УДК: 81’37 DOI: 10.31249/epl/2022.02.01

Литвинова Т.А.

Тематическое моделирование корпуса блогов на русском языке с учетом гендера автора: текст и контекст[*1]

Воронежский государственный педагогический университет, Россия, Воронеж, centr_rus_yaz@mail.ru

Аннотация. Тематическое моделирование относится к одному из видов методов исследования семантической организации текста и широко используется как для решения различных прикладных задач, так и для теоретических изысканий в области социологии, психологии, наукометрии и др. Однако в собственно лингвистических и социолингвистических исследованиях методы тематического моделирования используются не столь широко. Кроме того, классическое тематическое моделирование основано на анализе встречаемости слов в рамках документа и не учитывает локальную сочетаемость слов. В работе представлены результаты сравнительного исследования текстов блогов на русском языке с использованием метода латентного размещения Дирихле, примененного к матрицам двух видов: терм-документной (модель «Текст») и матрице совместной встречаемости слов в рамках контекстного окна (модель «Контекст»), с учетом гендера авторов, а также результаты эксперимента по классификации текстов по гендеру их авторов на основе вероятностей распределения тем. Были получены более высокие значения метрик качества тематического моделирования для моделей, построенных на матрицах совместной встречаемости слов («контекстах»). Высокая точность классификации текстов по гендеру

[1] Исследование выполнено в Воронежском государственном педагогическом университете при поддержке гранта Российского научного фонда № 21-78-10148 «Моделирование значения слова в индивидуальном языковом сознании на основе дистрибутивной семантики».

авторов указывает на то, что в текстах жанра «блог», предполагающего активное конструирование авторской идентичности, в том числе гендерной, присутствует ярко выраженный гендерный сигнал. Сравнение набора тем, вероятности распределения которых в текстах вносят наибольший вклад в классификацию, показало, что тематическое моделирование, выполненное на матрицах совместной встречаемости слов, позволяет выявить особенности семантической организации текстов, дополняющие результаты, полученные при традиционном тематическом моделировании.

Ключевые слова: семантика текста; компьютерная семантика; тематическое моделирование; гендерная атрибуция; блоги; русскоязычные корпусы текстов.

Поступила: 07.04.2022 Принята к печати: 16.08.2022

Litvinova T.A.

Topic modeling of the corpus of blogs in Russian with respect to author's gender: text and context*

Voronezh State Pedagogical University,
Russia, Voronezh, centr_rus_yaz@mail.ru

Abstract. Topic modeling aims to analyze the semantic organization of a text and is widely used both to deal with various applied problems and to conduct theoretical research in the field of sociology, psychology, scientometrics, etc. However, in linguistic and sociolinguistic studies topic modeling methods are used less frequently. Moreover, classical topic modeling is based on the analysis of the occurrence of words within a document and does not take into account the co-occurrence of words within context windows. The paper presents the findings of a comparative analysis of blog texts in Russian using Latent Dirichlet allocation applied to matrices of two types: a term-document matrix (Text model) and a term co-occurrence matrix (Context model), taking into account the gender of authors, as well as the results of an experiment on classifying texts by the gender of their authors based on the probabilities of the distribution of topics in texts. Higher values of topic modeling quality metrics were obtained for models built on term co-occurrence matrices («contexts»). High accuracy of the classification of texts by the author's gender reveals a clear gender signal in blogs – text

genre which involves active construction of the author's identity, including gender. Comparison of a set of topics the distribution probabilities of which in documents and contexts make the greatest contribution to the classifier showed that topic modeling performed on the term co-occurrence matrices makes it possible to identify features of the semantic organization of texts that complement the results obtained with traditional topic modeling.

Keywords: text semantics; computer semantics; topic modeling; gender attribution; blogs; Russian-language text corpora.

Received: 07.04.2022 Accepted: 16.08.2022

Введение

Одним из важнейших направлений современных исследований в области семантики и психолингвистики является изучение психологически актуального для носителей языка содержания слова [Пищальникова, 2019]. В качестве основных методов в таких исследованиях используются экспериментальные психолингвистические методы (ассоциативный эксперимент, семантическое шкалирование и т.д.), а также модели и методы дистрибутивной семантики, основанные на идее о том, что слова, которые используются в похожих контекстах, близки по значению (дистрибутивная гипотеза [Sahlgren, 2008]). Подобные модели позволяют получить векторные представления слов в некотором низкоразмерном пространстве (word embeddings), при этом семантически близкие слова будут иметь и близкие векторы. Модели векторной семантики активно разрабатываются в последние годы, однако хорошо известным существенным недостатком подобных моделей является «отсутствие интерпретируемости компонент построенных векторов» [Потапенко, 2018, с. 4], тогда как, по справедливому утверждению А.А. Потапенко, проводившей параллели с когнитивными науками, «векторные представления должны быть сильно разреженными, а их компоненты должны соответствовать отдельным семантическим признакам кодируемого понятия» [там же, с. 3].

Параллельно с векторными моделями семантики с конца 90-х годов XX в. активно развиваются методы тематического моделирования (topic modelling, далее также – ТМ), позволяющие

осуществлять кластеризацию слов и документов по темам, при этом каждая тема описывается вероятностным распределением на множестве слов. С лингвистических позиций темы представляют собой «кластеры слов, для которых характерна семантическая близость в корпусе текстов» [Митрофанова, 2015, с. 148], при этом тематическое моделирование, как правило, не учитывает локальную встречаемость слов.

Модели векторной семантики не отличаются концептуально от моделей ТМ, однако они применяются по отношению к матрицам разных типов. Вместо терм-документной матрицы (term-document matrix) на вход моделей векторной семантики подается матрица совместной встречаемости слов (term co-occurrence matrix).

Важным достоинством ТМ является интерпретируемость его результатов. К настоящему моменту разработан ряд моделей ТМ [Apishev, 2020], при этом самой популярной моделью для решения прикладных задач является латентное размещение Дирихле (Latent Dirichlet Allocation, LDA), предложенное Д. Блэй и соавторами [Blei, Ng, Jordan, 2003]. Существенным достоинством LDA является то, что извлечение тем из коллекции документов не требует какого-либо предварительного анализа корпуса.

ТМ используется для анализа текстов разных жанров: от научных текстов [Сравнение содержания …, 2019] до сказок [Митрофанова, 2015], однако наиболее широко методы ТМ используются для анализа текстов различных онлайн-сообществ: российской блогосферы [Кольцова, Маслинский, 2013], онлайн-сообществ мам [Кавеева, 2018], региональных сообществ [Методы тематического …, 2019]. Отдельным направлением использования ТМ выступает поиск вредоносного контента в Сети [Золотарев, Шарнин, Клименко, 2016], анализ настроений пользователей соцсетей [Чижик, 2021], особенностей построения имиджа известного лица или бренда территории [Гальченко, 2021], а также построение семантического профиля интернет-пользователя [Misztal-Radecka, 2018].

Следует отметить, однако, что ТМ в собственно филологических исследованиях используется редко (см., например, работу: [Кудин, 2021], в которой показана эффективность названной методологии для анализа концептов).

ТМ для сравнительного анализа семантической структуры текстов мужчин и женщин на русском языке, насколько нам известно, до настоящего времени не использовалось (см. об использовании методов ТМ для профилирования автора в работе: [Overview of the 4th Author …, 2016]). Кроме того, не проводилось сравнение резуль-

татов анализа текстов с использованием «классического» и «локального» ТМ, т.е. ТМ, примененного не к терм-документной матрице, а к матрице совместной встречаемости слов в контекстном окне.

Цель нашего исследования – сравнительный анализ семантической структуры текстов блогов на русском языке в гендерном аспекте с использованием методов тематического моделирования, примененных последовательно к матрицам разных типов: терм-документной и матрице совместной встречаемости слов в контекстном окне. Таким образом, научная новизна настоящего исследования определяется как его материалом, так и постановкой задачи.

Экспериментальное исследование текстов блогов с использованием методов тематического моделирования

Методология исследования

Материал исследования. В качестве материала исследования использовался созданный нами корпус текстов блогов на русском языке. Авторов-женщин в анализируемом корпусе – 966, мужчин – 1398. Алгоритм сбора корпуса описан нами в работе: [Litvinova, Sboev, Panicheva, 2018]. Для целей настоящего исследования корпус был подвергнут ручному анализу: удалялись профили, в которых сведения о поле автора не совпадали с полом автора, эксплицитно выраженным через соответствующие грамматические формы; профили авторов, чьи тексты целиком состояли из заимствованных текстов (новостей и т.д.). Очищенный таким образом корпус доступен в созданной нами базе данных RusIdiolect, специально предназначенной для идиолектных исследований [Litvinova, 2021].

В настоящем исследовании число авторов было сбалансировано по полу путем случайного отбора блогов 966 авторов-мужчин.

Предобработка корпуса. Хорошо известно, что обработка текста перед проведением дальнейших экспериментов является важным этапом работы, которому необходимо уделять самое пристальное внимание [Rodriguez, Spirling, 2022].

Обработка корпуса текстов для тематического моделирования проводилась по общепринятой схеме [Митрофанова, 2015]:

1. Осуществляется токенизация корпуса, после чего удаляются элементы, не являющиеся словом (цифры, знаки препинания, иные символы). Текст приводится к строчному виду. Далее исключаются токены, встречающиеся менее чем в десяти документах корпуса.

Данные процедуры были проведены с использованием пакета quanteda [Quanteda: An R package …, 2018].

2. Создается список стоп-слов, куда входят служебные слова, местоимения, некоторые наречия, числительные и т.д.

Общепринятого списка стоп-слов русского языка, применимого для всех задач в области автоматической обработки текстов, не существует, поэтому исследователи формируют свой список под конкретную задачу. В основу нашего списка был положен список стоп-слов, предложенный в пакете quanteda [Quanteda: An R package …, 2018], который был модицифирован в соответствии с особенностями корпуса и задачами исследования.

3. Проводится лемматизация корпуса (в нашем исследовании лемматизация осуществлялась с использованием морфоанализатора Treetagger [Schmid, 1994]).

4. Корпус разбивается на документы сообразно их первоначальной логической структуре (в нашем случае документ приравнивался к тексту блога; все блоги одного автора объединялись в один текст).

Средняя длина текста после описанной выше обработки составила 1246 токенов (медианное значение – 1089 токенов).

Анализ текстов с использованием алгоритма LDA. Тематическое моделирование осуществлялось нами на материале матриц двух типов: 1) терм-документной матрицы (модель «Текст»); 2) матрицы совместной встречаемости слов, которая строилась с использованием метода skip-grams с размером окна skipgram_window, равным пяти (т.е. подсчитывалось, сколько раз слово A встречается среди пяти слов до и после слова B) (модель «Контекст»). Подобный подход учитывает локальную встречаемость слов, тогда как традиционное ТМ учитывает встречаемость слов в рамках всего текста (т.е. весь текст рассматривается как контекст).

Оба типа матриц строились нами с использованием пакета textmineR [Jones, 2021]. Тематическое моделирование также проводилось с использованием названного пакета.

Тематическое моделирование предполагает определение числа тем. Вопрос о том, какое число тем следует использовать, является дискуссионным. Существует несколько подходов к его решению, которые можно разделить на две основные группы: 1) подходы, основанные на качественном анализе (интерпретируемости тем, соответствии решения поставленной задаче), и 2) подходы, ориентированные на количественные показатели (следует, однако, отметить, что общепринятого набора метрик, на которые следует ориентироваться при выборе числа тем, нет).

Мы использовали подход к оценке моделей, основанный на анализе значений коэффициента R^2 (отражает качество «подгонки» модели, т.е. насколько хорошо данные соответствуют модели) и метрики качества модели probabilistic coherence, реализованной в пакете textmineR [Jones, 2021].

Строились модели с числом тем от 10 до 100 с шагом 10, и для каждой модели вычислялись значения R^2 и probabilistic coherence, после чего для дальнейшего анализа были выбраны модели с самыми высокими значениями названных метрик. Далее на основе отобранных моделей вычислялись показатели вероятности распределения тем тета (Θ). Отметим, что в случае с матрицей совместной встречаемости слов Θ соответствуют эмбеддингам слов в вероятностном пространстве.

Мы рассматривали десять лемм в рамках каждой темы (применительно как к документам, так и к контекстам) с наиболее высокими значениями phi (Ф), т.е. нами рассматривался состав тем на уровне ядра.

Важно отметить, что в теме могут присутствовать леммы как из одного текста, так и из разных текстов с близким содержанием. Таким образом, в темах объединена лексика с общими дистрибутивными свойствами – «это слова, которые встречаются в сходном контекстном окружении и имеют близкие или смежные (не обязательно совпадающие) значения» [Митрофанова, 2015, с. 148].

При интерпретации результатов анализа следует помнить о том, что используемая нами тематическая модель LDA имеет вероятностный характер, а значит, наполнение темы не всегда полностью соответствует метке, которая обобщает значения основной части лемм из темы и назначается либо вручную, либо автоматически на основе набора правил.

Классификация текстов на основе вероятности распределения тем. Для оценки степени различительности выделенных тем в текстах авторов-мужчин и авторов-женщин проводился классификационный эксперимент. В качестве признаков (features) были использованы вероятности распределения выделенных тем в текстах (т.е. значения Θ), полученные при помощи пакета textmineR.

В качестве классификатора нами был выбран алгоритм машинного обучения «случайный лес» (random forest), поскольку он позволяет проводить классификацию текстов достаточно быстро и выделять наиболее значимые признаки (после тюнинга модели были выбраны параметры ntree=500, mtry=4, min n=50).

Выборка была разделена на тренировочную (90%) и тестовую (10%), при этом в тестовой выборке сохранялось распределение авторов по возрасту.

Результаты частотного анализа корпуса текстов блогов. Перед тем как провести тематическое моделирование, мы выполнили частотный анализ текстов.

На первом этапе мы вычисляли абсолютную частоту леммы в корпусе (т.е. число вхождений леммы некоторого слова во все документы коллекции) и встречаемость данной леммы в документах текстовой коллекции IDF, причем чем больше такая встречаемость, тем меньше IDF: леммы с низким IDF встречаются в большем числе текстов, с высоким – «концентрированно», т.е. в небольшом числе текстов.

Сравнение десяти самых частотных лемм (по показателям абсолютной частоты и IDF) в блогах авторов-мужчин и авторов-женщин показывает, что качественный состав рассматриваемых лемм совпадает (табл. 1).

Таблица 1

Самые частотные леммы в текстах блогов авторов-мужчин и авторов-женщин

Лемма	Ранг леммы по абсолютной частоте		Ранг леммы по IDF	
	Блоги авторов-женщин	Блоги авторов-мужчин	Блоги авторов-женщин	Блоги авторов-мужчин
Свой	1	1	1	1
Человек	2	2	2	3
Самый	3	4	3	4
Время	4	3	4	2
Знать	5	7	5	8
Жизнь	6	6	6	10
Говорить	7	8	7	6
Сказать	8	9	8	7
Хотеть	9	10	9	9
Стать	10	5	10	5

Далее было проведено взвешивание признаков на основе разностей вероятностей их встречаемости в блогах авторов-мужчин и авторов-женщин *prob_lift*:

$$prob\_lift = (\text{lemma} \mid \text{gender}_j) - P(\text{lemma}).$$

Подобный метод позволяет увидеть отличия в лексическом составе текстов блогов авторов-женщин и авторов-мужчин (табл. 2).

Таблица 2

Леммы с наибольшей разницей вероятностей встречаемости в текстах блогов женщин и мужчин

Леммы в блогах авторов-женщин с наибольшим значением prob_lift	prob_lift	Леммы в блогах авторов-мужчин с наибольшим значением prob_lift	prob_lift
Любить	0,000799	Россия	0,000492
Мама	0,000705	Город	0,000367
Хотеть	0,000705	Система	0,000281
Дитя	0,000614	Страна	0,000278
Знать	0,000588	Власть	0,000270
Муж	0,000552	Являться	0,000247
Жизнь	0,000552	Российский	0,000231
Говорить	0,000503	Украина	0,000213
Ребенок	0,000498	Область	0,000196
Хотеться	0,000431	Выборы	0,000207

Наличие подобных различий в вероятности встречаемости лемм позволяет ожидать и получение значимых результатов классификации текстов по гендеру авторов на основе вероятностей тем, а также сформировать представление о возможных наименованиях тем.

Результаты тематического моделирования текстов блогов. Мы строили последовательно несколько моделей с числом тем от 10 до 100 с шагом 10. Наиболее высокие значения метрик R^2 и probabilistic coherence были получены для значения k=50 (табл. 3).

Таблица 3

Метрики качества моделей LDA

Метрики качества моделей	Модель «Документ», к=50	Модель «Контекст», к=50
R^2	0,271391	**0,6255889**
Probabilistic coherence (среднее значение)	0,12587	**0,18940**
Probabilistic coherence (медианное значение)	0,12221	**0,19830**

Как показали результаты экспериментов, метрики качества при разных значениях k выше у моделей, построенных на контекстах.

Анализ тем с наиболее высоким значением критерия probabilistic coherence (табл. 4–5) позволяет заключить, что они являются хорошо интерпретируемыми (названия тем присвоены автоматически).

Таблица 4

Темы с наиболее высоким значением критерия probabilistic coherence (модель «Текст»)

Тема	Probabilistic coherence	Наиболее релевантные леммы
Выборы	0,301	*выборы, власть, кандидат, депутат, Россия.*
Рецепт	0,224	*масло, рецепт, тесто, яйцо, сахар.*
Фильм	0,211	*фильм, кино, герой, сюжет, главный.*
Театр	0,196	*театр, спектакль, выставка, искусство, художник.*
Суд	0,193	*суд, закон, право, дело, статья.*
Россия	0,192	*Россия, страна, Украина, российский, русский.*
Семья	0,187	*дитя, ребенок, семья, мама, муж.*
Книга	0,173	*книга, читать, автор, текст, история.*
Музыка	0,163	*песня, концерт, музыка, группа, альбом.*
Лес	0,163	*лес, земля, дерево, цветок, солнце.*

Таблица 5

Темы с наиболее высоким значением критерия probabilistic coherence (модель «Контекст»)

Тема	Probabilistic coherence	Наиболее релевантные леммы
Выборы	0,312	*закон, суд, область, выборы, право.*
Карта	0,307	*карта, телефон, документ, система, сайт.*
Деньги	0,304	*рубль, деньги, тысяча, цена, стоить.*
Процесс	0,298	*процесс, уровень, метод, система, результат.*
Церковь	0,289	*церковь, святой, храм, монастырь, бог.*
Школа	0,279	*школа, класс, точка, курс, мера.*
Россия	0,269	*россияне, страна, российский, власть, русский.*
Досуг	0,268	*фильм, играть, песня, игра, музыка.*
Дорога	0,265	*час, ехать, город, дорога, машина.*
Чтение	0,264	*книга, фильм, история, автор, читать.*

Отметим более высокие значения критерия probabilistic coherence у тем, выделенных моделью «Контекст». Качественный анализ тем также указывает на то, что темы модели «Контекст» легче интерпретировать.

Результаты классификационного эксперимента. В таблице 6 представлены результаты классификационного эксперимента с использованием в качестве признаков вероятностей распределения в текстах тем, полученных моделями «Текст» и «Контекст» с числом тем k=50. Для оценки качества классификационных моделей нами использовались следующие метрики: точность (*Accuracy*), т.е. процент верно определенных классов (в нашем случае – пол автора); точность (*Precision*), т.е. процент документов, отнесенных классификатором к написанным мужчинами (женщинами) и действительно являющимися написанными мужчинами (женщинами); полнота (*recall*), т.е. процент текстов авторов-мужчин (женщин) из всех текстов авторов-мужчин (женщин), гендер авторов которых классификатор определил верно.

Нами также вычислялась F1-мера, которая объединяет в себе информацию о полноте и точности модели и представляет собой взвешенное среднее этих показателей:

$$F1 = 2*(Recall * Precision) / (Recall + Precision).$$

Таблица 6

Результаты классификации текстов блогов по гендеру авторов

Метрика / Модель	Точность (Accuracy)	Точность (Precision)	Recall (полнота)	F1
Модель «Контекст»	77%	82%	76%	0,79
Модель «Текст»	80%	84%	79%	0,81

Как показывает таблица, признаки, извлеченные при тематическом моделировании, позволяют получить достаточно высокие результаты классификации текстов по гендеру авторов, что указывает на яркие различия в семантической организации текстов мужчин и женщин в таком жанре, как блог.

Признаки с наибольшим вкладом в модель для классификации текстов блогов по гендеру авторов, отобранные на основе значения индекса Джини (Gini coefficient), представлены в табли-

цах 7–8 (полужирным шрифтом выделены более высокие средние значения вероятности распределения тем в текстах).

Таблица 7

Признаки с наибольшим вкладом в классификатор (модель «Текст»)[1]

Тема	Индекс Джини	Наиболее релевантные леммы	Среднее значение тета в блогах женщин	Среднее значение тета в блогах мужчин
Дом	65.15795	*Мама, говорить, бабушка, любить, спать.*	**0,03684269**	0,02232226
Семья	49.96488	*Дитя, ребенок, семья, мама, муж.*	**0,02340248**	0,01526117
Чувства	38.49610	*Красивый, свой, любимый, любить, новый.*	**0,02361807**	0,01720897
Жизнь	31.88639	*Свой, знать, жизнь, друг, стать.*	**0,05867773**	0,04536635
Россия	30.86174	*Россия, страна, Украина, российский, русский.*	0,01475709	**0,02572685**
Война	29.00843	*Война, военный, армия, бой, время.*	0,0084614	**0,01295493**
Выборы	28.04075	*Выборы, власть, кандидат, депутат, Россия.*	0,01286445	**0,0205552**
Любовь	26.14426	*Друг, любить, самый, человек, жизнь.*	**0,01761532**	0,01356702
Система	24.33733	*Время, самый, система, более, случай.*	0,0362257	**0,04903088**
Бизнес	22.99209	*Компания, рубль, бизнес, деньги, рынок.*	0,01197821	**0,01755713**

[1] Для определения признаков, вносящих наибольший вклад в классификационную модель (feature importances), использовался индекс Джини (Gini coefficient). Были отобраны десять признаков (вероятностей тем) с наивысшим значением данного критерия (столбец 2 табл. 7). Далее приведены средние значения вероятностей тем с наибольшим значением индекса Джини. Чем выше значение индекса Джини, тем более дискриминирующим является признак. Нами были получены высокие значения метрик классификации текстов по гендеру, что связывается нами с исследуемым жанром – блогом, который предполагает конструирование идентичности авторов, в том числе гендерной.

Таблица 8

Признаки с наибольшим вкладом в классификатор (модель «Контекст»)

Тема	Индекс Джини	Наиболее релевантные леммы	Среднее значение тета в блогах женщин	Среднее значение тета в блогах мужчин
Семья	58.45635	*Дитя, свой, мама, друг, жить.*	**0,02321177**	0,01594406
Россия	34.56085	*Россия, страна, российский, власть, русский.*	0,01205163	**0,01742567**
Страна	33.52930	*Страна, война, Германия, Россия, США.*	0,01375826	**0,02221102**
Закон	30.27474	*Закон, суд, область, выборы, право.*	0,01266438	**0,02175396**
Действия	26.43118	*Свой, сказать, говорить, сидеть, пойти.*	**0,04683075**	0,0389084
Работа	24.50501	*Работа, компания, работать, помощь, проект.*	0,01496351	**0,02086453**
Начать	24.09626	*Свой, пытаться, долго, стать, начать.*	**0,07292902**	0,06328536
Время	22.55550	*Время, час, свой, человек, минута.*	**0,01847433**	0,01597144
Система	20.74106	*Процесс, уровень, метод, система, результат.*	0,03825561	**0,05131745**
Рецепт	20.56382	*Масло, вода, город, чай, яйцо.*	**0,02422345**	0,01638216

Сравнительный анализ тем, выделенных моделями «Текст» и «Контекст», показывает, что среди наиболее значимых для классификации тем выявляется много совпадений: для женщин более характерны темы «Семья», «Дом», для мужчин – «Россия», «Система», «Выборы». Между тем тематическое моделирование на уровне контекстов позволяет увидеть более тонкие различия в семантической организации текстов блогов авторов-женщин и авторов-мужчин, например выявить более характерные для женщин темы «Действия», «Начало и попытка действия», уточнить значение темы времени (у женщин она выражена через обозначения времени, у мужчин входит в тему «Система») и т.д.

Основные итоги и перспективы продолжения исследования

Проведенное нами исследование сбалансированного по полу авторов корпуса текстов блогов с использованием методов тематического моделирования показало перспективность применения названных методов для сравнительного исследования особенностей семантической организации текстов мужчин и женщин. Предпринятое нами впервые сравнение результатов применения методов тематического моделирования к матрицам разных видов (терм-документной и матрице совместной встречаемости слов) продемонстрировало комплементарность названных подходов. Их совместное применение позволяет более детально проанализировать особенности семантической организации текстов лиц разных групп, при этом темы, выделенные моделью на основе контекстов (т.е. учитывающей локальную встречаемость слов), являются более интерпретируемыми.

Несмотря на то, что в настоящее время большой популярностью пользуются модели дистрибутивной семантики, методы тематического моделирования также заслуживают внимания, поскольку их неоспоримым и весьма важным для лингвистических исследований преимуществом является интерпретируемость результатов, полученных с их помощью. Наше исследование показало перспективность использования тематического моделирования, учитывающего локальную встречаемость слов.

Направления наших дальнейших исследований связаны с расширением корпуса текстов, номенклатуры исследуемых жанров, а также расширением спектра используемых методов, в первую очередь за счет новых методов, комбинирующих тематическое моделирование и дистрибутивные семантические модели.

Список литературы

Гальченко Д.А. Тематическое моделирование как средство выявления репрезентации бренда территории // Смоленский филологический сборник. – 2021. – № 13. – С. 151–160.

Золотарев О.В., Шарнин М.М., Клименко С.В. Семантический подход к анализу террористической активности в сети интернет на основе методов тематического моделирования // Вестник Российского нового университета. Серия Сложные системы: модели, анализ и управление. – 2016. – № 3. – С. 64–71.

Кавеева А.Д. Онлайн-сообщества мам в социальной сети «ВКонтакте» : структура и тематика // Казанский социально-гуманитарный вестник. – 2018. – № 6(35). – С. 39–42.

Кольцова О.Ю., Маслинский К.А. Выявление тематической структуры российской блогосферы: автоматические методы анализа текстов // Социология: методология, методы, математическое моделирование. – 2013. – № 36. – С. 113–139.

Кудин А.М. Chekhov digital: цифровые методы изучения концептов в текстах произведений А.П. Чехова // Смоленский филологический сборник. – 2021. – № 13. – С. 132–141.

Методы тематического моделирования, их развитие и применение для контента, циркулирующего в региональных онлайн-сообществах / *Телегин Е.Н., Чапурин Е.Ю., Разинкин К.А., Плотников Д.Г., Попов А.В.* // Информация и безопасность. – 2019. – Т. 22, № 3. – С. 325–344.

Митрофанова О.А. Тематическое моделирование корпуса «народных русских сказок А.Н. Афанасьева» // Структурная и прикладная лингвистика. – 2015. – № 11. – С. 146–154.

Пищальникова В.А. Эксперимент как составная часть методологии сопоставительных исследований // Вестник Волгоградского государственного университета. Серия 2: Языкознание. – 2019. – Т. 18, № 3. – С. 52–63.

Потапенко А.А. Семантические векторные представления текста на основе вероятностного тематического моделирования : автореф. дис. … канд. физ.-мат. наук. – Москва, 2018. – 20 с.

Сравнение содержания коллекций научных журналов на основе разработанных тематических моделей и методики Т4С / *Краснов Ф.В., Хасанов М.М., Дементов А.В., Шварцман М.Е.* // Cloud of Science. – 2019. – Т. 6. – № 3. – С. 334–348.

Чижик А.В. Исследование динамики общественного настроения в социальных сетях с использованием методов тематического моделирования // International Journal of Open Information Technologies. – 2021. – Т. 9, № 12. – С. 21–29.

Apishev M.A. Effective implementations of topic modeling algorithms // Труды ИСП РАН. – 2020. – Т. 32, вып. 1. – С. 137–152.

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation // Journal of Machine Learning Research. – 2003. – Vol. 3. – P. 993–1022.

Jones T. textmineR: Functions for Text Mining and Topic Modeling. – 2021. – URL: https://cran.r-project.org/web/packages/textmineR/index.html (дата обращения: 01.04.2022).

Litvinova T. RusIdiolect: A New Resource for Authorship Studies // Lecture Notes in Networks and Systems. – 2021. – Vol. 186. – P. 14–23.

Litvinova T., Sboev A., Panicheva P. Profiling the Age of Russian Bloggers // Communications in Computer and Information Science. – 2018. – Vol. 930. – P. 167–177.

Misztal-Radecka J. Building semantic user profile for polish web news portal // Computer Science. – 2018. – Vol. 19. – P. 307–332.

Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations / *Rangel F., Rosso P., Verhoeven B., Daelemans W., Potthast M., Stein B.* // Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org, Évora, Portugal. – 2016. – Vol. 1609. – P. 1–15.

Quanteda: An R package for the quantitative analysis of textual data / *Benoit K., Watanabe K., Wang H., Nulty P., Obeng A., Müller S., Matsuo A.* // Journal of Open Source Software. – 2018. – Vol. 3(30). – P. 774.

Rodriguez P.L., Spirling A. Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research // The Journal of Politics. – 2022. – Vol. 84(1). – P. 101–115.
Sahlgren M. The Distributional Hypothesis // Rivista di Linguistica. – 2008. – Vol. 20(1). – P. 33–53.
Schmid H. Probabilistic Part-of-Speech Tagging Using Decision Trees // Proceedings of International Conference on New Methods in Language Processing. – Manchester : UK, 1994. – URL: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf

References[1]

Gal'chenko, D.A. (2021). Tematicheskoe modelirovanie kak sredstvo vyjavlenija reprezentacii brenda territorii. *Smolenskij filologicheskij sbornik*, 13, 151–160.
Zolotarev, O.V., Sharnin, M.M., Klimenko, S.V. (2016). Semanticheskij podhod k analizu terroristicheskoj aktivnosti v seti internet na osnove metodov tematicheskogo modelirovanija. *Vestnik Rossijskogo novogo universiteta. Serija: Slozhnye sistemy: modeli, analiz i upravlenie*, 3, 64–71.
Kaveeva, A.D. (2018). Onlajn-soobshhestva mam v social'noj seti «VKontakte»: struktura i tematika. *Kazanskij social'no-gumanitarnyj vestnik*, 6(35), 39–42.
Kol'cova, O.Ju., Maslinskij, K.A. (2013). Vyjavlenie tematicheskoj struktury rossijskoj blogosfery: avtomaticheskie metody analiza tekstov. *Sociologija: metodologija, metody, matematicheskoe modelirovanie*, 36, 113–139.
Kudin, A.M. (2021). Chekhov digital: cifrovye metody izuchenija konceptov v tekstah proizvedenij A.P. Chehova. *Smolenskij filologicheskij sbornik*, 13, 132–141.
Telegin, E.N., Chapurin, E.Yu., Razinkin, K.A., Plotnikov, D.G., Popov, A.V. (2019). Metody tematicheskogo modelirovanija, ih razvitie i primenenie dlja kontenta, cirkulirujushhego v regional'nyh onlajn-soobshhestvah. *Informacija i bezopasnost'*, 22(3), 325–344.
Mitrofanova, O.A. (2015). Tematicheskoe modelirovanie korpusa «narodnyh russkih skazok A.N. Afanas'eva». *Strukturnaja i prikladnaja lingvistika*, 11, 146–154.
Pishhal'nikova, V.A. (2019). Jeksperiment kak sostavnaja chast' metodologii sopostavitel'nyh issledovanij. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Serija 2, Jazykoznanie*, 18(3), 52–63.
Potapenko, A.A. (2018). *Semanticheskie vektornye predstavlenija teksta na osnove verojatnostnogo tematicheskogo modelirovanija* (Unpublished Doctoral dissertation). Moscow: Russian Academy of Science.
Krasnov, F.V., Hasanov, M.M., Dimentov, A.V., Shvarcman, M.E. (2019). Sravnenie soderzhanija kollekcij nauchnyh zhurnalov na osnove razrabotannyh tematicheskih modelej i metodiki T4 C. *Cloud of Science*, 6(3), 334–348.
Chizhik, A.V. (2021). Issledovanie dinamiki obshhestvennogo nastroenija v social'nyh setjah s ispol'zovaniem metodov tematicheskogo modelirovanija, *International Journal of Open Information Technologies*, 9(12), 21–29.

[1] Здесь и далее источники в разделе References оформлены в стиле APA 6th edition.

Apishev, M.A. (2020). Effective implementations of topic modeling algorithms. *Trudy ISP RAN*, 32(1), 137–152.

Blei, D., Ng, A., Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022. Retrieved from https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Jones, T. (2021). *textmineR: Functions for Text Mining and Topic Modeling*. Retrieved from: https://CRAN.R-project.org/package=textmineR

Litvinova, T., Sboev, A., Panicheva, P. (2018). Profiling the Age of Russian Bloggers. *Communications in Computer and Information Science*, 930, 167–177. DOI: 10.1007/978-3-030-01204-5_16

Litvinova, T. (2021). RusIdiolect: A New Resource for Authorship Studies. *Lecture Notes in Networks and Systems*, 186, 14–23. DOI: 10.1007/978-3-030-66093-2_2

Misztal-Radecka, J. (2018). Building semantic user profile for Polish web news portal. *Computer Science*, 19(3), 307–332. DOI: 10.7494/csci.2018.19.3.2753

Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B. (2016). Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. In *Working Notes Papers of the CLEF 2016 Evaluation Labs*. Retrieved from http://ceur-ws.org/Vol-1609/16090750.pdf

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., Matsuo, A. (2018). Quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30), 774. DOI: 10.21105/joss.00774

Rodriguez, P.L., Spirling, A. (2022). Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research. *The Journal of Politics*, 84(1), 101–115. DOI: 10.1086/715162

Sahlgren, M. (2008). The Distributional Hypothesis. *Rivista di Linguistica*, 20(1), 33–53. Retrieved from https://www.italian-journal-linguistics.com/app/uploads/2021/05/Sahlgren-1.pdf

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of International Conference on New Methods in Language Processing*. Retrieved from https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf