обозначающую степень подобия слову в левой части записи. В качестве допустимой замены принимались слова с оценкой > 0.6, т.е. обладающих достаточной степенью подобия применительно к задаче. Каждое из выбранных слов также приводится к начальной форме.

Шаг 4. Сокращение разнообразия слов, используемых в ответах. Удаление малозначимых слов с точки зрения задачи. На базе отобранных пар (слово + список слов для замены), производится замена всех подобных между собой слов на единый вариант словоупотребления. В качестве принимаемого варианта используется первый, найденный по тексту. Также из предложений удаляются те части речи, которые не несут ценной информации (исходя из экспертной оценки, полученной на шаге 1 в этапе А). Для данной задачи – это глаголы, числа и какие-либо токены на латинице.

Для того чтобы применить алгоритм для разделения данных на отдельные группы, нужно привести их в векторный вид. В работе использовалась модель “мешок слов” [7], основная идея которой состоит в том, что смысл и сходство кодируются в виде вектора частотами появлений слов в документе. В этой схеме кодирования каждый документ представляется как мультимножество составляющих его лексем, а значением для каждой позиции слова в векторе служит счетчик соответствующего слова. Значения могут быть простыми целочисленными счетчиками, как на рис. 1, или взвешиваться общим количеством слов в документе.

Рис. 1. Представление частот слов в виде вектора с помощью модели “мешок слов”

Для свободных ответов использовалась модель векторизации “мешок слов” [8]. После преобразования получается корпус (структурированное представление данных) вида [[(2, 1), (3, 1), (4,

1), (5, 1), (6, 1)] …], в котором каждый кортеж массива является словом. Первый элемент кортежа id – это уникальный индификатор слова, второй элемент количество употребления слова.

К строенным векторным представлениям применяется латентное размещение Дирихле (LDA). Он принадлежит семейству порождающих вероятностных моделей, в которых темы представлены вероятностями появления каждого слова из заданного набора. Документы, в свою очередь, могут быть представлены как сочетания этих тем. Уникальная особенность моделей LDA состоит в том, что темы не обязательно должны быть различными и слова могут встречаться в нескольких темах; это придает некоторую нечеткость определяемым темам, что может пригодиться для совладения с гибкостью языка (рис. 2). Тремя основными входными данными для LDA [9] являются словарь, корпус и количество тем.

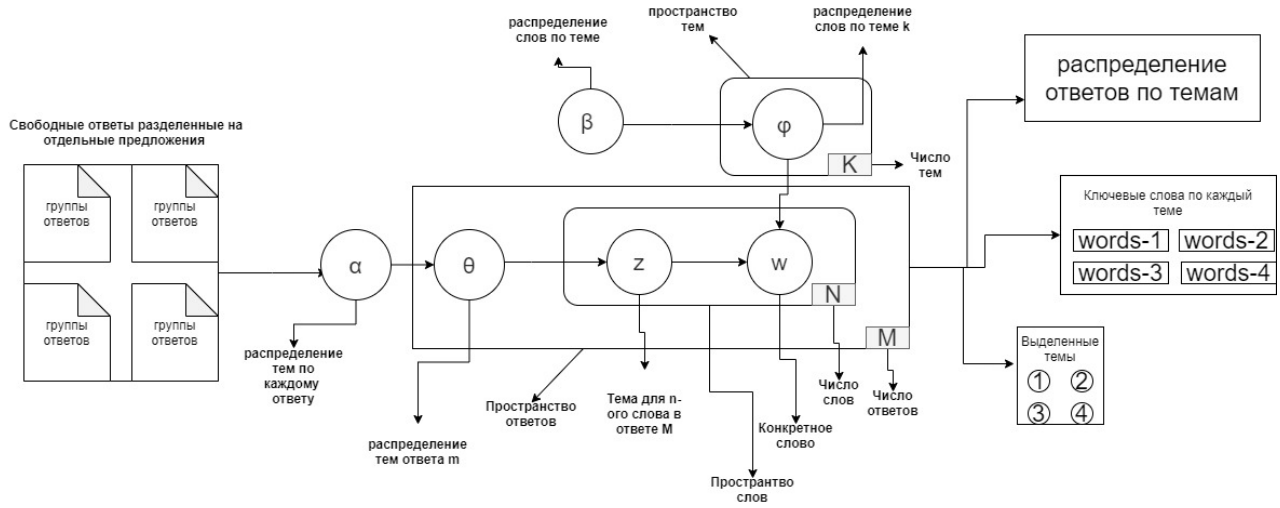

Рис. 2. Алгоритм работы LDA

Применялся алгоритм из пакета (https://github.com/RaRe-Technologies/gensim) для вопроса (1), пакет из (https://github.com/seanmowz/lda-topic-model) применялся для вопроса (2). Метод латентного размещения Дирихле дает наблюдаемое слово или лексему, по которому можно определить вероятную тему, распределение слов в каждой теме и сочетание тем в документе.

На выходе алгоритма LDA ключевые слова по каждой теме. Чтобы оценить насколько алгоритм отработал успешно, нужна экспертная оценка и метрика когерентности тем, оценивающая, насколько часто наиболее вероятные слова темы встречаются рядом в данных документах. Когерентность темы определяется как средняя совместная встречаемость двух слов по всем парам k наиболее вероятных слов темы [8]. Совместная встречаемость оценивается как поточечная взаимная информация (PMI) по документам, в которых встречаются оба слова. При оценивании эксперт должен выделить: в каждой теме были ключевые слова, которые значимы для отдельной темы; после разделения предложений на отдельные группы, общий смысл темы сохранялся; в ключевых словах должны присутствовать наиболее значимые для группы слова. Метрика принимает значения от 0 до 1; чем выше, тем лучше. Показывает, насколько хорошо ключевые слова темы отличает ее от остальных тем.

Поле получения модели тематического моделирования нужно разграничить свободные ответы с помощью применения полученной модели. Так как выходной результат работы модели LDA – это соотнесение каждого ответа к многоклассовой метки, то для каждого ответа выдается массив кортежей, где первый элемент – номер темы, а второе – принадлежность к тем.

Каждый свободный ответ получает вероятностное соотношение к теме, но так же для каждой темы, можно вывести массив кортежей ключевых слов, которые тоже имею вероятностный коэффициент, [(0, '0.260*"педагог" + 0.180*"директор" + 0.048*"служба" + 0.047*"внеурочный" + 0.029*"медиация" + 0.025*"медицинский" + 0.024*"часы" + 0.021*"питание" + 0.016*"обязанность" + 0.015*"урок"'), (1, '0.119*"сопровождение" + 0.101*"мероприятие" + 0.054*"дежурство" + 0.047*"директор" + 0.030*"лагерь" + 0.025*"комиссия" + 0.022*"летний" + 0.021*"замена" + 0.020*"педагог" + 0.018*"педагогический"'), (2, '0.100*"урок" + 0.065*"участие" + 0.052*"общеобразовательный" + 0.045*"педагог" + 0.043*"конкурс" + 0.038*"право" + 0.026*"соцсеть" + 0.025*"обучение" + 0.024*"организатор" + 0.023*"общественный"')], первый элемент кортежа номер темы, второй элемент – это ключевые слова с вероятностью появления в теме.

При просмотре ключевых слов и полных предложений становится возможным интерпретировать тему. Для того чтобы распределить ответы по темам, выводиться соотношение соотнесение к теме, массивы сортируются и выбирается максимальный коэффициент. Если интерпретация тем удовлетворяет эксперта, то можно дальше работать с данными, если нет, то возвращаемся на этап А для повторной валидации данных.

Результаты

Применение разработанной методики на рассматриваемом наборе данных опросы дало следующие результаты.

В вопросе (1) после очищение от нулевых значение, односложных слов, не валидных значений составила 3694 ответов.

В вопросе (2) после очищение от нулевых значение, односложных слов, не валидных значений составила 16564 ответов.

Данные с id и ответами были записаны в csv файл для дальнейшей обработки.

После предобработки ответов:

1. После разбиения ответов на отдельные предложения по вопросу (1) из 3694 ответов получилось 4980 предложения, по вопросу (2) из 16700 ответов получилось 25838 предложения.

2. Данные приведены в формат списка, каждое слово приняло нормальную форму. Длина словаря стоп-слов для вопроса (1) составило 644 слова, для вопроса (2) составило 859 слов.

3. После применения алгоритма по нахождению синонимов сформирован словарь синонимов для вопроса (1), который включал 546 пар слов, для вопроса (2) словарь составил 1865 пар слов.

При создании корпуса и словаря при прохождении всех шагов из этапа А итоговый словарь для вопроса (1) из 1646 сократился до 138 слов. Для вопроса (2) удалили все числа и глаголы, но редкие слова остались с частотой 1–2, итоговый словарь из 4943 сократился до 3079 слов.

На выходе алгоритма выдаются количество тем и ключевые слова, относящиеся к данным темам. К примеру, для вопроса (1) выделилось 4 темы, для вопроса (2) выделилось 7 тем.

Ключевые слова для первой темы для вопроса (1): 'директор', 'урок', 'заместитель', 'руководитель', 'секретарь', 'профсоюз', ' служба ', 'совет', 'нагрузка', 'администратор'.

Ключевые слова для первой темы для вопроса (2): 'психолог', 'ставка', 'педагог', 'школа', 'количество', 'часы', 'образовательный', 'необходимый', 'учебный', 'психология'.

При оценивании данных тем мера когерентности для вопроса составила (1) 0.541, а для вопроса (2) 0.657. Такие результаты означают, что часть данных разместилась не идеально, но так как ключевые слова отделили смысл тем друг от друга и интерпретировался общий смысл, это хороший результат.

По ответам на вопрос (1), получилось выделить 4 группы такие как:

1. Руководители и помощники руководителей.
2. Дежурные, воспитатели и вожатые в лагерях, библиотекари.
3. Преподаватели различных учебных профилей и репетиторы разных профилей.
4. Общественная деятельность, организаторы внеурочных мероприятий.

По группе 3 выделены конкретные совмещаемые должности (см. рис. 3).

По ответам на вопрос (2) выделено 7 групп.

1. Увеличение числа ставок в школе и выделение часов в учебном плане.
2. Создание центров и обеспечение методической поддержки.
3. Увеличение оплаты труда.
4. Проведение мероприятий по повышению квалификации.
5. Техническое оснащение кабинета и обеспечение диагностическими материалами.
6. Общее развитие системы.
7. Создание единой нормативной базы, документации и отчетности.

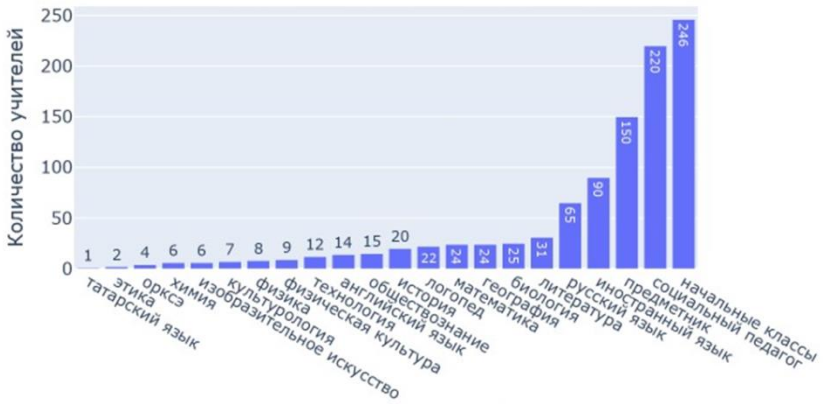

Рис. 3. Количество респондентов, совмещающих обязанности с должностью педагога-психолога

Использование других данных опроса, позволяет определять различные зависимости. На рис. 5 представлено количество педагогов-психологов, которые хотели пройти курсы по повышению квалификации, сгруппировано по возрасту.

Рис. 5. Распределение по возрастным группам педагогов, желающих пройти курсы по повышению квалификации

Заключение

В работе предложена методика анализа результатов массовых веб-опросов с открытыми вопросами, разработанная на основе использования методов обработки текстовых данных и LDA (латентное размещение Дирихле).

Сформированная методика позволила обработать результаты проведенного федерального профессионального опроса педагогов-психологов в системе образования.

Полученные результаты могут быть использованы для веб-исследований в системе образования с использованием цифровых платформ.

Литература

1. *Barakhnin V. B., Duisenbayeva A. N., Kozhemyakina O. Y., Yergaliyev Y. N., Muhamedyev R. I.* The automatic processing of the texts in natural language. Some bibliometric indicators of the current state of this research area. // Journal of Physics: Conference Series. 2018. Vol. 1117, No. 1. – P. 012001.
2. *Buenaño-Fernandez D., González M., Gil D., Luján-Mora S.* Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach // IEEE Access. 2020. V. 8. – P. 35318-35330.
3. *Blei D. M., Ng A. Y., Jordan M. I.* Latent dirichlet allocation // Journal of machine Learning research. 2003. V. 3. – P. 993-1022.
4. *Finch W. H., Hernández Finch M. E., McIntosh C. E., Braun C.* The use of topic modeling with latent Dirichlet analysis with open-ended survey items // Translational Issues in Psychological Science. 2018. V. 4. No. 4. – P. 403.

5. *Никульчев Е. В., Ильин Д. Ю., Колясников П. В., Исматуллина В. И., Захаров И. М., Малых С. Б.* Разработка открытой цифровой платформы масштабных психологических исследований // Вестник РФФИ. – 2019. №. 4. С. 105–119.
6. *Копотев М.* Введение в корпусную лингвистику: учеб. пособие — М.: Animedia Company, 2014.
7. *Wallach H. M.* Topic modeling: beyond bag-of-words // Proceedings of the 23rd international conference on Machine learning. – 2006. – P. 977-984.
8. *Mehrotra R., Sanner S., Buntine W., Xie L.* Improving LDA topic models for microblogs via tweet pooling and automatic labeling // Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. – 2013. – P. 889-892.
9. *Mimno D. Wallach H. M., Talley E., Leenders M., McCallum A.* Optimizing semantic coherence in topic models // Proceedings of the conference on empirical methods in natural language processing. – Association for Computational Linguistics, 2011. – P. 262-272.