Исследование динамики общественного настроения в социальных сетях с использованием методов тематического моделирования

А.В. Чижик

***Аннотация*— Цифровые технологии обусловили формирование нового уровня социокультурного пространства, который выразился в перманентном присутствии в повседневности феномена виртуальной реальности. Она является мотивирующим инструментом социальных и политических преобразований общества, которые заключаются в ускорении ритма причинно-следственной связи «общественное настроение→общественное мнение→социальное действие массы». Понимание изменившегося темпоритма социальных явлений актуализирует проблему формирования систем детекции текущего общественного настроения, что может явиться базисом обратной связи между властью и обществом. То есть мониторинг общественного настроения как реакции на социальную и политическую ситуацию.**

В статье рассматриваются две техники тематического моделирования – LDA и LSA. Они применяются к набору новостных анонсов, которые были опубликованы в период 2020 и 2021 года в социальных сетях официальных средств массовой информации.

Целью проводимого анализа является обнаружение наиболее освещаемых в СМИ тем. На следующем этапе проводится анализ динамики обсуждения этих тем в неформальных публичных беседах пользователей социальных сетей.

Таким образом, тематическое моделирование используется как прикладной метод, при этом на первый план выступает задача качественного выделения тематических кластеров, так как в дальнейшем это повлияет на репрезентативность выводов об общественном настроении.

***Ключевые слова*— тематическое моделирование, векторная модель, латентный семантический анализ, латентное размещение Дирихле, LDA, LSA.**

Постановка проблемы

Большая часть социальной реальности находится за пределами повседневной жизни человека, однако индивиду требуется быть осведомленным о происходящих событиях, более того требуется сформировать индивидуальное отношение к проблемам социального, политического и культурного характера, примкнув тем самым к той или иной социальной группе. В качестве основного инструмента, позволяющего получить сведения о социальной реальности, в таком случае выступают средства массовой коммуникации, это значит, что основных источников информации – три: СМИ (в контексте использования их основного канала коммуникации), социальные медиа в формате коммуникации «один ко многим» (например, посты в официальных группах СМИ, публикации лидеров мнений и т.д.) и отдельные пользователи ИКТ (коммуникация «один к одному», например, репост новости в личном сообщении). Очевидно, что информационный повод перетекает из одного типа источника в другой достаточно оперативно, при этом «обрастая» новыми деталями и теряя постепенно нейтральный контекст (если таковой и был изначально). При этом в силу удобности коммуникативного акта посредством социальных сетей обсуждение информационных поводов (а часто и само получение информации) происходит именно в них. В таких условиях индивид не может оценить сведения на предмет достоверности и находится в режиме доверия или недоверия к своему источнику информации. Основная цель при получении информации перетекает из области когнитивного в сферу эмоционального: получая новость, пользователь социальной сети должен оперативно решить, к какой позиции он примкнет в обсуждении события. Отметим, что концепция обсуждения вроде бы не является обязательной при получении информации, однако социальные медиа в целом провоцируют индивида к ответу (в том числе публичному). Следовательно, проектирование картины реальности в сознании индивидов переходит от средств массовой информации (как это было на предыдущем этапе функционирования постиндустриального общества) к социальным группам. Значит, сложные социальные явления окружаются комплексом стабильных ассоциаций и стереотипов благодаря коммуникативному действию, и в наибольшей степени за счет его протекания в онлайн-среде.

Исходя из основных моделей коммуникации внутри социальных медиа, очевидно, что начальная точка формирования эмоционального поля вокруг социального процесса – индивидуальный

Статья получена 20 ноября 2021.

Чижик Анна Владимировна, кандидат культурологии, Санкт-Петербургский Государственный Университет, старший преподаватель, ORCID 0000-0002-4523-5167 (a.chizhik@spbu.ru)

Статья подготовлена по итогам выступления на Международной объединённой конференции «Интернет и современное общество» (IMS-2021).

эмоциональный отклик, то есть индивидуальное настроение. Без поддержки социальной группой оно оказывается быстро угасающим явлением, не влияющим на социокультурные тренды общества. Однако, как только индивид получает одобрение от окружающей среды можно говорить о формировании устойчивых суждений на уровне некоторой социальной группы, и прежде всего маркером такой ситуации выступает социальное настроение. Б.Ф. Поршнев отмечал, что оно, являясь складывающимся из индивидуальных настроений целым, характеризует социум во всем его многообразии.

В качестве типов социального настроения можно выделить групповое, политическое и массовое [1], каждый из них влияет на состояние общества; крайними формами его отражения быть революционные действия, митинги и восстания. «Как правило, социальное настроение – это эмоциональное отношение к тому, что стоит на пути, кто мешает, или, напротив, кто помогает воплощению желаемого в жизнь» [2]. Это определение обуславливает важный для дальнейшей интерпретации факт: у настроения есть субъект и объект. Значит, возможны две ситуации: 1) индивид, получая из доступного для него источника информацию, стремится ее передать своей социальной группе, наделив ее маркером своего отношения; 2) индивид является созерцателем мнения социальной группы (или нескольких групп и нескольких мнений) об информационном поводе и решает, примкнуть ли к нему посредством вступления в диалог.

Значит, мы можем выделить корень зарождения социального настроения в онлайн-среде, им является информационный повод, транслируемый СМИ. Дальнейшая реакция общества рождается в публичном диалоге индивидов друг с другом. Выходит, актуальная повестка дня социальной группы должна отличаться от той, что позиционируется СМИ.

Чтобы глубже понять, роль социальной группы в формировании социокультурного ландшафта стоит рассмотреть процесс формирования социального настроения через теории коммуникативного действия Дж. Хабермаса и социального действия М.Вебера. Социальное действие по Веберу определяется через два взаимосвязанных компонента – «ориентацию на другого» и «субъективный смысл» [3]. Это значит, что любое действие индивида подчиняется логичным индивидуальным мотивам, то есть является осмысленно ориентированным, при этом оно может быть признано социальным, если процесс соотнесения взаимных ожиданий может быть декомпозирован из общего потока индивидуальной логики.

Влияние концепции социального действия М. Вебера прослеживается в логике структурно-функционального анализа Т. Парсонса, который воспринимает мотивированное поведение человека как ключ к запуску «социальной системы действия» [4], являющуюся базисом процессов социодинамики общества. При декомпозиции поведения индивида Парсонс выделял элементарное действие, которое обязательно включает в себя цель и ситуацию (контекст, в котором индивид действует), а также нормативную ориентацию, которая перекликается с категорией общественно-ориентированного действия, обозначенной М. Вебером. В ней можно выделить два компонента – ценностностные интенции и мотивационную составляющую действия. Наиболее интересной является мотивационная составляющая, в которой можно выделить чувственную, когнитивную и оценочную ориентации на объекты социальной системы.

Развитие идей Вебера и Парсонса легло в основу теории коммуникативного действия Ю. Хабермаса [5, 6], который в качестве цели коммуникации (как социального действия) видел координацию действий индивидов, что означало соотнесение ожиданий на уровне знаний, норм и интенций индивидов внутри социальной группы. Система, выступающая гарантом стабильного функционирования общества, по мнению Хабермаса, стремится аккумулировать в себе смыслы и символы, тем самым подчиняя их себе и подменяя процесс согласования общих целей процессом манипулирования и подчинения. Анализируя интерсубъективные структуры социального опыта субъектов, которые обеспечивают возможность взаимопонимания, ученый отталкивается от понятия «мира повседневной жизни», введенного Э. Гуссерлем [7]. Очевидно, что коммуникативные процессы при этом становятся постоянным фоном социального познания и психологических рефлексий индивида и задают рамки любого взаимодействия. В противопоставление исходному понятию «Lebenswelt», которое предполагало доминирование совокупности психических состояний [8], вслед за Ж.-П. Сартром Хабермас настаивает на присутствии в конструировании интенций субъекта (как своеобразного полюса всех актов сознания) влияния знаний и опыта. Процессы категоризации и регуляторы поведения влияют на структуры общения, которые позволяют в свою очередь сформировать некоторое усредненное представление об объективном мире, являющееся общим для социальной группы, достигшей взаимопонимания между своими членами. Исходя из этого возникает предположение, что постепенная трансформация структуры общения приведет к переформатированию рациональности в коммуникативную.

Обращаясь к современному этапу развития постиндустриального общества, справедливо заметить, что интенсивное развитие ИКТ обусловливает принципиальные и динамические изменения в жизни людей и общества в целом, это означает, что коммуникативное пространство, созданное посредством активного вхождения социальных медиа в повседневную жизнь индивидов, детерминирует формирование новой социокультурной парадигмы и актуализацию идей коммуникативной рациональности уже на почве виртуальной реальности. При этом стоит отметить, что механизмы работы социальных сетей являются изначальным базисом процессов коммуникации, таким образом, относительно идеалистическая концепция Хабермаса в современной формации общества получила новый виток развития на

фоне возможности проведения большого количества коммуникаций в онлайн-среде (что в частности расширяет масштабы коммуникативных актов, а также количество участников активного диалога).

На этом этапе анализа коммуникативного социального действия становится важным ввести еще одно понятие, а именно – социальное пространство, которое поможет обозначить суперпозиции акторов коммуникации относительно друг друга.

С одной стороны, в самом общем смысле под социальным пространством можно понимать сферу взаимодействия индивидов, с другой, П. Бурдье (а также Г. Зиммель [9] или П. Сорокин [10]) выделял важное свойство социального пространства – его структуру, задаваемую статусами социальных акторов. Очевидно, что и в том, и в другом случае социальное пространство зависит от социодинамики общества, в рамках которой коммуникация становится туннелем движения смыслов в социальном времени и пространстве. Значит, формирование социального пространства происходит при активном влиянии коммуникативного пространства. Обращаясь к анализу современной ситуации, мы также можем констатировать смешение виртуальной реальности и мира реального, а это значит, что социальное пространство суть порождение и продолжение коммуникативного. Опорой такого вывода могут служить исследования Н. Лумана, который отмечал, что «элементарный процесс, конституирующий социальное как особую реальность, есть процесс коммуникации» [11]. В таком случае мы можем заключить, что социальная система, которая стремится к упорядочиванию и стабильности, является совокупностью коммуникативных актов, которые создают сеть коммуникации. При этом каждая коммуникация предполагает самоописание, а это значит, что перед нами предстает самореферентная система, которая хоть и имеет связь с окружающим миром, но не детерминирована им, располагая достаточной силой собственных системных процессов (базис которых в социальном настроении, а одно из логичных конечных проявлений – формирующиеся стереотипы и паттерны поведения, одобряемые большинством) для построения структур.

Такой конструкт социума актуализируется и в работах Мануэля Кастельса, который констатирует, что социум обладает «специфической формой социальной организации, в которой благодаря новым технологическим условиям, возникающим в данный исторический период, генерирование, обработка и передача информации стали фундаментальными источниками производительности и власти» [12].

Таким образом, коммуникативное пространство в его современном понимании может быть рассмотрено как образующая основа формирования социального настроения, а социальное пространство может быть определено как коммуникативная конструкция.

Ниже будет приведен анализ дискуссий в социальных сетях (на примере анализа переписки пользователей в публичном Telegram-чате), которые возникали на протяжении годового цикла (с 1 января 2020 г. по 20 марта 2021 г.) в качестве реакции на новостную повестку, формируемую средствами массовой информации. Для того, чтобы проанализировать новостной фон за указанный период, был собран специальный набор данных из социальной сети ВКонтакте, состоящий из постов, которые публиковались на странице официального аккаунта «Ленты.ру». Отметим, что было взято только одно СМИ, так как цель была собрать релевантную заданному периоду информационную повестку, которая транслировалась бы через социальные сети. Для этого было достаточно выделить средне-нейтральное СМИ федерального масштаба, стабильно представленное в русскоязычных социальных медиа. Такой подход к фиксации повестки дня позволил выдвинуть гипотезу о том, что целевая аудитория, получающая новости из крупной социальной сети, является по своему составу (демографическому в первую очередь) аппроксимируемой до состава целевой аудитории публичных чатов из соседней социальной сети. Отметим, что в первую очередь нас интересовали переживание событий и рефлексии, а не первый эмоциональный импульс, получаемый при восприятии информационного сообщения, именно с этой целью набор данных, состоящий из диалогов, был собран не под постами СМИ (что тоже технически возможно), а в нейтральном пространстве. Итогом такого анализа является попытка оценки потенциала онлайн-среды как публичного поля для стихийно возникающих дискуссий и обсуждений с точки зрения ее функциональной значимости в рамках формирования структуры общественной интеграции, через анализ которой можно детектировать тренды социального настроения.

Классической задачей в области обработки естественного языка является тематическое моделирование, цель которого – поиск скрытой структуры данных (создание модели коллекции текстовых документов). Такая модель определяет набор тем, которые содержатся в серии документов, что позволяет рассортировать эти документы по различным тематическим категориям. Поскольку количество тем неизвестно, то эта задача относится к пулу задач обучения без учителя (unsupervised learning): на входе присутствует немаркированный набор данных, алгоритм должен самостоятельно провести их логическую классификацию. Таким образом, тематическое моделирование очень похоже на проблему кластеризации данных. С помощью моделирования тем, по сути, происходит группировка текстов, при этом кластеры приобретают интерпретацию как тематические категории. Основное отличие состоит в том, что, оказываясь в категориях тематического моделирования, приходится перейти от более традиционного евклидова векторного пространства к некоторому абстрактному пространству слов. Методы тематического моделирования можно разделить на две основных группы – алгебраические и вероятностные. Латентно-семантический анализ (LSA) относится к алгебраическим методам, а среди вероятностных наиболее популярным является латентное размещение

Дирихле (LDA). Так как LSA и LDA основаны на очень разных математических процедурах, то, очевидно, что в зависимости от типа входных текстовых данных они будут иметь разную степень успеха. При этом алгоритм использования их в рамках прикладной задачи может быть достаточно схож.

Набор данных

Для эксперимента была собрана коллекция новостных анонсов, опубликованных на официальной странице интернет-издания «Лента.Ру» в социальной сети «ВКонтакте» за период с 1 января 2020 г. по 20 марта 2021 г. Общее количество слов в исследуемой коллекции – 1766152 слов, среднее количество слов в одном анонсе – 11. В необработанном виде новостные анонсы являются серией текстовых строк, сопровождаемых датой публикации (также была собрана техническая информация об опубликованных постах, однако в этом исследовании она не является значимой). Ниже приведен фрагмент сформированного датасета (рис. 1).

	id	date	likes	reposts	views	comments	text
11	4847468	16-03-2021	69	20	10378	49	Администрация Байдена попыталась связаться с С...
12	4847421	16-03-2021	38	17	8613	20	Роскомнадзор обещает заблокировать Twitter в...
13	4847406	16-03-2021	8	5	6194	2	NaN
14	4847399	16-03-2021	20	16	7398	2	Глава Минвостокразвития предложил построить Ti...
15	4847374	16-03-2021	52	15	8319	2	В конце января пользователи Reddit обвели вокр...
16	4847349	16-03-2021	148	92	12015	9	Немного омолаживающих процедур с Таймыра. Свеж...
17	4847320	16-03-2021	33	18	8346	8	Протесты в Мьянме, вид от первого лица. Там во...

Рис. 1. Фрагмент исследуемого набора данных

Так как прикладной целью исследования было выявление корреляции между новостной повесткой в СМИ и формирующимся общественным мнением, то был собран еще один набор данных за тот же временной период, который раскрывал присутствующие в социальных сетях обсуждения обычных людей на предложенные темы. Источником послужил публичный Telegram-чат новостного канала Mash – «MACX», в котором участники на фоне официальных новостей обсуждают текущие события в мире, высказывая свою позицию (выводы), а также дополнительно освещая тему (факты). Этот датасет включил в себя 56171 слов, среднее количество слов в одном сообщении - 13. Формат коллекции – текст и дата его публикации (то есть, аналогичен описанному выше).

Как правило, тематическое моделирование предполагает достаточно длинные текстовые объекты в качестве исследуемой единицы (например, полный текст статьи), это связано с тем, что большее количество слов в документе помогает четче очертить потенциальную тему, а также составить объемный тематический словарь. Однако специфика новостных анонсов позволяет ожидать надежное ядро семантическое содержания за счет лаконичности дискурса, свойственного этому типу журналистского контента. Также необходимо отметить, что общий объем собранных записей обеспечил достаточную глубину анализа данных.

Построение векторного представления текста

Для того чтобы текст было возможно использовать в качестве входных данных в любом алгоритме, необходимо преобразовать языковую сущность (слова, предложения, параграфы или текст в его полном объеме) в набор чисел (числовой вектор). Оперируя такими векторами, в частности, становится легко представить в геометрическом пространстве близость слов друг к другу. Такой вектор называется word embedding. В случае данного исследования процессу векторизации подвергались по отдельности каждый новостной анонс (и далее – каждое сообщение пользователя из второго набора данных). Самый простой подход к векторизации модель «мешка слов» (Bag-of-words model, BoW), в рамках которой пренебрегается порядок слов, составляется словарь присутствующих слов, таким образом, каждое слово становится возможным превратить в вектор по длине такого словаря. Такой вектор показывает, сколько раз каждое слово из словаря встречается в конкретном документе.

Этот способ векторизации называется one-hot-encoding, и, в целом, дает необходимое количество операций над закодированным текстом для того, чтобы его можно было успешно проанализировать. Так, например, можно сложить векторы всех слов в предложении и получить вектор суммы. Также такой набор векторов дает информацию о том, насколько часто в предложении встречаются разные слова. К тому же векторы предложений можно сравнивать между собой. Распространенной альтернативой этого метода является использование статистической меры TF-IDF, которая вычисляет относительную частоту слов в документе по сравнению со всем корпусом, помогая таким образом оценить важность каждого слова внутри конкретного документа. Данный метод полезен при анализе коллекции неравномерных по длине текстов в качестве противодействия большим значениям, которые более длинные документы имели бы по сравнению с короткими, если бы использовались необработанные подсчеты. Однако исследуемый набор данных обладает двумя важными характеристиками, во-первых, в нем присутствуют тексты примерно одинаковой длинны, во-вторых, новостные анонсы являются короткими текстами, поэтому использование метода TF-IDF, скорее всего, приведет не к улучшению, а к снижению качества векторизации в рамках дальнейшей задачи тематического моделирования. Поэтому для этого исследования был выбран наивный BoW подход, который в конечном итоге принес терм-документную матрицу (document-term matrix), где каждая строка соответствует новостному анонсу, а каждый столбец – отдельному слову. Отметим, что при кодировании текстовой информации из набора данных были отсечены стоп-слова (например, предлоги и союзы) с целью обеспечения большей репрезентативности.

Обзор моделей LSA и LDA

Скрытый семантический анализ (LSA) был предложен для задач тематического моделирования в

2004 году [13].В основе метода лежит идея о том, что слова будут встречаться в похожих частях текста, если они имеют одинаковое значение. На вход в LSA модель поступает матрица, состоящая из m документов и n слов (созданная с применением ранее описанных методов, или любых других способов векторизации текстовых данных). Затем происходит процедура факторизации матрицы: алгоритм раскладывает матрицу по сингулярным значениям, благодаря чему получаются три новые матрицы (на рис.2 отображена суть процесса разложения), линейная комбинация которых является достаточно точным приближением к исходной матрице. Основная идея заключается в том, что матрица (topic matrix), получившаяся при перемножении новых ортогональных, которая содержит только k первых линейно независимых компонент исходной матрицы, отражает структуру зависимостей, которые латентно присутствовали в исходной матрице. Каждая из n строк этой матрицы представляет собой документ, а каждый из первых k столбцов соответствует теме. Тогда (i, j)-тая запись может считаться мерой присутствия темы j в документе i.

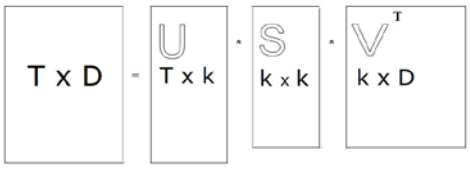

Рис. 2. Разложение матрицы размерности (T X D) на матрицу термов U размерности (T X k), матрицу документов V размерности (k X D) и диагональную матрицу S размерности (k X k), где k – количество сингулярных значении диагональной матрицы S.

Чтобы отсортировать документ по тематической категории, достаточно узнать наибольшее значение каждой строчки (argmax), которое будет соответствовать наиболее широко представленной теме. Отметим, что количество тематических категорий (параметр k), на которое алгоритм будет делить тексты, является задаваемым параметром.

Латентное размещение Дирихле (LDA) было представлено в 2003 году [14].как генеративная вероятностная модель для коллекций дискретных данных. Важный идейный момент LDA заключается в том, что вероятностные модели удобно понимать и представлять в виде порождающих процессов (generative processes), то есть последовательно описывать, как порождается единица данных, а имено каждое слово в документе (указывая вероятностные распределения). Основа метода – предположение о том, что в каждом документе смешаны разные темы, а в каждой теме – присутствует определенное распределение слов. Интуитивно прочитывается два уровня агрегирования: 1) распределение по категориям (к примеру, новости об экономике, политические новости и т.п.), 2) распределение слов внутри категории (например, «деньги» и «акции» актуальны текстах об экономике и финансах). Поэтому документы рассматриваются как распределения вероятностей по скрытым темам, а эти темы - как распределения вероятностей по словам. При этом существует большое количество слов, которые появляются в текстах любой тематики с одинаковой вероятностью. Поэтому удаление стоп-слов и для этого метода – важный шаг реализации алгоритма.

Теоретическое обоснование LDA полагается на использование понятия взаимозаменяемости (теорема де Финетти [15]), используя которую можно получить внутридокументную статистическую структуру через смешанное распределение. Итак, метод предполагает, что процесс порождения каждого слова состоит в том, чтобы сначала выбрать тему по распределению, соответствующему документу, а затем выбрать слово из распределения, соответствующего этой теме. То есть, чтобы отсортировать новостные анонсы по тематическим кластерам, LDA обращается к априорным значениям распределения Дирихле, использует вариационные байесовские методы для вывода скрытых параметров распределения, которые затем характеризуют различные темы.

Как и в случае с LSA, количество тем является гиперпараметраметром, который задается модели на входе. Результат работы алгоритма представляется в форме матрицы, но каждая из строк теперь представляет собой распределение вероятностей, определенное по темам для каждого документа. Поэтому (i,j)-тое значение этой тематической матрицы может интерпретироваться как вероятность того, что заголовок i принадлежит теме j (точнее, как доля слов в заголовке, относящихся к теме j). Для получения оценочной категории темы каждого новостного анонса, необходимо вычислить наибольшее значение каждой строчки.

ОПИСАНИЕ ЭКСПЕРИМЕНТА

Текстовые данные были предобработаны в следующей последовательности: разбиение текстов на токены; удаление спецсимволов, ссылок и пунктуации; удаление стоп-слов и лемматизация токенов. Далее текст был векторизован при помощи CountVectorizer (библиотека scikit-learn), который возвращает закодированные вектора с длиной всего словаря (поэтому векторы разреженные) и информацией, сколько раз каждое слово появилось в документе. Так возникает терм-матрица, которая будет подаваться на вход в оба алгоритма тематического моделирования.

Первым этапом анализа стало выявление наиболее частотных слов в наборе данных (без учета стоп-слов), что дало возможность оценить словарь исходных данных (рис. 3).

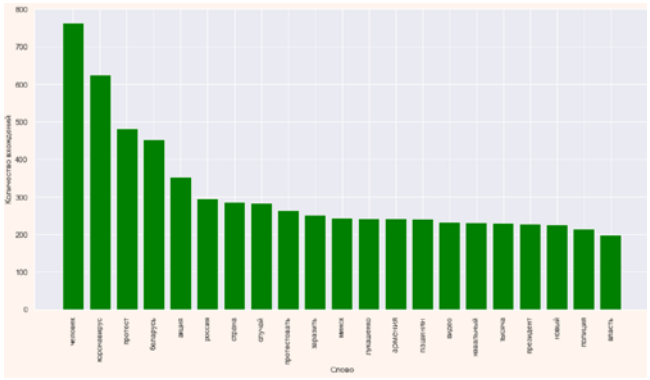

Рис. 3. 20 наиболее встречаемых в наборе данных слов

Получившаяся диаграмма показывает, что, во-первых, проведенной предобработки текстовых данных оказалось достаточно, так как наиболее часто встречающиеся слова выглядят интерпретируемо, во-вторых, интуитивно прочитывается несколько тем.

Латентно-семантический анализ

Модель LSA реализовывается с помощью TruncatedSVD (библиотека scikit-learn). Количество искомых тем было выбрано эмпирически – 20 (это же число будет использоваться и для модели LDA). Взяв argmax для каждого новостного анонса в получившейся матрице, были получены и отсортированы прогнозы тем для всех объектов в выборке. В каждом выделившемся тематическом топике были найдены наиболее часто встречающиеся слова (для более легкой дальнейшей интерпретации) – рис.4.

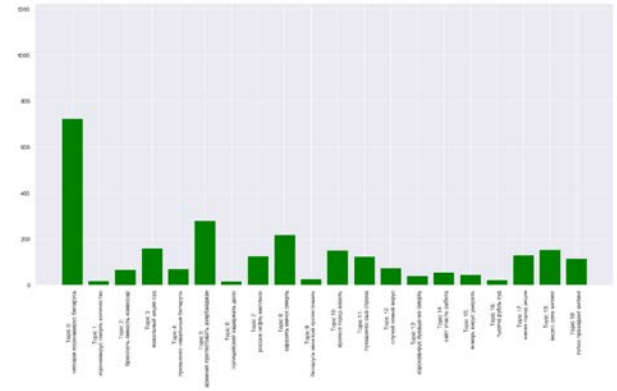

Рис. 4. Результат работы LSA, в каждой найденной теме выделены три наиболее частотных слова для визуализации

На гистограмме видно, что модель LSA в целом определила некоторые темы, которые интуитивно очерчивались и благодаря общему анализу частот слов. При этом распределение тем неравномерно, что свидетельствует о том, что одни темы более распространены, чем другие, в новостных репортажах.

Далее полученные вектора были преобразованы с использованием техники нелинейного снижения размерности t-SNE [16] для их отображения в двухмерное пространство. Таким образом, 20-мерные тематические векторы были сжаты в 2-мерные представления, чтобы выделить кластеры (рис. 5).

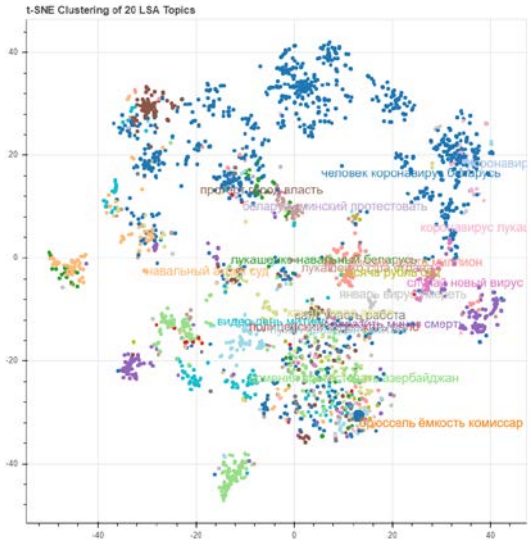

Рис. 5. Визуализация полученных тематических кластеров (модель LSA)

Хотя полученные выше тематические категории казались в целом согласованными, диаграмма рассеяния показывает, что разделение между этими категориями условное, есть участки, где кластера накладываются друг на друга. К такому результату в частности могло привести то, что один и тот же термин мог быть одинаково важен для нескольких тем одновременно.

Латентно-семантический анализ

LDA также реализован с использованием библиотеки scikit-learn (LatentDirichletAllocation класс). Также как описано выше для LSA по итогу работы алгоритма был вычислен argmax каждой записи в матрице, чтобы получить прогнозируемую категорию темы для каждого новостного анонса. Затем эти тематические категории были охарактеризованы по наиболее часто используемым словам, что проиллюстрировано на гистограмме (рис. 6).

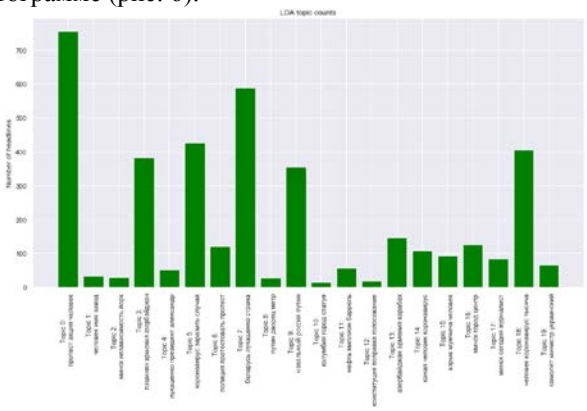

Рис. 6. Результат работы LDA, в каждой найденной теме выделены три наиболее частотных слова для визуализации

Получившиеся результаты отличаются от результатов, полученных с помощью LSA: выделенные темы более последовательны, к тому же распределение тем выглядит более убедительно. Вероятно, это следствие вариационного алгоритма Байеса, который начинается с равных априорных значений для всех

категории и только постепенно обновляет их по мере прохождения через набор данных.

Для интерпретируемого сравнения LDA с LSA полученная с использованием этого метода тематическая матрица также была спроецирована в двухмерное пространство (рис. 7).

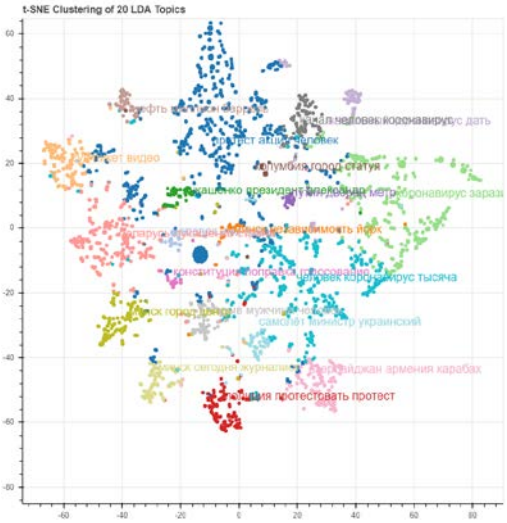

Рис. 7. Визуализация полученных тематических кластеров (модель LDA)

Отображение кластеров в двумерное пространство четко показывает, что LDA сработало для анализируемых данных гораздо лучше: тематические кластеры четко разделены между собой, к тому же каждая тема отсортирована в почти непрерывные области (инверсивная картина наблюдалась на рис. 5).

В качестве завершающего этапа исследования было решено выделить 4 наиболее популярные темы, детектированные алгоритмом LDA, чтобы обратиться ко второму набору данных с целью выявления динамики обсуждения в Telegram-канале тем, которые настолько активно обсуждались в СМИ (рис. 8). Поиск проводился по ключевым словам, выделенным на предыдущем этапе.

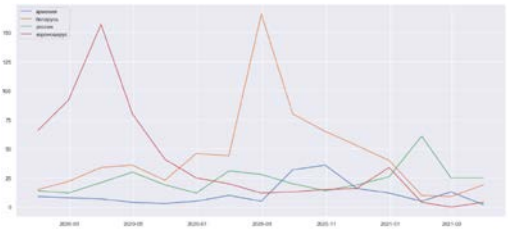

Рис. 8. Активность обсуждения тем протестов в России, Армении и Беларуси в сравнении с динамикой развития темы короновируса (анализ настроений при этом не учитывался)

Выводы

Скрытый семантический анализ (LSA) и скрытое распределение Дирихле (LDA) использовались для определения присутствующих в новостных анонсах тем. Модели LSA не удалось добиться большого разделения между выделенными кластерами, хотя выделенные темы выглядели достаточно интерпретируемыми, если ориентироваться на наиболее частотные слова. В то же время алгоритм LDA продемонстрировал большой потенциал для подобных исследований, добившись хорошего разделения между темами. Для дальнейшей оптимизации LDA-модели видится целесообразным выбор количества тематических групп производить путем оптимизации показателей качества. Например, использовать показатель согласованности, который измеряет семантическое сходство между наиболее часто встречающимися словами в теме. Максимально увеличив показатель согласованности, удастся добиться еще более хорошего качества реализованной модели.

Полученный график активности пользователей социальных сетей показывает, что СМИ оказывают влияние на динамику обсуждений той или иной темы (на рис. 8 видны пики максимального внимания и наибольшей апатии к каждой из выделенных тем, которые совпадают со временем появления новых релевантных информационных поводов), однако при этом наблюдается способность индивидов к самостоятельной оценке актуальности освещаемой средствами массовой информации повестки дня (это отчетливо видно по общему распределению количества обсуждений каждой из тем).

Формируемое посредством коммуникативного социального действия социальное настроение массы (или, по крайней мере, социальной группы) является продуктом взаимодействия индивидуальных сознаний, при этом начальный этап формирования социального сознания, функционирующего как единое целое, находится в точке, где каждый отдельный индивид, получая информационный повод, испытывает острое эмоциональное состояние. Оно, привлекая внимание на социальном, политическом или культурном уровне, заставляет индивидов соединиться в единое целое – массу. То есть базис социального настроения – эмоции, затем индивид соотносит себя с доступными социальными группами, и только на третьем этапе переживания события включаются рациональные интенции.

Роль средств массовой информации в процессах формирования массового сознания видится в эпоху постиндустриального общества все более ощутимой, так как именно они являются основными поставщиками информационных поводов, которые впоследствии становятся предметом социального коммуникативного акта. При этом необходимо отметить, что, потребляя контент, индивиды воплощаются в образе соучастника процесса создания интерпретации освещаемого события, общественное мнение и общественное настроение в таком случае становятся гибкой системой взглядов, формируемых достаточно спонтанно. Но интерпретации событий в совокупности с социальным настроением как новым компонентом сообщения являются мощным инструментов, позволяющим приводить социокультурный ландшафт к трансформации.

Библиография

[1] Ядов В.А. Социальные и социально-психологические механизмы формирования социальной идентичности личности // Мир России. 1995. №3- 4. – С.158-181.

[2] Поршнев Б. Ф. Социальная психология и история. М, 1979. – 232 с.
[3] Weber M., Basic Concepts in Sociology. 2000. 123 p.
[4] Parsons T. The structure of social action. 1937. 753 p.
[5] Habermas J., Theorie des kommunikativen Handelns. Frankfurt, 1991. 385 p.
[6] Habermas J. The Idea of the Theory of Knowledge as Social Theory // J. Habermas. Knowledge & Human Interests. 1987. pp. 102-317.
[7] Husserl E., Logical Investigations. International Library of Philosophy. 2001. vol. 1&2. 362 p.
[8] Sartre J.-P., La Transcendance de l'Ego. 1936. 87 p.
[9] Зиммель Г. Социология пространства. Избранное: в 2 т. М., 1996. Т. 2. 607 с.
[10] Сорокин П. А. Социальная стратификация и мобильность // Человек. Цивилизация. Общество. М., 1992. С. 295-425.
[11] Луман Н. Социальные системы. Очерк общей теории. Спб, 2007. 641 с.
[12] Кастельс М. Информационная эпоха: экономика, общество и культура. М., 2000. С. 41-42.
[13] Bellegarda J.R. Latent Semantic Language Modeling for Speech Recognition, Mathematical Foundations of Speech and Language Processing, IMA, V 138. pp 73-103. 2004.
[14] Blei D., Ng A., Jordan M. Latent Dirichlet allocation. Journal of Machine Learning Research 3. pp 993-1022. 2003.
[15] Barlow R. E. Introduction to de Finetti (1937) Foresight: Its Logical Laws, Its Subjective Sources. Breakthroughs in Statistics. pp 127-133. 1992.
[16] Maaten L.J.P., Hinton G.E. Visualizing Data Using t-SNE. Journal of Machine Learning Research. Pp 2579-2605. 2008.

Чижик Анна Владимировна, кандидат культурологии, Санкт-Петербургский Государственный Университет, старший преподаватель, ORCID 0000-0002-4523-5167 (a.chizhik@spbu.ru)

The Detecting Dynamics of Public Opinions in Social Media using thematic modeling methods

A.V. Chizhik

***Abstract*— Digital technologies have led to the formation of the new level of socio-cultural space, which is expressed in the permanent presence of the phenomenon of virtual reality in everyday life. It is the main motivating tool for social and political transformations of society, which consist in accelerating the logical bound "public mood → public opinion → social action of the masses". Understanding these changes in social phenomena actualizes the problem of forming systems for detecting the current public mood, which can be the basis of feedback between the authorities and society, or monitoring public mood as a reaction to the social and political situation.**

This study would work on topic modeling focused on the algorithm employing Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA). The data collection of news announcements, that were published between 2020 and 202, is used as the main data resours with unstructed text. The stages of preprocessing include cleansing, stemming, and stop words. The advantages of LSA are fast and easy to implement. LSA, on the other hand, doesn't consider the relationship between documents in the corpus, while LDA does. This study shows that LDA gives a better result than LSA.

***Keywords*— topic modeling, text embeddings, LDA, LSA.**

REFERENCES

[1] Yadov V.A. Social and socio-psychological mechanisms of the formation of a person's social identity // World of Russia. 1995. No. 3-4. - P.158-181.
[2] Porshnev BF Social psychology and history. M, 1979 .-- 232 p.
[3] Weber M., Basic Concepts in Sociology. 2000. 123 p.
[4] Parsons T. The structure of social action. 1937. 753 p.
[5] Habermas J., Theorie des kommunikativen Handelns. Frankfurt, 1991. 385 p.
[6] Habermas J. The Idea of the Theory of Knowledge as Social Theory // J. Habermas. Knowledge & Human Interests. 1987. pp. 102-317.
[7] Husserl E., Logical Investigations. International Library of Philosophy. 2001. vol. 1&2. 362 p.
[8] Sartre J.-P., La Transcendance de l'Ego. 1936. 87 p.
[9] Simmel G. Sociology of space. Selected works: in 2 volumes, Moscow, 1996.Vol. 2.607 p..
[10] Sorokin PA Social stratification and mobility // Man. Civilization. Society. M., 1992.S. 295-425.
[11] Luhmann N. Social systems. An outline of the general theory. SPb, 2007.641 p.
[12] Castells M. Information age: economy, society and culture. M., 2000.S. 41-42.
[13] Bellegarda J.R. Latent Semantic Language Modeling for Speech Recognition, Mathematical Foundations of Speech and Language Processing, IMA, V 138. pp 73-103. 2004.
[14] Blei D., Ng A., Jordan M. Latent Dirichlet allocation. Journal of Machine Learning Research 3. pp 993-1022. 2003.
[15] Barlow R. E. Introduction to de Finetti (1937) Foresight: Its Logical Laws, Its Subjective Sources. Breakthroughs in Statistics. pp 127-133. 1992.
[16] Maaten L.J.P., Hinton G.E. Visualizing Data Using t-SNE. Journal of Machine Learning Research. Pp 2579-2605. 2008.

Anna V. Chizhik, Saint Petersburg State University, senior lecturer, ORCID 0000-0002-4523-5167 (a.chizhik@spbu.ru)