СОВРЕМЕННОЕ СОСТОЯНИЕ И ТЕНДЕНЦИИ РАЗВИТИЯ СОЦИАЛЬНОЙ ПСИХОЛОГИИ

ТЕМАТИЧЕСКИЙ АНАЛИЗ ДИСКУССИЙ С ПРИМЕНЕНИЕМ МЕТОДА ЛАТЕНТНОГО РАЗМЕЩЕНИЯ ДИРИХЛЕ*

©2022 г. Н.А. Алмаев*, О.В. Мурашева**

*Доктор психологических наук, профессор, ведущий научный сотрудник, лаборатория психологии речи и психолингвистики, Федеральное государственной бюджетное учреждение науки Институт Психологии РАН, г. Москва, улица Ярославская, 13к1; e-mail: almaev@mail.ru

**Кандидат психологических наук, научный сотрудник, там же; e-mail: murashevaov@yandex.ru

DOI: 10.38098/ipran.sep_2022_25_1_03 **Поступила в редакцию 20 декабря 2021 г.**

Аннотация. Проведена оценка возможности приложения Латентного размещения Дирихле (Latent Dirichlet Allocation, LDA) к анализу дискуссий в «Живом Журнале» (ЖЖ) на примере комментариев пользователей в трех блогах по проблемам ковид-диссидентства и антиваксерства за ноябрь 2021 г. с тэгами «коронавирус», «covid-19». Алгоритм LDA был реализован в экосистеме языка Python в составе пакетов *scikitlearn*. Для автоматизированной обработки данных использовался интернет-ресурс ЖЖ, формат которого способствует откровенности высказываний, что и требуется для изучения мотивации посредством контент-анализа текстов обсуждений. Парсинг содержания осуществлялся в отношении непосредственно HTML страниц ЖЖ, без использования API, что представляется важным для тех интернет площадок, у которых API отсутствует либо малофункционален. Полученные результаты показали чувствительность LDA к содержанию тем и способность отражать их близость. На основе однозначных биграмм могут быть созданы рекомендаторы или автоматические резюме. Однако при поисках глубинной мотивации антиваксерства и ковид-диссидентства в самих темах обнаруживается много информационного шума, случайных биграмм с низкой содержательностью, не интерпретируемых вне контекста предложения. Причина этого в стохастическом подходе выделения слов в документе – «мешок слов». Для дальнейшего смыслового наполнения данной методики представляется целесообразным перейти к выделению суждений: необходимо включить уровень синтаксического разбора предложения в первый этап обработки текста – токенизацию, и передавать на дальнейшую векторизацию коллекции суждений, т.е. биграммы, связанные отношением субъект-предикат.

Ключевые слова: тематический анализ, латентное размещение Дирихле, ковид-диссидентство, антиваксерство, парсинг, мотивация, контент-анализ, социальные сети, биграммы, коллекции суждений.

* Работа выполнена по Госзаданию № 0138-2022-0004.

Развитие машинного обучения (Machine Learning, ML) в области обработки естественного языка (Natural Language Processing, NLP) в последние годы характеризуется значительными техническими достижениями. Их основой выступает технология векторизации, т.е. придания отдельным словам численных значений, характеризующих частоту совместного употребления этих слов, полученных на основе обработки значительных корпусов текстов. Благодаря данной технологии, а также основанной на ней технологии BERT, при которой вектора рассчитываются не только для отдельных слов, но и для целых выражений и даже предложений, намного улучшился автоматический перевод, классификация текстов с помощью нейросетей (как с супервизией, так и без нее), и даже боты стали коммуницировать несколько более «осмысленно». Вместе с тем, данные успехи можно охарактеризовать, скорее, как научно-практические, они в большей степени связаны с применением знания, а не с получением его. Ввиду не только научной, но и социальной важности возникающих в данной связи вопросов, следует отнестись и к самой данной практике. В рамках ее сам факт обучения нейронной сети зачастую предлагается воспринимать как некое самостоятельное достижение, затеняя тем самым вопрос, чему именно она обучилась?

Например, из недавнего доклада в стенах президиума РАН: «Современные модели ИИ пока затронули только анализ фейков. Заняться СМИ им мешает то, что когда дело касается более сложных способов манипуляции общественным сознанием и приемов информационной войны, возникают нетривиальные лингвистические конструкции, которые требуют серьезной экспертизы со стороны гуманитариев – лингвистов, политологов, журналистов, социологов, психологов. Мы хорошо понимаем, как решать эти задачи, если есть хорошо сделанная разметка текстов, – объясняет суть

проблемы Константин Воронцов»[5]. В контексте выступления сказанное звучит, как предложение усилить идеологическую цензуру посредством ботов, обученных на материале деятельности цензоров-людей; но проблема гораздо шире и распространяется также и на область собственно науки. Адекватность разметки контент-анализа должна быть экспериментально проверена (ведь какие-то утверждения в ее рамках делаются на основании общих соображений, какие-то наследуются от предшественников и т.п. (подробнее см.: Алмаев, Градовская 2002; Алмаев и др., 2016), а не просто приниматься на веру. Без экспериментальной проверки любое экспертное мнение, не более чем просто мнение, либо гипотеза. Причем бремя доказательства адекватности системы (как, во-первых, контент-анализа, так и, во-вторых, его реализации средствами ML) лежит на том, кто ее предлагает.

Чем же технология векторизации и различные подходы на ее основе могут помочь собственно науке, т.е. быть использованы для получения новых знаний, в частности, в области социальной психологии? Изучение общественно значимых явлений, находящих свое выражение в дискурсах и нарративах значительного количества людей, требует обработки внушительных объемов текстовой информации, ведь обычно каждый участник дискуссии генерирует далеко не одно сообщение по теме. Соответственно, актуальным является применение к текстам методов статистики, обеспечивающих уменьшение количества переменных, подобных факторному и кластерному анализам, но адаптированных для текстовых сообщений. Одним из таких методов является Латентное размещение Дирихле (Latent Dirichlet Allocation, LDA), алгоритм которого впервые был предложен Дэвидом Блэйем с соавторами (Blei et al., 2003). LDA применяется в основном в англоязычной IT сфере для тематического анализа новостей и разработки рекомендаторов (программ,

[5] http://www.ras.ru/news/shownews.aspx?id=3ee5718a-2dc0-4857-b1c0-2fe90f872f7d#content

находящих материалы по определенным темам и предоставляющих ссылки на них). Наблюдаемые появления слов в корпусе документов рассматриваются в LDA как проявление некоторого числа латентных тем, которые надо реконструировать. Авторы подчеркивают, что данная задача не имеет однозначного решения, тем не менее, его можно аппроксимировать благодаря повторению ряда шагов, сводящихся к нахождению минимумов неравенства Куллбака-Лейблера (Kullback-Leibler) с помощью байесовских методов. Фактически, полученное в начале процесса предварительное решение в ходе выполнения алгоритма постоянно проверяется и уточняется на новом материале вплоть до достижения минимального расхождения в указанном выше неравенстве. Таким образом, в основе LDA лежит весьма продвинутый алгоритм оптимизации, что делает его предпочтительным по точности и консистентности результатов по сравнению с родственными методами LSA (latent semantic analyses – латентный семантический анализа) и NMF (Nonnegative matrix factorization – неотрицательная матричная факторизация). При этом сам подход к выделению тем чисто стохастический, документ понимается, как «мешок слов» («bag of words»). Алгоритм LDA был реализован в таких средах как R, MATHLAB, а также в экосистеме языка Python, в последней – в составе пакетов *scikitlearn* и *gensim*. Мы пользовались скриптами программ, входящих в состав *scikitlearn*[6] (Pedregosa et al., 2011).

Хотя сам по себе алгоритм LDA применим к любым наборам текстов, для решения проблем, связанных с контекстом, в частности, полисемии, вектора должны быть получены на основе обработки значительных корпусов текстов, а не просто взяты из текущей выборки. Это порождает чисто технические проблемы совместимости различных версий пакетов как с корпусами предварительно размеченных на русском языке текстов, так и между собой.

[6] https://scikit-learn.org/stable/modules/decomposition.html#latent-dirichlet-allocation-lda

Поскольку разработчики независимы друг от друга и постоянно совершенствуют свой продукт, версии часто не стыкуются (причем далеко не все проблемы решаются автоматически)[7].

Задача

Поскольку нам не удалось обнаружить применение алгоритмов LDA в исследовательских целях к русскоязычным текстам, основной задачей данной работы является первичная оценка приложимости указанного метода к анализу дискуссий в социальных сетях. Прежде всего, необходимо выяснить, насколько адекватно могут быть выделены с его помощью темы дискуссий, насколько полученные решения отличаются между собой для различных корпусов обсуждений, насколько устойчивы при различном количестве тем (в LDA количество тем задается априори). И главное, – насколько хорошо темы поддаются содержательной интерпретации.

МЕТОДИКА

Выбор социальных медиа

Обычно для автоматизированной обработки данных используют те социальные медиа, которые предоставляют более развитый прикладной программный интерфейс (Application Program Interface, API); это, прежде всего, Facebook (ФБ) и Twitter (Бонцанини, 2018). Однако ФБ характеризуется сильной цензурой, причем, двоякого рода. Во-первых, большинство участников фигурируют там под своими настоящими именами, следовательно, для многих людей запись в ФБ равносильна публичному высказыванию от своего имени, некому заявлению. Во-вторых, руководство ФБ проводит жесткую цензурную

[7] В приведенном тексте нашей программы (блокнот Jupyter) https://yadi.sk/d/5_bbhGPAfKJZ8g в комментариях к установке пакетов указаны их версии, надо установить в активное окружение именно их, иначе программа может не работать.

политику и может заморозить аккаунт на 30 дней и более, если на какие-либо сообщения приходят жалобы со стороны других пользователей. Фигурирующий в нашем исследовании А. Рощин систематически размещает свои тексты параллельно и в livejournal (ЖЖ), и в ФБ и в последнем регулярно подвергается банам. Количество откликов на его статьи в ФБ примерно в 10 раз меньше, чем в ЖЖ. Для Twitter характерно использование коротких сообщений, соответственно, значительную роль в ней приобретают ссылки на разного рода сторонний контент (в частности, визуальный), обработка которого в настоящее время далека от унификации. В свою очередь ЖЖ характеризуется анонимностью (в смысле практической невозможности установить личность блоггера по его нику), к тому же во многих журналах разрешены анонимные (в буквальном смысле) комментарии, практически полным отсутствием централизованной цензуры (пользователи сами банят друг друга при необходимости) и своеобразной системой рейтинга, привязанной к публикационной активности, побуждающей писать больше. Такая система способствует откровенности высказываний, что и требуется для изучения мотивации посредством контент-анализа текстов обсуждений. Кроме того, занимаясь данной темой, мы имели целью выработать методику парсинга (автоматизированного сбора общедоступного контента в Интернете) и обработки данных легко приложимую к анализу содержания текстов различных дискуссионных площадок, могущих вообще не предоставлять никакого API, но иметь ценность с точки зрения социальной психологии, а вовсе не одной только ЖЖ.

События последних двух лет, связанные с пандемией COVID-19, позволили обнаружить такие социально значимые проблемы, как ковид-диссидентство и антиваксерство. Антиваксерство – это общественное движение, которое известно еще со времен разработки первых вакцины (конец XVIII – начало XIX вв.). Его представители аргументируют в пользу

бессмысленности и опасности различных, а иногда и вообще любых вакцин. В данном контексте имеется ввиду агитация только против прививок от COVID-19 штаммов до Омикрона (исследование проводилось до его обнаружения). Ковид-диссидентство – понятие, которое вошло в обиход в период современной пандемии. Ковид-диссидент – это человек, который в той или иной степени не верит в существование коронавируса и отрицает серьезность угрозы от заражения вирусом Sars-Cov-2 (штаммами до Омикрона). Описанные выше явления оказывают влияние на отношение общественности к последним достижениям в области медицины и здравоохранения. Исследования общественного мнения о COVID-19 проводятся по всему миру[8]. В связи с социальной значимостью указанных проблем были взяты обсуждения в журналах трех различных пользователей ЖЖ за ноябрь с тэгами «коронавирус», «covid-19». А. Рощин (ник lj.sapojnik, 6 место в общем рейтинге) занимает ярко выраженные ковид-диссидентские и антиваксерские позиции, в ЖЖ его блог едва ли не главное прибежище адептов и того, и другого. А. Колыбанов (lj.kolybanov, 63 место) сам перенес ковид, как и многие его знакомые, характеризуется умеренным, реалистичным подходом, критикует антиваксерство и ковид-диссидентство, однако далек от оскорблений и радикальных призывов. Б. Рожин (lj.colonellcassad, 2 место) – тема ковида находится на периферии его интересов, он занимает по ней позицию близкую к официальной. Его журнал взят в качестве «фона» ввиду огромного количества комментариев под каждым постом. С содержанием обсуждений у А. Рощина (lj.sapojnik) и А. Колыбанова (lj.kolybanov) мы были знакомы ранее, а с содержанием обсуждений у Б. Рожина (lj.colonellcassad), нет.

Парсились (непосредственно в формате HTML) и обрабатывались только комментарии, но не исходные тексты упомянутых блоггеров. Для наших задач

[8] https://wapor.org/resources/covid-19-public-opinion-research/

парсинг был осуществлен с помощью программы «Content Downloader» в варианте «Ultima», включающем специальный блок для обработки событий на Java Script (JS), позволяющий имитировать действия пользователя (разворачивать комментарии, ожидать их загрузку и т.п.). Программа характеризуется широким функционалом и может быть настроена под парсинг материалов с различных площадок. Результаты сохранялись в формате csv, подвергались дополнительной очистке от тегов, ссылок, и т.п. с помощью макроса VBA Excel, форматировались (для лучшего прочтения «пандами» комментарии разделялись «;»), а также просматривались в полуавтоматическом режиме на предмет лишнего контента (сообщений от известных ботов, ответов пользователей ботам, разного рода остатков форматирования). Разумеется, при необходимости такие действия могут быть осуществлены и с помощью средств самого Python’а, но в любом случае, когда API ресурса недоступен или ограничен в функциональности, не без обработки JS. Следует отметить, что далеко не все комментарии были сгружены в ходе парсинга, хотя на любой странице они полностью раскрывались соответствующим модулем. Проблема надежности парсинга, несомненно, важна для репрезентативности исследования, но в данном случае, в рамках первоначального ознакомления с методом, не принципиальна.

РЕЗУЛЬТАТЫ

Представим результаты для обсуждений каждого из блоггеров от 4 тем и более в таблицах (см. табл. с 1 по 5, рис. с 1 по 5).

Четыре темы антиваксерского дискурса можно охарактеризовать как 1 – демонстративность, 2 – сомнения в ковиде/официальной статистике, 3 – QR коды, 4 – конспирология (см. рис. 1).

Таблица 1.

А. Рощин (lj.sapojnik), 4 темы.

№	Блоггер: А. Рощин (lj.sapojnik). Темы.	Характеристика
1	Topic 0: [('жёлтый звезда', 12.418704379968478), ('звезда грудь', 12.41732628127572), ('россия даже', 7.474331986855239), ('сразу сказать', 7.300053804166196), ('алексей мыслить', 7.183065635080682), ('принять решение', 7.180117273495982), ('сказать надо', 7.1689643823140194), ('митинг протест', 6.988724371575143), ('полностью вакцинировать', 6.8419232311218785), ('год назад', 6.815283257652249)]	Демонстративность
2	Topic 1: [('covid 19', 16.895266416972333), ('какой нибудь', 11.815486937468764), ('вакцинация covid', 8.018434759565928), ('настоящий время', 6.873281412780997), ('слабый вакцина', 6.748671374261476), ('российский статистика', 6.386420111656788), ('сотня тысяча', 5.750198704194731), ('куриный код', 5.714957445727213), ('40 умерших', 5.649864327559168), ('человек умереть', 5.487569791432816)]	сомнения в ковиде/официальной статистике
3	Topic 2: [('qr код', 30.343893405562525), ('куар код', 22.075484934091598), ('три раз', 7.925177362126371), ('общественный транспорт', 7.233624391433668), ('сделать прививка', 6.612092332467021), ('отключить куар', 6.468865831213588), ('отлючили два', 6.468734183995417), ('перешёл улица', 6.467381811930657), ('начать дальше', 6.467049573932777), ('код месяц', 6.467034046386109)]	QR коды
4	Topic 3: [('мировой правительство', 11.159030170749597), ('вакцина грипп', 6.92835416757633), ('со сторона', 5.83625985885718), ('главный вопрос', 5.797127508333894), ('каждый год', 5.727059096906791), ('больной человек', 4.872244431163681), ('сей пора', 4.610409821131893), ('вопрос почему', 4.391474837504981), ('каждый день', 4.193524811094562), ('год через', 4.1060543715004725)]	конспирология

Примечания: по каждой теме (Topic) представлены основные термины в виде двух слов (биграмм) числовое значение биграммы – это ее вклад в общую тему, ближайший аналог факторной нагрузки какого-либо пункта теста в факторном анализе.

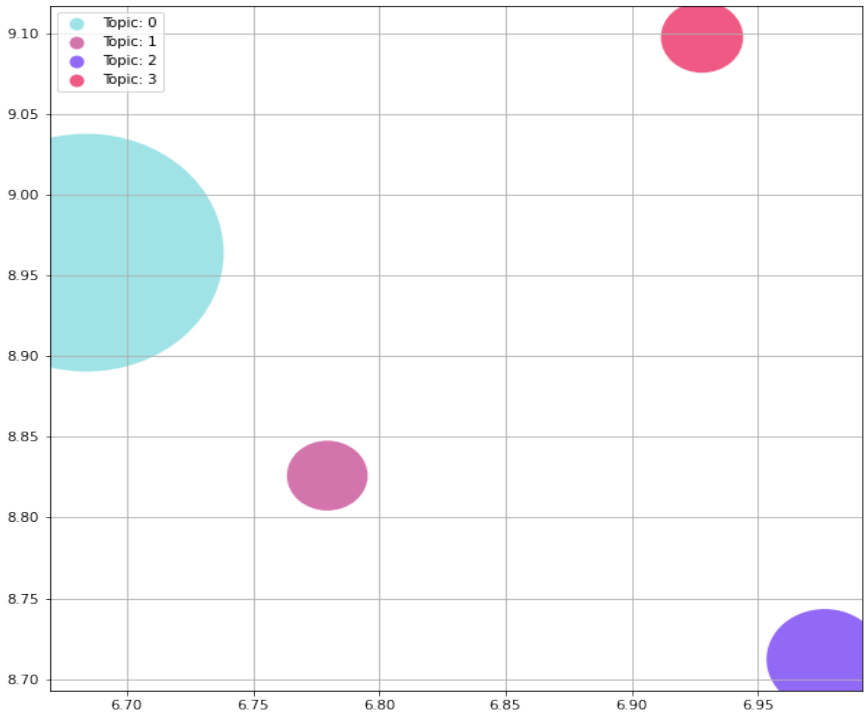

Рис 1. Четыре латентные темы антиваксерского дискурса, ЖЖ А. Рощин (lj.sapojnik)
В обсуждениях журнала А. Колыбанова (lj.kolybanov) темы таковы (см. табл. 2).

Таблица 2.
А. Колыбанов (lj.kolybanov), 4 темы.

№	Блоггер: А. Колыбанов (lj.kolybanov). Темы.	характеристика
1	Topic 0: [('год назад', 7.023554455824928), ('красный зона', 4.212927430394821), ('причина смерть', 4.001136379158566), ('известный человек', 3.7442921574973482), ('стать инвалид', 3.1374493786916546), ('начало пандемия', 3.105665257143985), ('30 год', 3.080565479018121), ('иметь возможность', 3.0612389603608863), ('чуть ли', 3.0519199063973126), ('крайний мера', 3.0236097609879495)]	споры сторонников и противников вакцинации
2	Topic 1: [('расширять кругозор', 6.0345064236919885), ('рано поздно', 5.746595718853801), ('стройный рядами', 4.34672121108201), ('нормальный человек', 3.9902538328846067),('решить вопрос', 3.0601380209967495), ('знать сколько', 2.972432195839799), ('естественный иммунитет', 2.9254542074164083), ('против население', 2.911061391464715), ('население хотя', 2.826711301803102)]	
3	Topic 2: [('миллион человек', 7.729016922427531), ('два год', 6.33086609773991), ('знакомых умереть', 5.377743454269429), ('20 век', 5.067698077877296), ('вряд ли', 5.039015076066029), ('группа риск', 4.521202643131755), ('вакцина защищать',	защита прививок

	4.427941785218852), ('переделывать вакцина', 3.3954475898512566), ('больший часть', 3.3756086361695825), ('часть население', 3.3737125619054753)]	
4	Topic 3: [('данном случай', 5.082635586747307), ('сей пора', 4.196406694562701), ('после вакцинация', 3.983254769789713), ('через месяц', 3.9002279750347917), ('очень много', 3.680146235087448), ('иметь право', 3.6686171308817594), ('вирус именно', 3.4208358480044336), ('год назад', 2.9550244018370915), ('вопрос почему', 2.822863487148012), ('фашизм коммунизм', 2.8047955327141327)]	споры сторонников и противников вакцинации

Первые две темы похожи на споры ваксеров и антиваксеров, третья явно в защиту прививок, четвертая – опять-таки спорная (см. рис. 2).

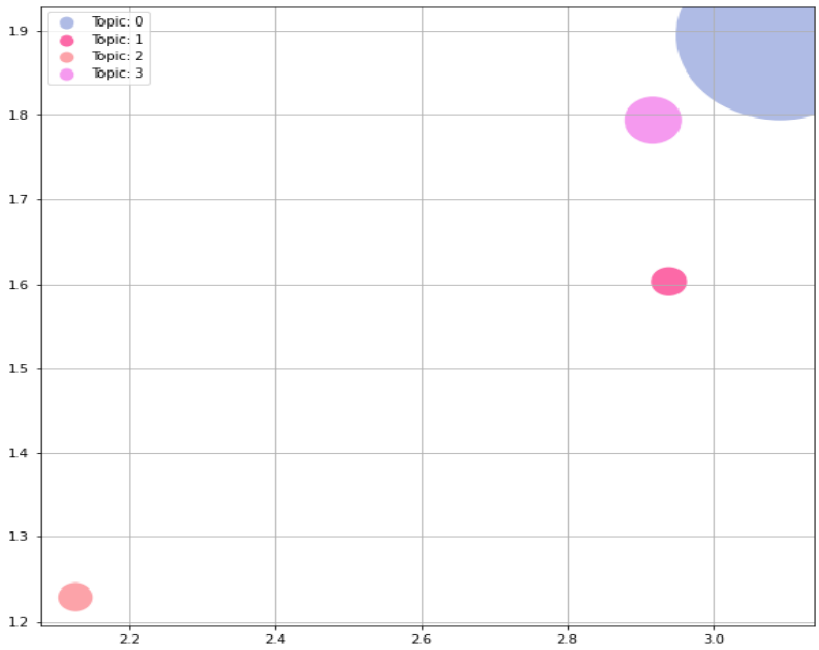

Рис 2. Четыре латентные темы обсуждений с превалированием споров. ЖЖ А. Колыбанова (lj.kolybanov)

Интересно при этом, что четко провакцинная тема (topic 2) находится дальше всего от стартовой.

В обсуждениях журнала Б. Рожина (lj.colonellcassad) начальная тема направлена четко против антиваксеров, а остальные можно интерпретировать как фиксирующие преимущественно спорный контент (см. табл. 3, рис. 3).

Таблица 3.

Б. Рожин (lj.colonellcassad), 4 темы.

№	Блоггер: Б. Рожин (lj.colonellcassad). Темы.	характеристика
1	Topic 0: [('qr код', 21.399527817207463), ('логика факт', 4.2555178005326875), ('вакцина против', 3.7265295089253905), ('новый штамм', 3.2905135675325305), ('вполне интиллигентно', 3.0043057294630464), ('дело антиваксеры', 2.9935319458707363), ('какой дело', 2.99061695144894), ('факт отбиваются', 2.9833077440128717), ('антиваксеры вымереть', 2.978006476908014), ('без маска', 2.975957988824218)]	против антиваксеров
2	Topic 1: [('обязательный вакцинация', 7.826289110106452), ('неделя карантин', 5.305136439769955), ('хронический заболевание', 4.862138559962834), ('пцр тест', 4.83025529421795), ('любой вакцина', 3.747637898008674), ('борьба эпидемия', 3.6968799359982345), ('сидеть дома', 3.6135470138602206), ('ну хотеть', 3.547685509858903), ('заболеть ковидом', 3.5328316071063255), ('метод борьба', 3.2250278520097186)]	споры сторонников и противников вакцинации
3	Topic 2: [('со сторона', 4.93574084446639), ('10 год', 4.0208254013066576), ('какой нибудь', 3.0188900413754745), ('500 тысяча', 1.8832698181513008), ('выступать против', 1.7807799166618472), ('объяснить почему', 1.1262108464275038), ('без прививка', 0.9717656798973411), ('год ну', 0.9657353837543424), ('надо семья', 0.6485386719959304), ('отношение резко', 0.6463490530312364)]	
4	Topic 3: [('год назад', 5.879245748232525), ('достаточно вдумчивый', 3.985633867707237), ('несколько 90', 3.982416175525766), ('всекроме короновирус', 3.979044285441171), ('блог лкт', 3.9751626332858288), ('предпринимать мир', 3.97184373105231), ('африка тоже', 3.971390061872983), ('особенно азия', 3.962312175842014), ('берег ждать', 3.9525071967359344), ('анализ событие', 3.952324495235007)]	

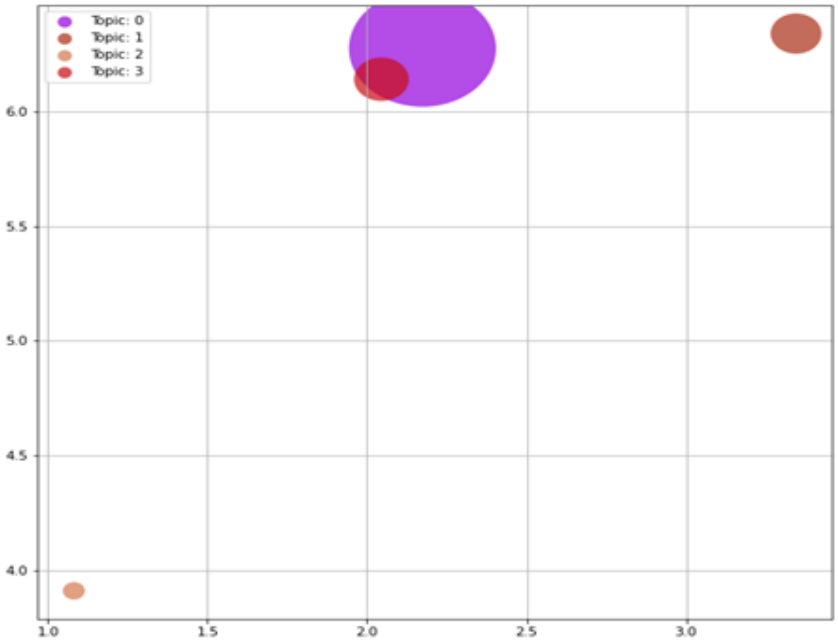

Рис 3. Латентные темы преимущественно провакцинных обсуждений, ЖЖ Кассада по методу LDA

Данное решение можно использовать для сравнения LDA и NMF (последнее совпадает с LSA), для обсуждений в блоге Б. Рожина (lj.colonellcassad) оно будет таковым (см. табл. 4).

Таблица 4.
NMF/ LSA для обсуждений в блоге Б. Рожина (lj.colonellcassad).

№	Блоггер: Б. Рожин (lj.colonellcassad). Темы.	характеристика
1	Topic 0: [('год назад', 0.10387219485657988), ('якобы шенсов', 0.10313423899246597), ('мир всекроме', 0.10313423899246597), ('имвсегда место', 0.10313423899246597), ('какой гарантия', 0.10313423899246597), ('кормить отношение', 0.10313423899246597), ('короновирус моряк', 0.10313423899246597), ('летний старик', 0.10313423899246597), ('лкт радовать', 0.10313423899246597), ('логический анализ', 0.10313423899246597)]	споры сторонников и противников вакцинации
2	Topic 1: [('qr код', 0.8844188105300731), ('обязательный вакцинация', 0.26181804432377065), ('любой вакцина', 0.18472921457905367), ('лучше выйти', 0.15255381591875478), ('без qr', 0.11354056567698306), ('без	споры сторонников и противников вакцинации

	прививка', 0.09958034991140383), ('год ну', 0.09958034991140383), ('хронический заболевание', 0.08959734910587434), ('метод борьба', 0.08552800225152667), ('борьба эпидемия', 0.0853526359924846)]	
3	Topic 2: [('логика факт', 0.45948105437676773), ('дело антиваксеры', 0.30397889271413975), ('антиваксеры вымереть', 0.30397889271413975), ('антиваксеры вполне', 0.30397889271413975), ('интиллигентно опираться', 0.30397889271413975), ('факт отбиваются', 0.30397889271413975), ('вполне интиллигентно', 0.30397889271413975), ('какой дело', 0.30397889271413975), ('опираться логика', 0.30397889271413975), ('пара день', 0.15570264439555923)]	против антиваксеров
4	Topic 3: [('обязательный вакцинация', 0.6738429936828353), ('любой вакцина', 0.36628147100513125), ('хронический заболевание', 0.16472193343655425), ('метод борьба', 0.15308830427639997), ('борьба эпидемия', 0.1522536265497654), ('ну хотеть', 0.15225362423816546), ('вакцинация больший', 0.14810434269225334), ('уровень здоровье', 0.14748669166441786), ('без прививка', 0.1415768233764195), ('год ну', 0.1415768233764195)]	

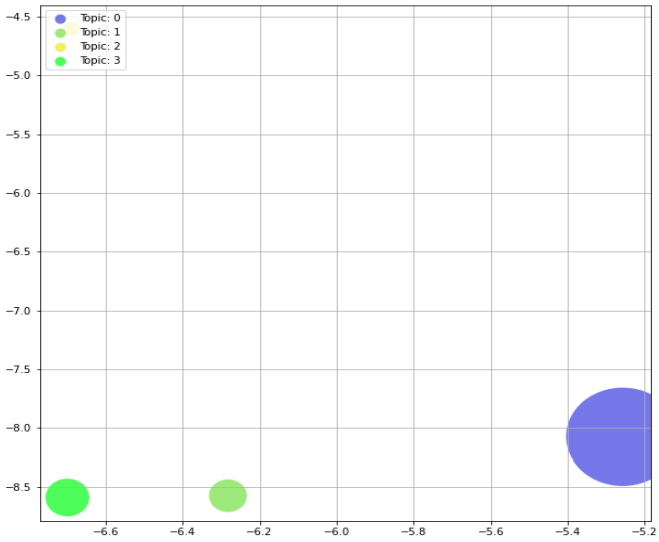

Рис 4. Латентные темы ЖЖ Б. Рожина (lj.colonellcassad), по методу NSF, (совпадает с LSA, тема № 2 находится под легендой)

Темы одного и того же корпуса обсуждений сгруппированы методами LDA и NSF/LSA по-разному, на первый взгляд отличия представляются

большими. Однако при более внимательном анализе нетрудно заметить, что биграммы не только повторяются, но и группируются во многом в те же самые темы, например, тема исходная в LDA решении, оказывается под № 2 в NSF решении и т.д. Тот же эффект сохранения наиболее частотных и устойчивых биграмм наблюдается и при увеличении числа тем (см. рис. 5).

Проиллюстрируем данный тезис восьмитемным решением для обсуждений у А. Рощина (lj.sapojnik), см. табл. 5.

Таблица 5.
Восьмитемное решение для обсуждений у А. Рощина (lj.sapojnik).

№	Блоггер: А. Рощин (lj.sapojnik). Темы.
1	Topic 0: [('qr код', 30.13940377398657), ('какой нибудь', 11.560044142839434), ('общественный транспорт', 7.246711689842176), ('очень похожий', 6.075339715352398), ('прошлый год', 5.477364886935312), ('очень хороший', 4.8594716731297485), ('противник прививка', 4.522930854023622), ('через месяц', 4.300853546509251), ('коллективный иммунитет', 4.300429720552111), ('прививка грипп', 4.230583262526659)]
2	Topic 1: [('covid 19', 16.056080772719028), ('мировой правительство', 11.320580032832119), ('три раз', 7.8622915342108115), ('40 умерших', 7.75181393868818), ('вакцинация covid', 7.387324274507218), ('настоящий время', 6.312977321167426), ('число зверь', 5.891499840196811), ('сотня тысяча', 5.486072586512252), ('человек человек', 5.080671359751267), ('sputnik литва', 4.844575184553542)]
3	Topic 2: [('куар код', 15.71820389981178), ('главный вопрос', 5.8898839321474075), ('сей пора', 4.945225393555085), ('каждый день', 4.350292215782014), ('следующий день', 3.3860878215786143), ('второй вообще', 3.369693913326353), ('участвовать медицинский', 3.267274298846831), ('хотеть получить', 3.2579185636617), ('вакцинировать человек', 3.2310465843853184), ('новый год', 2.506358121273106)]
4	Topic 3: [('куар код', 6.213860019763106), ('посмотреть начальник', 6.213113697451279), ('просто начать', 6.213078416347993), ('митинг отключить', 6.2129892594667595), ('день посмотреть', 6.212957986569074), ('неположенный место', 6.2129463280027055), ('начать дальше', 6.212942899136747), ('код месяц', 6.21294247246381), ('два день', 6.212917636362545), ('место отлючили', 6.21289410750806)]
5	Topic 4: [('раз больше', 5.432233359782512), ('раз чаще', 4.251840448614694), ('русский мир', 3.6376331800822324), ('всякий случай', 3.5628177292369734), ('ваксеров антиваксеров', 3.0865300889349134), ('человек год', 3.0862258611332143), ('сказать где', 2.884802711844654), ('два слово', 2.849395283714704), ('вообще ничто', 2.8390885953827136), ('знать ответ', 2.8058235246854673)]

6	Topic 5: [('слабый вакцина', 7.725325665071258), ('год назад', 6.979389349422735), ('вакцина грипп', 6.656042068167633), ('любой случай', 6.5253830965839805), ('со сторона', 5.667819702726925), ('каждый год', 5.558419704880613), ('куриный код', 5.478428421235608), ('бла бла', 5.416254460311184), ('больной человек', 4.666876519260984)]
7	Topic 6: [('звезда грудь', 12.127745497380134), ('жёлтый звезда', 12.127697605280025), ('россия даже', 7.215313401948817), ('сразу сказать', 7.16607079405715), ('алексей мыслить', 7.044004466645231), ('принять решение', 7.03819372739776), ('сказать надо', 6.999365264589717), ('митинг протест', 6.894827559366418), ('сделать прививка', 6.303316615493169), ('куаризации боец', 6.1272950818415195)]
8	Topic 7: [('полностью вакцинировать', 6.98381684759865), ('российский статистика', 6.230278966086847), ('человек умереть', 5.590392724939588), ('где теперь', 4.6772162947521325), ('самый главный', 4.629528090838196), ('почему тогда', 4.612254267755777), ('право человек', 4.517320116765871), ('там тоже', 4.318914543949402), ('оба сторона', 3.8817221864673965), ('конец конец', 3.8386339556037252)]

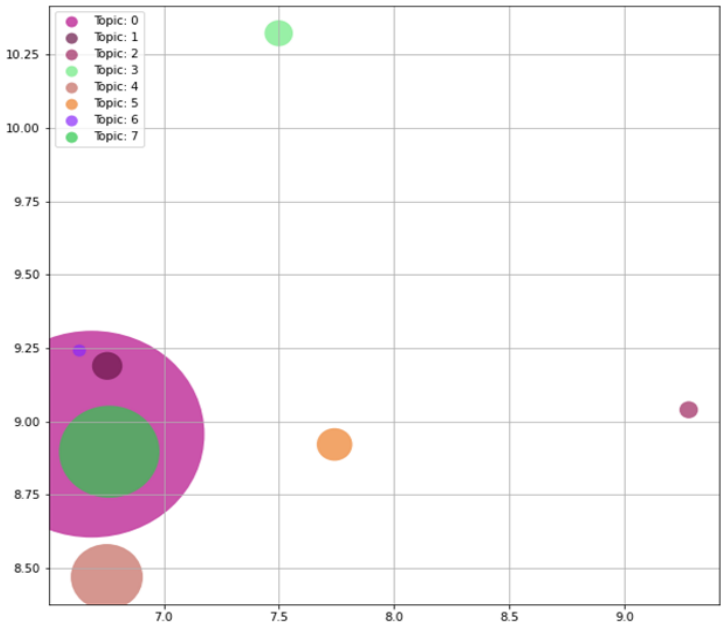

Рис 5. Эффект сохранения наиболее частотных и устойчивых биграмм

По большей части к прежним осмысленно интерпретируемым биграммам добавились устойчивые, но не поддающиеся содержательной интерпретации, например, содержащихся в темах 4 и 7.

То же самое наблюдается и при дальнейшем увеличении числа тем, например, при двадцатитемном решении у А. Рощина (lj.sapojnik) мы обнаружили следующие содержательные дополнения к Topic 1:

- [('covid 19', 16.056080772719028), ('мировой правительство', 11.320580032832119), ('число зверь', 5.891499840196811), ('sputnik литва', 4.844575184553542)]; – а также новую Topic 16: [('через месяц', 4.0762401722145825), ('следственный комитет', 2.869467865387611), ('дельта штамм', 2.670460926980069), ('официальный статистика', 2.6488120628673606), ('австрия израиль', 2.6488069217335335), ('израиль австрия', 2.6479667715155752), ('рано поздно', 2.027439439697083), ('никто даже', 1.9583506788509368), ('смертность ковида', 0.9164467764123355), ('со сторона', 0.0503678484164774 56)].

У А. Колыбанова (lj.kolybanov) при двадцатитемном решении его «фирменная» биграмма «Знакомых умереть», дополнилась следующими подробностями: Topic 17: [('знакомых умереть', 5.267000053788176), ('данном случай', 4.13930251807225), ('через месяц', 3.992244024609062), ('после вакцинация', 2.9034850498394444), ('пара неделя', 1.9883688726871924), ('раз больше', 1.037867103744398), ('семья колесниковых', 1.037864587674608), ('колесниковых под', 1.0378014658084795), ('год следующий', 1.0375626908072264), ('после прививка', 1.0011822287984637)].

У Б. Рожина (lj.colonellcassad) Topic 19: [('азия изза', 4.218694398588728), ('всекроме короновирус', 4.218397927298255), ('отношение резко', 4.21822173720705), ('несколько человек', 4.218153794179898), ('заболеют здесь', 4.218053774599463), ('население понять', 4.217989608359413), ('место судно', 4.217968933166473), ('изза бесконечныхлокдаунов', 4.2179572404799135), ('здесь больше', 4.217921773640806), ('торговый флот', 4.217863225237014)]. И т.п.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

В целом можно сказать, что Латентное размещение Дирихле показало свою эффективность в плане чувствительности к содержанию тем и способности отображать совместную часту их встречаемости. Если субъективно оценивать пользу применения LDA к обсуждениям, то представляется, что в отношении журнала Б. Рожина (lj.colonellcassad),с которым мы не были знакомы, она явна и несомненна, в то время как в отношении журналов А. Рощина (lj.sapojnik) и А. Колыбанова (lj.kolybanov), которые ранее читали, – нет. Действительно, метод в том виде, как он сейчас существует, вполне достаточен для первичного ознакомления с корпусом текстов. Исключительно с его помощью, безо всякого «чтения глазами» мы узнали, что журнал Б. Рожина (lj.colonellcassad) – оплот противостояния антиваксерам и даже познакомились с весьма специфической аргументацией: например, «антиваксеры вымереть» и предлагаемыми способами ее развития в дискурсе – «логика факт», «интелигентно опираться» (сохранена орфография оригинала) и т.п.

При любом количестве тем решения для трех блогов сохраняют свою уникальность и четко отличаются друг от друга. Содержательная идентификация тем возможна благодаря отдельным однозначным биграммам. Далее на основе полученных факторов могут быть, например, созданы рекомендаторы, которые (на основе косинусиального сходства) будут находить соответствующее содержание в интернете и поставлять пользователю ссылки. Наиболее содержательные из полученных терминов и биграмм могут трактоваться как ключевые слова и быть использованы для автоматического создания резюме, например, на основе конкорданса – выделения всех предложений со встречающимся словом.

Однако при поисках глубинной мотивации антиваксерства и ковид-диссидентства обнаруживается в основном «пустая порода». В самих темах много информационного шума, более или менее случайных биграмм с низкой содержательностью, не интерпретируемых вне контекста предложения и представляющих собой то ли субъекты без предикатов, то ли предикаты без субъектов. Причина этого, как думается, в чисто стохастическом подходе «мешка слов». Причем если в аналитических языках соседствующие слова, образующие биграмму, скорее всего, сказывают одно о другом, то в синтетических языках с почти произвольным порядком следования слов это не так.

Для дальнейшего смыслового наполнения данной методики представляется целесообразным перейти к выделению суждений в качестве первого уровня анализа текста и последующую обработку строить уже на основе суждений. Эта идея была высказана еще в (Almayev, 2019) безотносительно каких-либо конкретных средств программной реализации, и данный практический опыт применения LDA вполне ее подтверждает. Другими словами, следует включить уровень синтаксического разбора предложения уже в первый этап обработки текста – токенизацию, и передавать на дальнейшую векторизацию уже не коллекции слов, а коллекции суждений, т.е. биграммы, связанные отношением субъект-предикат. Иначе говоря, перейти от «мешка слов» к «мешку суждений». В принципе, существующие программные средства позволяют это сделать, основные корпуса русскоязычных текстов размечены с указанием частей речи, но, конечно, это отдельная задача, как в отношении разработки алгоритмов, так и в плане совместимости пакетов и потребления вычислительных ресурсов.

ВЫВОДЫ

1. Латентное размещение Дирихле может быть применено к задаче тематического анализа дискуссий: метод достаточен для первичного ознакомления с корпусом текстов без прочтения самого текстового корпуса исследователем.
2. С помощью LDA возможно релевантное выделение тем, которые сохраняют свою уникальность даже при значительном варьировании их количества.
3. Благодаря отдельным однозначным биграммам возможна содержательная идентификация тем.
4. Так как сам подход к выделению слов чисто стохастический (текст – это «мешок слов»), при поисках глубинной мотивации антиваксерства и ковид-диссидентства в самих темах обнаруживается много информационного шума, случайных биграмм с низкой содержательностью, не интерпретируемых вне контекста предложения (субъекты без предикатов, предикаты без субъектов).
5. Для дальнейшего смыслового наполнения данной методики представляется целесообразным включить уровень синтаксического разбора предложения в первый этап обработки текста – токенизацию, и передавать на дальнейшую векторизацию коллекции суждений (биграммы, связанные отношением субъект-предикат).

ЗАКЛЮЧЕНИЕ

В ходе исследования была оценена применимость алгоритмов LDA для тематического анализа русскоязычных текстов и получения новых знаний при изучение общественно значимых явлений (ковид-диссидентство, антиваксерство) в дискурсах и нарративах значительного количества людей.

Результаты работы свидетельствуют об адекватности выделения тем в корпусах текстов, об их отличии друг от друга и их устойчивости. Однако для дальнейшей работы в данной области при поисках глубинной мотивации можно считать целесообразным переход к выделению суждений (биграмм, связанных отношением субъект-предикат) в качестве первого уровня анализа текста.

СПИСОК ЛИТЕРАТУРЫ

Алмаев Н.А., Градовская Н.И. Субъективное шкалирование и конетнт-анализ в оценке эмоционально- аффективной компоненты дискурса // Психологические исследования дискурса. М.: ПерСэ, 2002. С. 18-39.

Алмаев Н.А., Мурашева О.В., Бессонова Ю.В., Киселева Н.И. Обобщенные шкалы контент-анализа проективных рассказов теста социальной мотивации (ТСМ). Описание и критериальная валидность. Часть 1. // Экспериментальная психология. 2016. Т. 9. № 4. С. 90-104. DOI:10.17759/exppsy.2016090408.

Бонцанини М. Анализ социальных медиа на Python. М.: ДМК пресс, 2018.

Almayev N. Content Analyses for the Psychological Purposes: Requirements to Software Supporting Tools // Proceedings of the 14th International Conference on Interactive Systems: Problems of Human-Computer Interaction. Ulyanovsk, Russia, September 24-27, 2019. P. 249-254. URL: http://ceur-ws.org/Vol-2475/short8.pdf (accessed 05.12.2021).

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

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2011.12(85):2825-2830.

BIBLIOGRAFICESKIJ SPISOK

Almaev N.A., Gradovskaya N.I. Sub"ektivnoe shkalirovanie i konetnt-analiz v otsenke emotsional'no- affektivnoi komponenty diskursa // Psikhologicheskie issledovaniya diskursa. M.: PerSe, 2002. S. 18-39.

Almaev N.A., Murasheva O.V., Bessonova Yu.V., Kiseleva N.I. Obobshchennye shkaly kontent-analiza proektivnykh rasskazov testa sotsial'noi motivatsii (TSM). Opisanie i kriterial'naya

validnost'. Chast' 1. // Eksperimental'naya psikhologiya. 2016. T. 9. № 4. C. 90-104. DOI:10.17759/exppsy.2016090408.

Bontsanini M. Analiz sotsial'nykh media na Python. M.: DMK press, 2018.

Almayev N. Content Analyses for the Psychological Purposes: Requirements to Software Supporting Tools // Proceedings of the 14th International Conference on Interactive Systems: Problems of Human-Computer Interaction. Ulyanovsk, Russia, September 24-27, 2019. P. 249-254. URL: http://ceur-ws.org/Vol-2475/short8.pdf (accessed 05.12.2021).

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

Pedregosa, F. Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É.. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2011.12(85):2825-2830.

THEMATIC ANALYSIS OF DISCUSSIONS USING THE LATENT DIRICHLET ALLOCATION**

N.A. Almaev*, O.V. Murasheva**

*Sc.D. (psychology), professor of RAS, leading research fellow; laboratory of psychology of speech and psycholinguist, Federal State Financed Establishment of science Institute of psychology, Russian academy of sciences; 13, , Yaroslavskaya str., Moscow, 129366; e-mail: almaev@mail.ru

**Ph.D. (psychology), research fellow; the same place; e-mail: olgalogatskaia@gmail.com

Summary. An assessment of the application of Latent Dirichlet Allocation (LDA) to the analysis of discussions in the LJ was carried out using the example of user comments in three blogs on the problems of covid-dissidence and anti-vaxing in November with the tags "coronavirus," "covid-19." The LDA algorithm was implemented in the Python language ecosystem as part of the scikitlearn packages. LJ ("Live Journal") was used to collect test cases and further automated data processing. The LJ system contributes to the frankness of statements, which is required to study motivation through content analysis of discussion texts. A methodology for parsing and data processing was developed, attached to the analysis of the content of texts of various discussion platforms. Parsing was carried out in relation to HTML pages of LJ, without using the API, which seems important for those Internet sites where the API is missing or low-functional. The results showed the effectiveness of LDA to the content of topics and the ability to reflect their proximity. On the basis of unambiguous bigrams, advisors or automatic summaries can be created. However, when looking for the deep motivation of the anti-vaxing and covid-dissidence movement, a lot of information noise is found in the topics themselves, random bigrams with low content that are not interpreted outside the context of the sentence. The reason for this is the stochastic approach of highlighting words in the document - "bag of words." For further semantic content of this technique, it seems advisable to move to the selection of judgments: it is necessary to include the level of parsing of the sentence in the first stage of text processing - tokenization, and transfer to further vectorization the collection of judgments, i.e. the bigrams associated with the subject-predicate relationship.

Keywords: thematic analysis, Latent Dirichlet Allocation, covid-dissidence, anti-vaxing, parsing, motivation, content analysis, social networks, bigrams, judgment collection.

** The study was fulfilled in accordance of the state assessment № 0138-2022-0004.