ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В ИССЛЕДОВАНИЯХ КОММУНИКАЦИИ

Бодрунова Светлана Сергеевна, профессор кафедры медиаменеджмента массовых коммуникаций, Институт «Высшая школа журналистики и массовых коммуникаций», Санкт-Петербургский государственный университет, г. Санкт-Петербург, Россия; e-mail: s.bodrunova@spbu.ru, ORCID 0000-0003-0740-561X

Блеканов Иван Станиславович, кандидат технических наук, заведующий кафедрой технологии программирования, факультет прикладной математики – процессов управления, Санкт-Петербургский государственный университет, г. Санкт-Петербург, Россия; e-mail: I.blekanov@spbu.ru, ORCID 0000-0002-7305-1429

Тарасов Никита Андреевич, младший научный сотрудник, факультет прикладной математики – процессов управления, Санкт-Петербургский государственный университет, г. Санкт-Петербург, Россия; e-mail: nkt.tarasov@yandex.ru, ORCID 0000-0002-9473-6179

«ДЕРЕВО МНЕНИЙ»: МЕТОД ДИНАМИЧЕСКОГО МЭППИНГА ОНЛАЙН-ДИСКУССИЙ НА ОСНОВЕ НЕЙРОСЕТЕВОГО ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ И АБСТРАКТИВНОЙ СУММАРИЗАЦИИ[1]

Современные модели формирования мнений в онлайн-среде, в т.ч. концепция кумулятивной делиберации, ставят вопросы о природе и динамике публичного консенсуса. Нейросетевые методы детекции мнений способны помочь в выявлении согласия/несогласия пользователей, включая складывание и разрушение консенсуса. Но пока не разработаны методики, учитывающие накопительную динамику мнений – их «мутацию», расхождение и слияние. Наша работа посвящена построению интерпретируемого «дерева мнений», в котором нарастают и изменяются мнения в онлайн-дискуссии. Кейсом послужило обсуждение на платформе Reddit 27-й Рамочной конференции ООН 2022 г. по изменению климата. Метод сочетает три элемента: выявление тем дискуссии, определение в них точек бифуркации (ветвления) и суммаризацию высказываний в рамках «ветвей». Это позволяет одновременно оценить тематический разброс «ветвей» и «мутацию» мнения в них. Метод сочетает пошаговое тематическое моделирование на основе BERTopic и абстрактивную суммаризацию на базе Pegasus, позволяя одновременный мэппинг хода обсуждения и мнений пользователей в «свернутом», «развернутом» и полярном виде на основе суммаризации разной длины.

***Ключевые слова**: кумулятивная делиберация, онлайн-дискуссия, тематическое моделирование, абстрактивная суммаризация, BERT, Reddit, COP27.*

Как цитировать: *Бодрунова С. С., Блеканов И. С., Тарасов Н. А.* «Дерево мнений»: метод динамического мэппинга онлайн-дискуссий на основе нейросетевого тематического моделирования и абстрактивной суммаризации // Вестник Московского университета. Серия 10. Журналистика. 2025. № 5. С. 179–208. DOI: 10.55959/msu.vestnik.journ.5.2025.179208

Svetlana S. Bodrunova, Doctor of Political Sciences, Professor, Department of Media Management and Mass Communications, Institute "School of Journalism and Mass Communications", St. Petersburg State University, Saint Petersburg, Russia; e-mail: s.bodrunova@spbu.ru, ORCID 0000-0003-0740-561X

Ivan S. Blekanov, Head of the Department of Programming Technologies, Faculty of Applied Mathematics – Control Processes, St. Petersburg State University, Saint Petersburg, Russia; e-mail: I.blekanov@spbu.ru, ORCID 0000-0002-7305-1429

Nikita A. Tarasov, Junior Research Fellow, Faculty of Applied Mathematics – Control Processes, St. Petersburg State University, Saint Petersburg, Russia; e-mail: nkt.tarasov@yandex.ru, ORCID 0000-0002-9473-6179

'OPINION TREE': A METHOD FOR MAPPING ONLINE DISCUSSIONS BASED ON NEURAL-NETWORK TOPIC MODELING AND ABSTRACTIVE SUMMARIZATION

So far, no neural-network-based methodologies that aim at online opinion detection allow for representing user discussions on social networks in forms that would simultaneously capture cumulation, shift, and dissipation of consensus. Such a method would allow for scrupulous tracing of the opinion dynamics (including polarization of views), shorten the time for evaluation of opinion dynamics, and help address several theoretical assumptions on the nature of cumulative opinions. We propose a method for construction of 'opinion trees' within user discussions on social networks. The case dataset features a Reddit discussion of the 27th UN Climate Change Conference (COP27/UNFCCC2022). The method includes three methodological steps, namely defining the topicality bifurcation points, measuring the 'thickness' of 'branches', and summarizing the meaning of individual 'branches', thus allowing for both topicality divergence assessment and quick enough opinion tracing. Our method integrates recursive BERTopic-based topic modeling and Pegasus-based abstractive summarization, allowing for opinions to be seen as 'folded', 'unfolded', and 'polar', as detected in summaries of varying length.

***Key words**: cumulative deliberation, online discussions, topic modeling, abstractive summarization, BERT, Reddit, COP27.*

To cite this article: Bodrunova S. S., Blekanov I. S., Tarasov N. A. (2025) "Derevo mneniy": metod dinamicheskogo meppinga onlayn-diskussiy na osnove neyrosetevogo tematicheskogo modelirovaniya i abstraktivnoy summarizatsii ['Opinion tree': a method for mapping online discussions based on neural-network topic modeling and abstractive summarization]. *Vestnik Moskovskogo Universiteta. Seriya 10. Zhurnalistika* 5: 179–208. DOI: 10.55959/msu.vestnik.journ.5.2025.179208

DOI: 10.55959/msu.vestnik.journ.5.2025.179208

Введение

Мнения пользователей в социальных сетях сегодня стали частью публичной сферы и делиберативного процесса. Делиберация до появления интернета и социальных сетей понималась как процесс обретения рационального и легитимного консенсусного решения по общественно-значимым вопросам путем дискурсивного «взвешивания» (*deliberare* (лат.) – «взвешивать») предложений разных участников дискуссии. Делиберация имела форму раундного обсуждения проблемы или конфликта в публичном пространстве вплоть до того момента, когда будет найдена точка согласия, желательно равноудаленная от участников диспута и при этом частично включающая их позиции. Такой процесс предполагал ограниченное число участников обсуждения, их желание прийти к компромиссу и отказ от иррациональности в аргументации, что гарантировало высокое качество обретаемого решения (Habermas, 1992; 1996).

Однако с появлением массовых публичных обсуждений в интернете и особенно в соцсетях общественная дискуссия, по мысли многих ученых, стала диссонантной и разобщенной (Pfetsch, 2018). В соцсетях она почти никогда не направлена на поиск консенсуса, излишне комплексна, диссипативна, дискретна, несвязна, нерациональна и малопредсказуема, что влияет на качество публичной делиберации и обретение согласия. Но соцсети также создали колоссальные возможности для изучения текущего общественного мнения и его динамики, добавив новое измерение к классическим взглядам на анализ позиций граждан. Выявление тематики онлайн-дискуссий и точек зрения пользователей на обсуждаемые проблемы проходит сегодня путь (одновременно) методологического усложнения и технической автоматизации.

Важной проблемой в области полуавтоматизированной оценки мнений стала проблема динамики и эволюции тем/мнений в рамках конкретной онлайн-дискуссии. С одной стороны, огромное количество работ, начиная с первых исследований структуры нарождающейся блогосферы, посвящено выявлению связи структуры дискуссии и ее тематики, часто смешанной с политической позицией пользователей, т. к. методы кластеризации на основе «мешка слов» или тональности речи допускают выделение кластеров на разных основаниях (Кольцова, Маслинский, 2013). При этом графовая визуализация показывает только финальную кластеризацию дискуссии, выдавая статику веб-графа за динамику мнений. Таким подходам недостает учета временных параметров обсуждения, свя-

занных с изменением тематики и мнений в текстах пользователей. Отдельные недавно возникшие научные зоны – например, изучение эволюции тематики (*topic evolution studies*) – пытаются разрешить эту проблему, но, кластеризуя тексты во времени, перестают учитывать теперь уже групповые связи в дискуссии. Подхода, который одновременно позволял бы выявить ветвление тематики и отражал бы мнения в динамике, складывающиеся в разных ветвях дискуссии, пока не разработано.

С другой стороны, структурно-тематико-временно́му картированию онлайн-обсуждений не уделялось должного внимания, т. к. само по себе оно часто не считается финальной задачей, а только подзадачей, решить которую нужно для ответа на вопросы в области социальных наук. При этом построение временного «дерева дискуссии» с тематикой и мнениями, «привязанными» к отдельным «ветвям», на наш взгляд, позволяет:

– точечно отследить динамику тем и поляризацию мнений по разным вопросам внутри больших наборов текстовых данных;

– критически сократить время, нужное исследователям для понимания динамики мнений, путем предоставления краткой информации о том, когда и какие мнения выделились и насколько они весомы;

– по-новому заглянуть внутрь онлайн-обсуждений и ответить на ряд теоретических вопросов о природе и динамике текущего общественного мнения, в т. ч. поставленных в рамках концепции кумулятивной делиберации.

Термин «кумулятивная делиберация» был предложен нами в 2021 г. и впоследствии развит (Бодрунова, 2023) как дополнение к классическим моделям делиберативного процесса именно для онлайновой публичной сферы. Концепция кумулятивной (накопительной) делиберации указывает на то, что характер мнений в онлайн-обсуждениях является в первую очередь накопительным и не следует паттернам рационального обсуждения «по кругу».

Она предлагает обновленный взгляд как на природу складывания мнений, так и на нормативную меру качества публичной коммуникации и участия в ней «обычных пользователей». Концепция ставит в т. ч. вопросы о том, как эндогенные (дискурсивные) и экзогенные (контекстуальные, платформенные, демографические) факторы, влияющие на ход дискуссии, связаны с паттернами и эффектами накопления мнений. Предлагаемый метод «древовидного» представления дискуссии может помочь в осознании связи

между структурой и скоростью «роста ветвей» дискуссии, с одной стороны, и факторами «ветвления» – с другой. Задачей-максимум при создании «дерева мнений» мы считаем поиск точек консенсуса в онлайн-обсуждении.

Для создания «дерева мнений» в онлайн-дискуссии нами предпринята попытка скомбинировать методы, которые ранее применялись для выявления смысловой структуры текстовых данных, и развернуть построение дискуссии во времени. Автоматизированные подходы к анализу текстовых корпусов включают в т. ч. методы, основанные на строгой и нестрогой кластеризации и различных видах суммаризации (создания резюме авторских текстов). Так, мы определяем точки «ветвления» дискуссии с помощью первичного применения нейросетевой тематической модели *BERTopic* (включающей в себя этапы кодировки, кластеризации, снижения размерности и выявления ключевых терминов тематик) с последующим построением структуры терминов во времени и, наконец, выявлением точек ветвления с помощью метода оценки схожести списков ключевых терминов. Для оценки «толщины ветви» мы суммируем число текстов, относящихся к теме «ветви». Для краткого описания смысла (темы и мнений) конкретной «ветви» применяется метод абстрактивной суммаризации с использованием предобученной нейросетевой модели *Pegasus* (Passali, Gidiotis, Chatzikyriakidis, Tsoumakas, 2021), дообученной нами для работы с пользовательскими текстовыми данными для трех языков; метод был апробирован ранее (Blekanov, Tarasov, Bodrunova, 2022). Об абстрактивной суммаризации как методе создания текстовых резюме подробнее см. ниже; здесь укажем, что идея суммаризации – в создании с нуля кратких версий текстов на основе предобучения нейромодели.

Для эксперимента был выбран кейс обсуждения на платформе *Reddit*. Тексты используются для тестирования методов абстрактивной суммаризации, т. к. платформа предоставляет пользователям право создать авторское резюме текста в формате *too long; didn't read* (*tl;dr*). Такие пользовательские резюме сравниваются с машинными резюме при тестировании метода. Наш выбор также обусловлен несколькими иными факторами, включая более удобный (пока) для нейромоделей английский язык, на котором шла дискуссия, среднюю длину текста, доступность данных и наличие разработанных нами инструментов для их сбора. Кейсом среднего объема, удобным одновременно для построения многих «ветвей» дискуссии и для охвата «дерева мнений» одним взглядом, стало обсуждение 27-й Конференцией сторон Рамочной конвенции

ООН об изменении климата (*COP27/UNFCCC2022*) в Шарм-эль-Шейхе (Египет) 6–18 ноября 2022 г. Эта тематика обсуждалась на *Reddit* с привлечением контекстуальной информации и иных вопросов повестки дня, что разнообразило картину «малых» тем и мнений внутри них.

Далее в статье мы аргументируем необходимость мэппинга онлайн-дискуссий в рамках концепции кумулятивной делиберации и ставим вопросы, на которые такой мэппинг может отвечать. Затем мы приводим краткий обзор применения машинных методов к задаче мэппинга онлайн-дискуссий. Под мэппингом здесь мы понимаем процесс картирования содержания дискуссии в сокращенном, удобном для чтения виде через одновременное представление ее тематики и позиций пользователей в выявленных и резюмированных «ветвях» обсуждения. Отдельно мы останавливаемся на методе абстрактивной суммаризации, который, в отличие от моделирования тем, ни в России, ни за ее пределами пока не применялся для картирования сетевых дискуссий. Затем мы описываем методологию, сбор данных, ход исследования и его итоги. В дискуссионной части мы оцениваем недостатки метода и возможные варианты его развития.

Обзор литературы

Накопление мнений в соцсетях: к вопросу о необходимости смыслового мэппинга онлайн-дискуссий. Как можно заключить из ранних исследований (King, Anderson, 1971; Noelle-Neumann, 1974), общественное мнение обладает накопительным характером с видимыми паттернами формирования большинства/меньшинства; сегодня к ним добавились паттерны взлета и слома обсуждений, поляризации и «эхо-камер», «перетекания через край», каскадного и вирусного распространения информации и т. д. Накопительным характером обладают многие классические медиаэффекты (Koch, Arendt and Maximilian, 2017). Но и классические концепции публичной делиберации (Habermas, 1996; Bohman, Rehg, 1997), и их левая, феминистская и иная критика почти не учитывали именно накопительную природу общественного мнения. Редким исключением здесь являются работы Шанталь Муфф, которая говорила о консенсусе как кристаллизации гегемонии (Mouffe, 2000). Это позволяет увидеть публичную дискуссию не только как институциональный процесс, но и как условный «Солярис», где каждый

новый момент времени можно вообразить как новое сочетание накопленных мнений, включая институциональные и «обычные» мнения.

С развитием соцсетей «айсберг» накапливаемых мнений, ранее бытовавших в устной форме, стал во многом доступен для отслеживания и изучения. Но накопительный характер мнений наложился на ставшее очевидным крайне низкое делиберативное качество людских высказываний. Отвечая на вопрос С. Ярди и д. бойд от 2010 г. о том, для чего люди приходят в Интернет – согласиться или поспорить (*to agree or to argue*) (Yardi, boyd, 2010), можно сказать: ни за тем, ни за другим. Разнообразие мотиваций и фрагментарность участия пользователей в дискуссиях указывают на то, что пользователи почти не имеют делиберативных целей. Но они тем не менее могут являться носителями мнений, высказывать их и, таким образом, *en masse* влиять на позиции институциональных акторов. Это требует пересмотра нормативности в теории публичной сферы (Бодрунова, 2023). При этом неделиберативный характер дискуссии (естественная речь, фрагментарное участие, накопительные мнения) не отменяет значимости агрегированных мнений для общего процесса делиберации. Таким образом, делиберация в целом приобретает накопительное измерение, в котором важными являются и паттерны, и эффекты накопления мнений, и факторы, определяющие их.

Мы видим в динамическом мэппинге потенциал не только для оптимизации работы с большими текстовыми данными, но и для анализа природы общественного мнения в его динамике. При оценке применимости метода мы будем иметь в виду следующие исследовательские вопросы: Как выглядит пользовательское мнение в суммаризированном виде? Интерпретируемо ли оно? Можно ли вычленить из резюме поляризованные мнения и, главное, найти точку консенсуса? При этом не хотелось бы ограничиваться качественной оценкой понятности суммаризаций, как это часто бывает в исследованиях текстовых массивов. Нам важно, какие типы мнений видит машина и можно ли рассчитывать на обнаружение точек консенсуса.

Построение «дерева»: методы мэппинга тематики и мнений в онлайн-дискуссиях. Идея картирования тематики и мнений в динамике еще в 1998 г. была разделена на три задачи: «(1) сегментирование потока данных в отдельные истории; (2) идентификация историй, которые будут обсуждаться в первую очередь; (3) <...>

нахождение всех последующих новостей в данной истории» (Allan, Carbonell, Doddington, Yamron et al., 1998: 1). Постепенно научная зона *topic detection and tracking* (*TDT*) разрасталась; в разное время для прослеживания эволюции тем+мнений в соцмедиа применялись разные методы автоматизированного анализа (тематическое моделирование, тональный анализ, иные варианты кластеризации текстов и их сочетания со структурным анализом соцсетей, локацией сообществ и т. д.).

Мы включим в обзор только те исследования в зоне *topic evolution studies,* которые напрямую учитывают временное измерение. Общая цель этой группы работ состоит в выявлении того, как изначально найденные для всего датасета темы эволюционируют во времени внутри себя (Alam, Ryu and Lee, 2017). Мы не будем рассматривать работы о динамике популярности отдельных тем; нас интересует иная динамика темы – динамика смысла (в сочетании, конечно, с динамикой популярности). Мы также не включаем в обзор работы по динамическому тематическому моделированию, которые предлагают заново разбивать тексты по темам на каждом новом временном отрезке; наша задача – определить точки ветвления уже известных тем самым простым и применимым способом.

Тем не менее динамическое моделирование важно нам, т. к. оно принципиально показало возможность построения генеративных моделей тематики (AlSumait, Barbará and Domeniconi, 2008). Такие модели позволяют как прослеживать изменения в тематике, так и выявлять вновь возникающие в датасете темы; но они редко применяются для реального анализа в силу сложности. Более простые (и часто более применимые) опции трекинга тематики исследуют семантические сдвиги внутри тем, например с регулярным шагом – от нескольких часов (Lu, Guo, Chen, Wang, 2022) до десятилетия (Blei, Lafferty, 2008).

С конца 2000-х гг. появлялись статьи, предлагающие одновременно выявление новых тем и анализ их эволюции на уровне ключевых слов (Ahmed, Xing, 2012). Сосредоточившись как на развивающихся, так и на возни́кающих темах, исследователи внедрили временную регуляризацию моделирования (Saha, Sindhwani, 2012) и метрики новизны и угасания тем (Huang, Peng, Wang, Cao et al., 2017). Но часто даже в новейших работах авторы упускали из вида то, как продолжают выглядеть основные, а не новые темы в каждый новый период времени.

В начале 2010-х случился «темпоральный поворот» в тематическом моделировании, и динамический анализ тематики, в т. ч. анализ трендов, стал набирать популярность (Hong, Yin, Guo, Davison, 2011), хотя под динамическим анализом понимались очень разные задачи. Одной из важных идей стало применение «скользящего окна» (*sliding window*) для текущего потока данных, где одни посты уходили из моделирования, а другие попадали в него. Так называемое инкрементное моделирование, позволяющее видеть нарастание темы, привнесло еще одну идею – использовать только ключевые тексты для формирования тем (Lee, Lakshmanan and Milios, 2013).

Несмотря на «темпоральный поворот», лишь небольшой ряд авторов подошел к идее сочетать динамику смысла и ее ветвление и их обоих – с иными аспектами дискуссии, помогающими понять ее смысл. Так, объединение метрик связности сети, тематического подобия и временной близости позволило отобразить динамику аспектов смысла в рамках одной темы; при этом магистральная тема могла быть сколь угодно широкой (Tang, Yang, 2011). Подобные работы сформировали линию анализа «контент и сеть» (*content-and-network approaches*), где поток текстов моделируется одновременно как динамическая сеть постов и как динамический кластер (Lee, Lakshmanan and Milios, 2014). Это позволяет видеть параллельное развитие темы и сети, которой она принадлежит. Еще один подход, близкий к идее мэппинга мнений в изменяющихся темах, соединяет моделирование тем и тональный анализ («настроение» пользователей) (Dermouche, Velcin, Khouas, Loudcher, 2014). Некоторые авторы привлекают контекст (например, встречающиеся в теме хэштеги), геолокацию, данные пользователей и другие параметры для более точного отслеживания тем.

Также была реализована важная идея относительной «толщины веток» при переплетении тем во времени (Saha, Sindhwani, 2012), в т. ч. в наших более ранних работах об относительной «выпуклости» (saliency) тем.Но и в них не была по-настоящему отражена эволюция тем как эволюция их содержания. Тема предполагалась неизменной/изменяющейся пренебрежимо мало; накопление и распад мнений внутри темы не анализировались с применением нейросетей.

Близко к идее древовидного мэппинга дискуссии подходит изучение тематических цепочек (*topic chains*) (Kim, Oh, 2011). Так, авторы видят «большие» темы как развивающиеся во времени цепочки микротем и экспериментируют с расхождением/схождением

цепочек, показывая, что при изменении параметров модели число цепочек варьирует от 1 (все темы слиты в одну цепочку) до 0 (цепочки не формируются), а между ними находится некоторое оптимальное число цепочек. Т. е. сам исследователь может моделировать их число и объем, и структура «ветвей» зависит от авторских настроек, как и в классическом тематическом моделировании. Микротемы при этом объединены в «большую» цепь по принципу схожести ключевых слов (и отделены от других «больших» тем). Увы, но цепи тем, обнаруженные Кимом и Оу, не дают понимания, когда и что изменяется внутри «большой» темы; об этом можно судить только по ключевым словам. Наш метод отличается тем, что вместо объединения разрозненных фрагментов датасета, написанных в разное время, он «ветвит» именно наблюдаемую тему, находя точки бифуркации тематики внутри нее, и находим мы эти точки через измерение схожести ключевых слов, но внутри темы, а не внутри целого датасета.

Некоторые рабочие группы подошли к идее ветвления мнения через идею подтем (*sub-topics*). Она не слишком отличается от изучения «выпуклости» тем, но сочетает ее с идеей возникновения новых мини-потоков сообщений внутри общей темы (Srijith, Hepple, Bontcheva, Preotiuc-Pietro, 2017). Идея «тематических деревьев» может быть обнаружена в работе, посвященной оптимизации тематического моделирования, но в ней только одно «дерево тем» реализуется в единицу времени (Kawamae, 2018). Одна ранняя работа демонстрирует древовидную диаграмму тематики и описывает подтемы в каждой теме, правда без применения автоматизации (Jo, Hopcroft & Lagoze, 2011). Более поздние варианты применяют «деревья тем» (*topic trees*) (Song, Huang, Qi, Li et al., 2016) и «эволюционные деревья» в научных данных с годичным шагом (Wang, Zhang, Zhou, Li et al., 2017) и показывают именно смысловые разветвления тем на более мелкие, но не позволяют ученым видеть содержание текстов. В нескольких работах присутствует идея мэппинга, но иного – через привязку развивающихся тем к карте мира (Cai, Tang, Yang, Huang, 2014). В 2017 г. был опубликован обзор статей по мэппингу эволюции тем, но для кроссмедийных наборов данных, т. е. включающих текст, видео, аудио (Zhou, Yu, Hu, Hu, 2017); в упомянутых там работах эволюция содержания исследуемых тем также не раскрывается.

В конце 2010-х гг. случился еще один «поворот» в динамическом тематическом моделировании – нейросетевой. Так, к выяв-

лению точек ветвления («подсобытий» – *sub-events*) стали применяться нейросетевые модели (Chen, Xu and Mao, 2018), в т. ч. на основе архитектуры *Transformer* (Gao, Peng, Wang, Zhang et al., 2020). Для улучшения качества моделирования используется также представление текста на основе его векторной стандартизации через т. н. «эмбеддинги» (*word embeddings*) (Momeni, Karunasekera, Goyal, Lerman, 2018) – стандартизированные замены редких слов. Их числовые векторы преобразуют семантические отношения в математический аналог; благодаря замене лексических единиц на их более частотные синонимы тексты легче поддаются кластеризации, и удается выстроить гибкую картину ветвления, слияния и угасания тем (Abulaish, Fazil, 2018). Показано, что нейросетевые модели более устойчивы, проще в использовании, лучше работают с различными видами данных, позволяют машине определять число тем (используя *HDBSCAN* или другие методы *smart*-кластеризации), допускают тонкую настройку (*fine-tuning*) модели и динамическое моделирование аспектов тем (Zhang, Fang, Chen, Namazi-Rad, 2022). В 2022 г. была предложена модель *BERTopic* на основе двух нейронных сетей – архитектуры *BERT* и процедуры эмбеддинга на базе *Transformer* (Grootendorst, 2022); мы настроили обе модели под задачу и применяем указанное сочетание, включающее кодировку, кластеризацию, снижение размерности датасета и выбор основных терминов (о применении *BERTopic* для «разбиения» и «наследования» тем см. также: Wang, Chen, Chen, Chen, 2024).

Пока ни один из описанных выше методов не дает возможности легко построить и прочитать настоящее «дерево мнений» – т. е. понять, какие темы есть, когда они разделяются, насколько они мощные, что в них обсуждается и как в них «мутируют» мнения. Недостатком *topic evolution studies* является то, что авторы работ ограничиваются классической репрезентацией тематики через ключевые слова и – реже – через облака слов или краткие аннотации, что не дает понимания смысла обсуждения без чтения большого числа оригинальных текстов. Как ни парадоксально, в работах об эволюции тем мы часто видим не меняющиеся/трендовые темы, а только модель в целом и оценку ее качества. Вероятно, поэтому большинство сложных методов анализа текстов на базе тематических моделей мало применяются в социогуманитарном знании.

«Дерево мнений», как сказано выше, не только может располагать темы и мнения относительно друг друга, но и позволяет про-

слеживать накопление мнений и делать это с достаточной степенью удобства. Для решения этой задачи мы объединили нейросетевое тематическое моделирование, позволяющее выявлять моменты разделения тем, векторную стандартизацию текстов и их абстрактивную суммаризацию, чтобы кратко изложить содержание «ветвей» дискуссии. О суммаризации стоит сказать отдельно.

Абстрактивная суммаризация и ее применение при описании сегментов дискуссии. Для изучения общественного мнения важно, чтобы была понятна не только тема, но и взгляды пользователей и их поляризация. Для этого мы предлагаем применять еще один нейросетевой метод – абстрактивную суммаризацию.

Суммаризация текстов была разработана в четырех формах, из которых сегодня широко применяются две – экстрактивная и абстрактивная. Первая кратко передает смысл текста или группы текстов через нахождение и представление наиболее значимых фрагментов (см., например, представление новости на *dzen.ru*). Абстрактивная суммаризация создает на основе длинного текста/группы текстов собственное резюме, как если бы человек кратко передавал смысл своими словами.

Идея суммаризации текстов в темах для более быстрого понимания тематической картины появилась еще в начале 2010-х. Первые попытки совместить эволюцию тем и суммаризацию текстов из социальных сетей строились с помощью экстрактивных методов: через выбор наиболее характерных предложений (Li, Li, 2013) или репрезентативных твитов для каждого временного сегмента – ручных (Brigadir, Greene and Cunningham, 2014), автоматизированных (Wang, Shou, Chen, Chen et al., 2015; Zhou, Kanhabua and Cristea, 2016), «взвешенных» на основе метрики *TF-IDF* (Alsaedi, Burnap and Rana, 2017) и др. Но в ранних работах суммаризации подвергались не тематические «деревья», а отдельные сквозные темы, найденные в данных. В 2016 г. к ним была применена идея одновременного разделения на сегменты и суммирования значений в каждом из них (Chang, Tang, Yin, Yamada et al., 2016), близкая идее «дерева мнений».

Абстрактивная суммаризация была применена к выстраиванию «линии темы» (*topic storyline*) только спустя годы. Она почти одновременно применялась для определения тематических сдвигов в дискуссиях в работе Гойала и коллег (Goyal et al., 2020) и в наших исследованиях (Blekanov, Tarasov, Bodrunova, 2022). Для картирования смысла онлайн-дискуссий абстрактивная суммаризация пока не использовалась, несмотря на массу экспериментов с наборами

данных из Reddit и Твиттера*. В начале 2020-х для русского языка всего в нескольких работах были предложены простые модели суммаризации фрагментов новостных (заголовков, лидов) и научных текстов; сегодня внимание к ним понемногу растет среди компьютерных лингвистов (Митрофанова, Адамова, Букреева, Голубев и др., 2024; Сорокина, 2024). Нам также известна одна кратко описанная попытка применить гибридную экстрактивно-абстрактивную модель к текстам *Telegram* (Головнин, Курганов, 2022). Мы также пока не используем абстрактивную суммаризацию для русского языка, но видим это как естественное продолжение исследований при применении предложенного метода.

Методология исследования

Дизайн исследования состоит из пяти этапов:

– сбор и подготовка данных (1);

– построение «дерева мнений» путем нейросетевого тематического моделирования – выявление макротем в датасете (2) и точек ветвления (3);

– определение «толщины ветвей» и визуализация «дерева мнений» (4);

– построение суммаризаций для отдельных «ветвей» и оценка их качества для решения задачи выявления мнений пользователей (5).

Кейс COP27/UNFCCC2022 и сбор данных. Как сказано выше, тестирование метода суммаризации пока удобнее проводить на англоязычных данных: суммаризация для русского языка требует существенной доработки моделей. Тестирование абстрактивной суммаризации чаще всего проводится на данных с платформы *Reddit* в силу наличия авторских суммаризаций формата "*tl;dr*". Для нас *Reddit* также был удобен: в отличие от платформ *Meta***, попытки загрузки данных с *Reddit* не блокируются и платформа не ограничена к использованию в России. Тексты на *Reddit* обладают большей средней длиной, чем на платформах микроблогов, что повышает качество и тематического моделирования, и абстрактивной суммаризации (хотя проблема большей вариабельности длины также должна быть решена).

С помощью разработанного нами веб-краулера по набору ключевых слов (*climate change, COP27, CCC2022)* были собраны сообщения пользователей на английском языке, опубликованные

* Принадлежит компании *Meta* (США), признанной экстремистской в РФ.

** Компания признана экстремистской в России.

с мая по ноябрь 2022 г. (полгода «вокруг» события), в количестве 54 565 (датасет 1). Для нашей задачи посты и комментарии не различались. Для подробного визуального представления и оценки качества суммаризации использовалась одна макротема («финансы») – 1341 текст (датасет 2).

Построение «дерева мнений». Общая схема обработки данных (*pipeline*), кратко описанная выше, включает следующие шаги (см. рис. 1):

1. Подготовка данных. Объединение текстовых данных каждого сообщения в датасете 1: внутри каждого сообщения заголовок и тело объединяются в один текст, фиксируется время публикации.

2. Подача подготовленных текстов сообщений (на рис. 1 обозначены как «сообщение») на вход нейросетевой тематической модели *BERTopic*.

3. Работа модели *BERTopic*. В данном исследовании использовалась собственная реализация и модификация концепции алгоритма *BERTopic* по причине отсутствия возможности модификации архитектуры и анализа результатов отдельных этапов. Приводимые далее шаги кратко описывают основные этапы реализации алгоритма.

3.1. Построение эмбеддингов для каждого сообщения с использованием предобученной нейросетевой модели *all-MiniLM-L6-v2*.

3.2. Сокращение размерности векторного представления эмбеддингов алгоритмом *UMAP*. Исходные эмбеддинги размерности 512 уменьшены до векторов размерности 50. Позволяет лучше определять схожесть текстов, что улучшает качество кластеризации путем снижения уровня семантического «белого шума» (лексики с крайне малым числом семантических связей).

3.3. Кластеризация сообщений на основе алгоритма иерархической кластеризации *HDBSCAN* для группировки сообщений по темам. Пространственная кластеризация *HDBSCAN* позволяет группировать сообщения по семантической близости. Это позволяет выделить тематические кластеры без фиксации количества тем заранее, а значит, определить не только темы дискуссии, но и их оптимальное количество без многочисленных прогонов и подбора числа тем исследователем (в отличие от алгоритмов моделирования, подобных латентному размещению Дирихле (*LDA*), где число тем задается исследователем произвольно).

3.4. Формирование макротем. С помощью процедуры агрегации тем (*Topic Reduction*) первоначальный набор тем объединяется в макрогруппы. Данный алгоритм объединяет темы с похожими ключевыми словами и используется в нашем исследовании для

получения основных «ветвей» «дерева мнений». Использовался подход с ручным отбором и объединением кластеров на основе примеров текстов из каждого кластера и их ключевых слов, получаемых с использованием меры *TF-IDF*. Учитывая проводимый пошаговый детальный контроль, стандартная процедура на основе косинусного расстояния не применялась.

3.5. Выделение ключевых слов для описания макротем. Все сообщения, входящие в одну макротему, объединяются в один документ с последующим выявлением в нем наиболее значимых слов с помощью алгоритма *TF-IDF*, где *TF* – частота термина в отдельном кластере, *IDF* – обратная частота документа во всей коллекции.

4. Итеративное построение «ветвей» внутри макротем:

4.1. Разбиение макротем на временные интервалы. Корпус сообщений внутри макротемы дополнительно разбивается на равные промежутки (параметр разбиения фиксированный и задается исследователем после претестов; в нашем случае он равен 30 промежуткам для каждой темы).

4.2. Выделение ключевых слов для каждого промежутка. Для описания темы каждого из 30 временных промежутков внутри кластера повторно вычисляется метрика *TF-IDF* (с фиксированным значением *IDF* по всему датасету 1, тогда как значение *TF* рассчитывается заново для каждого промежутка). Расчет направлен на поиск главных ключевых слов темы на данном промежутке; главными считаются пять первых слов с наивысшим значением *TF-IDF*.

4.3. Поиск точки бифуркации (ветвления) внутри макротемы. Точка ветвления выявляется путем последовательного сравнения и поиска существенных изменений главных ключевых слов во всех 30 временных промежутках. Изменение существенно, если не менее трех из пяти главных ключевых слов рассматриваемого промежутка различаются с прежними и изменение наблюдается не менее чем в двух последовательных временных промежутках (данное условие позволяет пропускать всплески специфичных сообщений, фокусируясь на общей динамике дискуссии). Число слов в данном случае выбирается эмпирически с ручной оценкой качества и количества разбиений. Вопрос выбора оптимального количества изменяющихся ключевых слов для поиска точек бифуркации находится на этапе исследования, так как зависит от большого ряда параметров. В число таких параметров, кроме доли изменяющихся ключевых слов, входит кодировщик текстов, алгоритм выделения ключевых слов, алгоритм выделения минимального размера окна.

4.4. Тематическое моделирование сообщений, следующих после точки ветвления. Повторяем п. 2 для сообщений, находящихся во временном интервале после точки бифуркации. Далее повторяем пп. 3.1–3.5. Процесс происходит рекурсивно до тех пор, пока алгоритм поиска точек бифуркации не перестанет обнаруживать точки внутри каждой подтемы.

5. Определение «толщины ветвей» путем дифференциации подтем по количеству сообщений в них. «Толщина ветви» на визуализации (см. рис. 2) варьирует в диапазоне от 10 пунктов (для ≤ 20 сообщений в «ветви») до 80 пунктов (для ≥ 300 сообщений в «ветви») с шагом в 10 пунктов (восемь возможных вариантов «толщины ветви»). Длина «ветви» определяется временем появления точки бифуркации и временем окончания обсуждения (на рис. 2 точки с точным временем завершения подтем не обозначаются, т. к. совпадают в пределах дня). Итоговая репрезентация онлайн-дискуссии имеет древовидную структуру, в основе («стволе») которой находится корпус текстов и его основные темы, а в «ветвях» представляются репрезентации тем, соответствующих минимальному разбиению.

6. Суммаризация «ветвей». Для каждой «ветви» на базе нейромодели *financial-summarization-pegasus* строится абстрактивная суммаризация для самых значимых сообщений «ветви» (если сообщений больше 100 – относящихся к теме с вероятностью 0,9 и выше). Претест метода (один способ суммаризации, так называемая «книжная модель») проводился нами ранее (Blekanov, Tarasov, Bodrunova, Sergeev, 2023) и показал, что важным параметром метода является заданная длина суммаризации; выбранная нами ранее длина приводила к частичной потере смысла за счет «углубленности в контекст» и утраты опознаваемой связи между предложениями в суммарном тексте. Краткие суммаризации (по одной на «ветвь») помогли показать общее течение дискуссии, но также не выявили разных мнений пользователей, так как выглядели в основном как заголовки (см. рис. 2). Поэтому в данной работе для одной из макротем под условным названием «финансы» мы сопоставили три опции суммаризации с различной длиной – краткую, длиной не менее 32 слов (2–3 предложения) и «книжную модель» – и оценили типы мнений, полученные в них. Для этого мы дополнительно разбили изучаемые «ветви» рассматриваемой темы на фрагменты фиксированной длины (для демонстрации различий трех моделей суммаризации была выбрана длина фрагмента в 10 сообщений), используя идею «цепочки тем» и превратив ее в «цепочку суммаризаций».

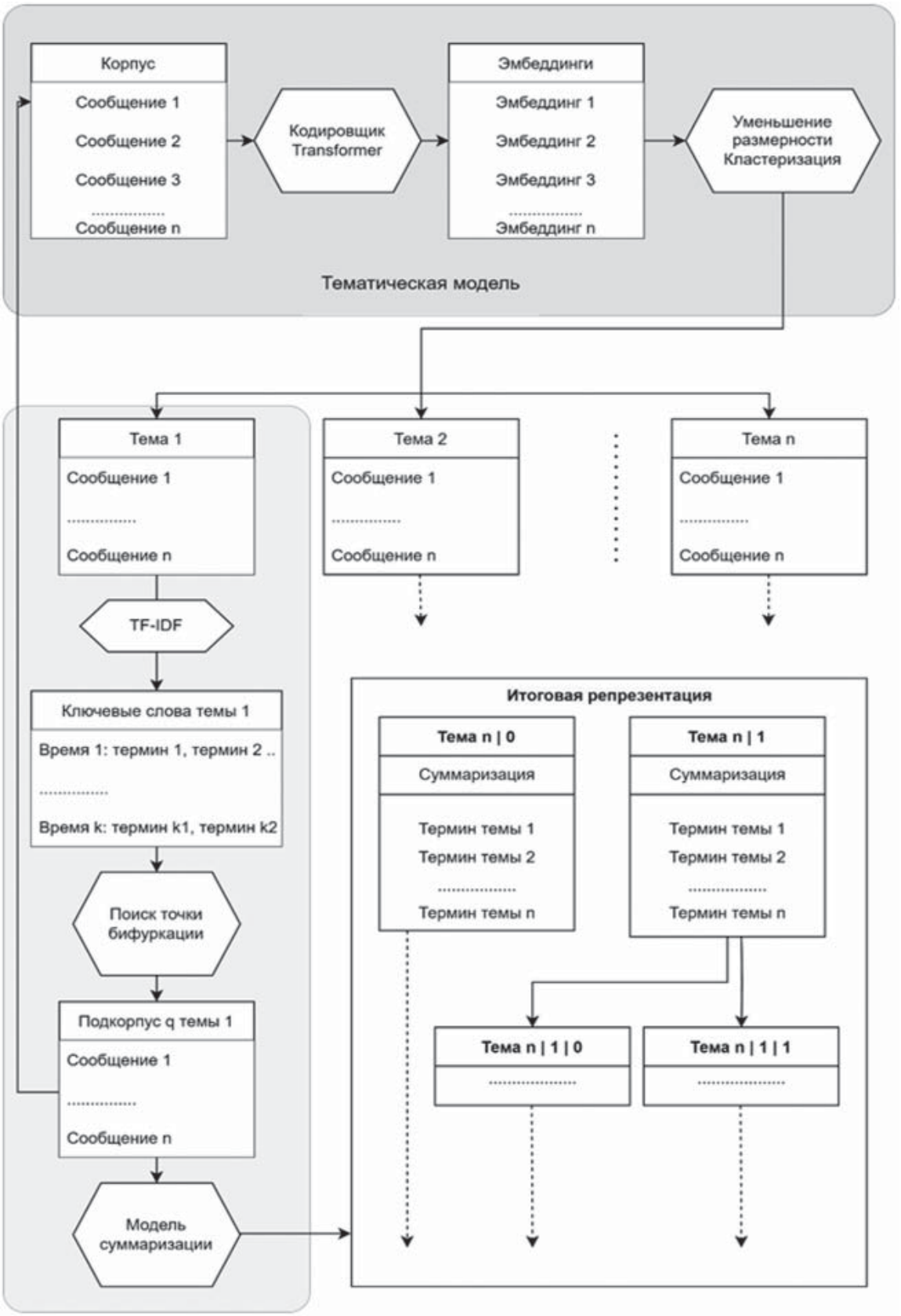


Рисунок 1. Общая схема работы метода (pipeline)

Результаты моделирования и суммаризации

Поскольку дискуссия велика, мы представим результаты мэппинга для небольшой ее части – макротемы «финансы» (датасет 2). Рисунок 2 «углубляется» в нее и картирует ее «ветви» в самом кратком виде (суммаризация каждой «ветви» в одном предложении), что позволяет охватить макротему (и наш метод) одним взглядом. Рисунок 3 представляет три подхода к суммаризации с выявленными мнениями пользователей на примере еще меньшего участка датасета 2 – совсем небольших «ветвей» 1|0 (22 сообщения) и 1|1 (36 сообщений) по тематике майнинга криптовалюты и его влияния на изменение климата. Обсудим отдельно итоги мэппинга тем и итоги поиска оптимальной суммаризации.

Мэппинг тем. В результате тематического моделирования датасета 1 на основе дообученного *BERTopic* были получены 72 темы (пп. 2.1–2.3 метода), затем сгруппированные в 10 макротем (п. 2.4). На рисунке 2 в иллюстративных целях представлена детализация одной (из 10 полученных) макротемы «финанасы» датасета 1, состоящей из пяти тем (0–4) с указанием пяти ключевых слов и числа сообщений для каждой.

Стоит обратить внимание на минимальное пересечение главных ключевых слов общей темы с ключевыми словами отдельных ее подтем в рамках макротемы «финансы». Это указывает на высокое качество разработанной тематической модели с точки зрения ее способности улавливать контекст подтемы и тем самым находить точки бифуркации.

Качество суммаризации. Как видно из рисунка 3, разные подходы к суммаризации дают разные результаты для мэппинга мнений. Краткие суммаризации могут улавливать «свернутые» мнения (мы не знаем, какие именно мнения стоят за указаниями на то, что дискуссия имеет место): *proof-of-stake consensus* (консенсус о том, что следует учитывать вклад [в трату энергии вместо доказательства участия в работах]), *considering cracking down* (рассматривает возможность борьбы [с энергоемким майнингом биткоина]), а также яркие сложившиеся позиции – *has been a total waste* ([электричество] потрачено полностью напрасно). Короткие суммаризации можно использовать для того, чтобы оценить качество выделения тем: например, видно, что тема 1|0 в первом столбце не поменяла содержание, т. е. гарантированно не давала точек бифуркации. При этом тема 1|1 не давала ветвления при моделировании (и сообщения действительно относятся к теме «биткоин *vs* изменение климата»), но дала шанс проследить основные обсуждаемые события.

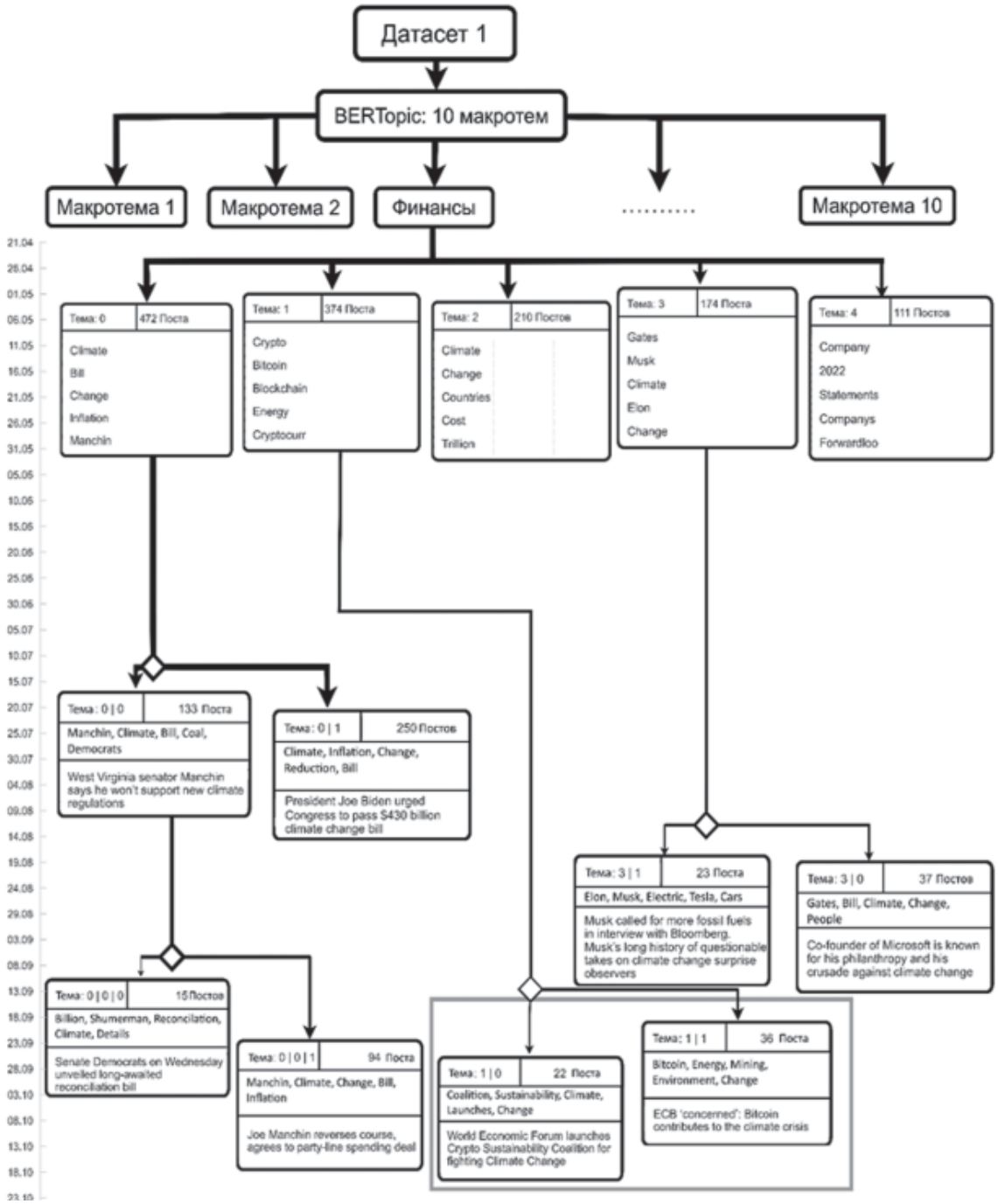


Рисунок 2. Визуализация результатов работы метода на макротеме «финансы»

Примечание. Ромбами обозначены точки бифуркации. Серой линией обведены подтемы 1|0 и 1|1, для которых ниже приведены примеры суммаризаций (см. рис. 3).

«Ветвь»	Подход к суммаризации		
	Краткая	Длиной не менее 32 слов	«Книжная модель»
1|0	World Economic Forum launches Crypto Sustainability Coalition.	Web3 technologies could be used to fight climate change, WEF says. WEF launches Crypto Sustainability Coalition to leverage Web3 technologies in change battle. Controversial new report on climate change to be published by White House.	New White House report on climate change could be bullish for these two coins. The weekend shift: Crypto, Billionaires, Climate Change ~~nans~~ New Crypto Sustainability Coalition to investigate potential of web3 technologies in fighting climate change - ~~Eurissaa~~ Review. Wef Launches Crypto Sustainability Coalition to ~~Convince Web3 Technologies to Fight Carbon Change Blockchain News~~. The World Economic Forum announces that it has launched a new group of "crypto sustainability co-ops" to fight climate change. They're joined by Stellarlumens and Aaronchi, who are joining forces to combat climate change with their new technology.
	World Economic Forum launches Crypto Sustainability Coalition with Stellar Lumens and Avalanche.	World Economic Forum (WEF) launches Crypto Sustainability Coalition with Stellar Lumens and Avalanche. Web3 technologies could be used to fight climate change. New York financial regulator says.	
1|1	Climate groups want Bitcoin to follow Ethereum in switching to Proof-of-Stake consensus.	~~Cardano and ban of PoW~~ President Biden signed executive order 14067 to support responsible digital asset development. Greenpeace and other climate groups start a big campaign to fight Bitcoin.	The ~~PoW~~ President has signed an executive order to support responsible, "artificial asset development" in line with the goals of the United States. In the document, he says that it could mean ban on Btc mining. According to this report, cardano is among the largest proof-of-stake blockchains around the world ~~and~~ potential positive effects of the Merge. ~~It could also mean a banning on Bitcoin~~. Because ~~miners~~ will no longer be financially incentived to run machines around ~~theclock~~, the network\'s energy consumption will drop by over 99% according to scientists at the Ethereum foundation. A new advertising campaign is underway against Bitcoin's climate change and other climate change groups. They are planning to launch a new campaign called "Change the code not the climate." This time, they want to use bitcoin as a tool to combat climate change. Their goal is to convince people to switch from proof of work to proof of stake so that they can mine more efficiently. As predicted, ~~ETH~~ price may grow to $5,000 after switching to the pox of the Ethereum chain. Roman ~~Nekrastrov~~, another founder of the ~~Enchry~~ Foundation, predicts that ether will reach $5,000 in the next month. He advises owners of the coin to host the token in the following three months. After these predictions, ethereal prices will rise to $5,000. Does Bitcoin really affect the climate change?
	Environmentalists point out that Bitcoin uses as much energy as a small country.	Environmentalists point out that Bitcoin uses as much energy as a small country. Bitcoin is not an enemy, but an ally in the fight against climate change: Harald Rauter, in /r/Bitcoin.	
	EU considering cracking down on energy-intensive crypto mining.	EU is considering a bill to limit the mining of energy-hungry cryptocurrencies. Bitcoin is down 1.1% in Asian morning deals on Wednesday, while Ethereum is down 1.9%.... Bitcoin energy consumption has jumped 41% in the past year.	
	Electricity used to process transactions has been a total waste.	Electricity used to process transactions has been a total waste. Is Bitcoin Cash 'indestructible' if governments crack down on PoW? Is it immune to government regulation if they crack down on PoW?	

Рисунок 3. Подходы к суммаризации и мнения в них («ветви» 1|0 и 1|1)

Примечание. Серый с подчеркиванием: указание на мнение; пунктир: выраженное мнение; темно-серый: полярное мнение («за/против»); вычеркнуто: контекстуальная неполнота, ошибки аббревиации, словоупотребления, грамматики.

Суммаризации средней длины и «книжные» суммаризации улавливают также более развернутые мнения, в т. ч. формулируя их в виде вопросов, отражая и ход событий. Как видно из рисунка 3, «средние» суммаризации уловили поляризацию взглядов: одни пользователи процитировали Харальда Раутера из сабреддита (выделенной дискуссии на *Reddit*) *Bitcoin* о том, что биткоин – союзник в борьбе против изменения климата, тогда как другие указали на напрасную трату энергии при майнинге. Также удалось увидеть, что дискуссия заканчивается постановкой важных вопросов, что необычно для сетевых обсуждений. Удалось уловить и почти все «свернутые мнения», за исключением важного вопроса о переходе от *proof-of-work* к *proof-of-state* при учете влияния отдельного майнера. Этот вопрос, однако, отражен в «книжной» версии суммаризации. Она подробнее отражает ход главных обсуждаемых в «ветке» событий и дает важный объяснительный контекст. Но в то же время в ней достаточно много «темных» мест, ошибок, нераскрытых аббревиатур. Модель уловила «противостояние, выраженное в действиях» (*campaign... “Change the*

code not the climate" – «кампания... "Меняй код, а не климат"»), но не уловила противостояния во мнениях. Таким образом, средняя длина суммаризации в рамках «цепочки мнений» видится оптимальным на данный момент способом уловить не только ход событий, но и мнения в дискуссии.

Метод позволил обнаружить паттерны, описываемые концепцией кумулятивной делиберации: накопление вопросов и поляризацию взглядов. Но наиболее значимым результатом стало предложенное машиной обозначение отсутствия консенсуса в конце ветви 1|1 через риторические вопросы (обозначены серым в двух моделях суммаризации). Использование моделью риторического вопроса как приема для демонстрации нерешенности вопроса, вероятно, проистекает из обучения машины на текстах СМИ, где журналисты используют подобные приемы для описания камня преткновения в конфликте.

Дискуссия и заключение

Представленный нами метод мэппинга содержания онлайн-дискуссии, названный «деревом мнений», в целом позволяет проследить тематику обсуждения, ход обсуждаемых событий и мнения – как пользователей, так и упоминаемых акторов конфликта. Метод задействует передовые методики нейросетевого анализа для тематического моделирования и суммаризации текстов. Его применение позволяет отслеживать паттерны накопления мнений и указывает на различные форматы выражения мнений в онлайн-обсуждениях, а также указывает на отсутствие консенсуса в дискуссиях. При применении на разных платформах он может показать влияние платформенных ограничений на динамику мнений.

При наличии полного датасета на начальном этапе моделирования выделяется спектр макротем; при мэппинге в реальном времени (с небольшим запаздыванием для сбора данных) на вход подаются данные за первый по времени фрагмент дискуссии. Для мэппинга мнений можно использовать суммаризации разной длины – в зависимости от потребностей ученых.

Ограничения метода, как и возможности его развития, связаны с несколькими проблемами.

Во-первых, разбиение на макротемы может давать темы, выделенные на разных основаниях (например, «финансы» и «консервативный дискурс»). Это «встроенное» ограничение тематического моделирования вряд ли удастся преодолеть; но для прослеживания мнений это может быть даже полезно.

Во-вторых, это вопрос пороговых значений при моделировании тем и выборе «окна» (объема текстов в «цепочке мнений»). Поиск точки бифуркации может быть более и менее «тонким»; пороги принадлежности текста конкретной «ветви» также могут быть заданы по-разному. Это наращивает гибкость метода для более точной суммаризации и для разных задач. Но это же делает метод более произвольным и зависимым от исследователя (хотя основные параметры моделирования выбирает машина).

В-третьих, нами не до конца решена проблема суммирования очень длинных постов. Для таких постов в дальнейшем мы протестируем гибридную экстрактивно-абстрактивную модель суммаризации.

В-четвертых, посты и комментарии можно различать, моделируя дискуссию еще более точно и суммируя ответы пользователей на основной контент.

В-пятых, ме тод также может быть расширен через построение связи тем/мнений с их носителями, т. е. через прослеживание разветвленных тем в онлайн-сообществах (Lin, Mei, Han, Jiang et al., 2011). В работах по эволюции тем пока не задействуются данные о социально-групповых чертах пользователей. Соединение «дерева мнений» с такими данными может дать уникальную по своей точности картину накопления мнений в «Солярисе» интернет-среды.

Однако, несмотря на указанные ограничения, метод демонстрирует возможности искусственного интеллекта по представлению онлайн-дискуссий для их обзора «с высоты птичьего полета». Это позволяет проследить динамику обобщенных позиций пользователей, выявить типы мнений и их поляризацию, а также подтвердить наличие или отсутствие значимого консенсуса по важным для общественности вопросам повестки дня.

Примечания

[1] Исследование выполнено за счет средств гранта «Накопительное общественное мнение и кумулятивная делиберация: теоретизация, методы анализа и оценка качества современного делиберативного процесса: 2025 г. этап 1» № 128786104.

Библиография

Бодрунова С. С. Кумулятивная делиберация: новая нормативность в изучении публичных сфер онлайн // Вестник Московского университета. Серия 10. Журналистика. 2023. №1. С. 87–122.

Головнин О. К., Курганов А. В. Автоматизированная система анализа комментариев пользователей в службе мгновенного обмена сообщениями Telegram // Труды Международной научно-технической конференции «Перспективные информационные технологии». Самара, 2022. С. 23–25.

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.

Митрофанова О. А., Адамова М. А., Букреева Л. А., Голубев Р. В. и др. Интеллектуальный анализ данных в корпусе текстов по корпусной и компьютерной лингвистике // International Journal of Open Information Technologies. 2024. № 12. С. 11–26.

Сорокина С. Г. Интеллектуальная обработка текстовой информации: обзор автоматизированных методов суммаризации // Виртуальная коммуникация и социальные сети. 2024. № 3. С. 203–222.

Abulaish M., Fazil M. (2018) Modeling topic evolution in Twitter: An embedding-based approach. *IEEE Access* 6: 64847–64857.

Ahmed A., Xing E. P. (2012) *Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream*. arXiv:1203.3463.

Alam M. H., Ryu W. J., Lee S. (2017) Hashtag-based topic evolution in social media. *World Wide Web* 20: 1527–1549.

Allan J., Carbonell J. G., Doddington G., Yamron J., Yang Y. (1998) Topic detection and tracking pilot study final report. Available at: people.cs.pitt.edu/~chang/265/proj10/sisref/1.pdf (accessed: 31.10.2025).

Alsaedi N., Burnap P., Rana O. (2017) Can we predict a riot? Disruptive event detection using Twitter. *ACM Transactions on Internet Technology* 17 (2): 1–26.

AlSumait L., Barbará D., Domeniconi C. (2008) On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *8th IEEE international conference on data mining*. IEEE. Pp. 3–12.

Blei D. M., Lafferty J. D. (2006) Dynamic topic models. In *23rd International conference on machine learning*. ACM. Pp. 113–120.

Blekanov I. S., Tarasov N., Bodrunova S. S. (2022) Transformer-based abstractive summarization for Reddit and Twitter: Single posts vs. comment pools

in three languages. *Future Internet 14* (3). Available at: https://www.mdpi.com/1999-5903/14/3/69/pdf (accessed: 31.10.2025).

Blekanov I. S., Tarasov N., Bodrunova S. S., Sergeev S. L. (2023) Mapping opinion cumulation: Topic modeling-based dynamic summarization of user discussions on social networks. In *International Conference on Human-Computer Interaction*. Cham: Springer. Pp. 25–40.

Bohman I. J., Rehg W. (1997) *Deliberative Democracy.* Cambridge: MIT Press.

Brigadir I., Greene D., Cunningham P. (2014) *Adaptive representations for tracking breaking news on Twitter*. arXiv:1403.2923.

Cai H., Tang Z., Yang Y., Huang Z. (2014) Eventeye: Monitoring evolving events from tweet streams. In *22nd ACM International conference on multimedia*. ACM. Pp. 747–748.

Chang Y., Tang J., Yin D., Yamada M., Liu Y. (2016) Timeline Summarization from Social Media with Life Cycle Models. In *IJCAI-16 Proceedings*. AAAI. Pp. 3698–3704.

Chen G., Xu N., Mao W. (2018) An encoder-memory-decoder framework for sub-event detection in social media. In *27th ACM international conference on information and knowledge management*. ACM. Pp. 1575–1578.

Dermouche M., Velcin J., Khouas L., Loudcher S. (2014) A joint model for topic-sentiment evolution over time. In *2014 IEEE International conference on data mining*. IEEE. Pp. 773–778.

Gao W., Peng M., Wang H., Zhang Y., Han W., Hu G., Xie Q. (2020) Generation of topic evolution graphs from short text streams. *Neurocomputing* 383: 282–294.

Goyal P., Kaushik P., Gupta P., Vashisth D., Agarwal S., Goyal N. (2020) Multilevel event detection, storyline generation, and summarization for tweet streams. *IEEE Transactions on Computational Social Systems* 7(1): 8–23.

Grootendorst M. (2022) *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv:2203.05794.

Habermas J. (1996) *Between facts and norms: Contributions to a discourse theory of law and democracy.* Boston: MIT Press.

Hong L., Yin D., Guo J., Davison B. D. (2011) Tracking trends: incorporating term volume into temporal topic models. In *17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. Pp. 484–492.

Huang J., Peng M., Wang H., Cao J., Gao W., Zhang X. (2017) A probabilistic method for emerging topic tracking in microblog stream. *World Wide Web* 20: 325–350.

Jo Y., Hopcroft J. E., Lagoze C. (2011) The web of topics: discovering the topology of topic evolution in a corpus. In *20th International conference on world wide web*. ACM. Pp. 257–266.

Kawamae N. (2018) Topic chronicle forest for topic discovery and tracking. In *11th ACM International conference on web search and data mining*. ACM. Pp. 315–323.

Kim D., Oh A. (2011) Topic Chains for Understanding a News Corpus. In: Gelbukh, A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2011. Lecture Notes in Computer Science, vol 6609. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-19437-5_13

King A. A., Anderson F. D. (1971) Nixon, Agnew, and the 'silent majority': A case study in the rhetoric of polarization. *Western Speech* 35 (4): 243–255.

Koch T., Arendt F., Maximilian L. (2017) Media effects: Cumulation and duration. In P. R ssler, C. A. Hoffner, L. Zoonen (eds.) *The International Encyclopedia of Communication*. DOI: 10.1002/9781118783764.wbieme0217.

Lee P., Lakshmanan L. V., Milios E. E. (2013) *Event evolution tracking from streaming social posts.* arXiv:1311.5978.

Lee P., Lakshmanan L. V., Milios E. E. (2014) Incremental cluster evolution tracking from highly dynamic network data. In *2014 IEEE 30th International conference on data engineering.* IEEE. Pp. 3–14.

Li J., Li S. (2013) Evolutionary hierarchical dirichlet process for timeline summarization. In *51st Annual Meeting of the Association for Computational Linguistics.* Volume 2: Short Papers. ACL. Pp. 556–560.

Lin C. X., Mei Q., Han J., Jiang Y., Danilevsky M. (2011) The joint inference of topic diffusion and evolution in social communities. In *2011 IEEE 11th International conference on data mining* IEEE. Pp. 378–387.

Lu X., Guo Y., Chen J., Wang F. (2022) Topic change point detection using a mixed bayesian model. *Data Mining and Knowledge Discovery* 36: 146–173.

Momeni E., Karunasekera S., Goyal P., Lerman K. (2018) Modeling evolution of topics in large-scale temporal text corpora. *International AAAI Conference on web and social media* 1 (12): 656–659.

Mouffe C. (2000) *The Democratic Paradox.* New York: Verso.

Noelle Neumann E. (1974) The spiral of silence a theory of public opinion. *Journal of Communication* 24 (2): 43–51.

Passali T., Gidiotis A., Chatzikyriakidis E., Tsoumakas G. (2021) Towards human-centered summarization: A case study on financial news. In *First workshop on bridging human-computer interaction and natural language processing* ACL. Pp. 21–27.

Pfetsch B. (2018) Dissonant and disconnected public spheres as challenge for political communication research. *Javnost – The Public* 25 (1–2): 59–65.

Saha A., Sindhwani V. (2012) Learning evolving and emerging topics in social media: A dynamic NMF approach with temporal regularization. In *5th ACM International conference on web search and data mining* ACM. Pp. 693–702.

Song J., Huang Y., Qi X., Li Y., Li F., Fu K., Huang T. (2016) Discovering hierarchical topic evolution in time stamped documents. *Journal of the Association for Information Science and Technology* 67 (4): 915–927.

Srijith P. K., Hepple M., Bontcheva K., Preotiuc-Pietro D. (2017) Sub-story detection in Twitter with hierarchical Dirichlet processes. *Information Processing & Management* 53 (4): 989–1003.

Tang X., Yang C. C. (2011) Following the social media: Aspect evolution of online discussion. In *Social Computing, Behavioral-Cultural Modeling and Prediction: 4th International Conference Proceedings*. Springer. Vol. 4. Pp. 292–300.

Wang Z., Chen J., Chen J., Chen H. (2024) Identifying interdisciplinary topics and their evolution based on BERTopic. *Scientometrics* 129: 7359–7384. DOI: 10.1007/s11192-023-04776-5.

Wang Z., Shou L., Chen K., Chen G., Mehrotra S. (2015) On summarization and timeline generation for evolutionary tweet streams. *IEEE Transactions on Knowledge and Data Engineering* 27 (5): 1301–1315.

Wang P., Zhang P., Zhou C., Li Z., Yang H. (2017) Hierarchical evolving Dirichlet processes for modeling nonlinear evolutionary traces in temporal data. *Data Mining and Knowledge Discovery* 31: 32–64.

Yardi S., boyd d. (2010) Dynamic debates: An analysis of group polarization over time on Twitter. *Bulletin of Science, Technology & Society* 30 (5): 316–327.

Zhang Z., Fang M., Chen L., Namazi-Rad M. R. (2022) *Is neural topic modelling better than clustering? An empirical study on clustering with contextual embeddings for topics*. arXiv:2204.09874.

Zhou H., Yu H., Hu R., Hu J. (2017) A survey on trends of cross-media topic evolution map. *Knowledge-Based Systems* 124: 164–175.

Zhou Y., Kanhabua N., Cristea A. I. (2016) Real-time timeline summarisation for high-impact events in Twitter. In *22nd European conference on artificial intelligence*. IOS Press. Pp. 1158–1166.

Bibliography

Abulaish M., Fazil M. (2018) Modeling topic evolution in Twitter: An embedding-based approach. *IEEE Access* 6: 64847–64857.

Ahmed A., Xing E. P. (2012) *Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream*. arXiv:1203.3463.

Alam M. H., Ryu W. J., Lee S. (2017) Hashtag-based topic evolution in social media. *World Wide Web* 20: 1527–1549.

Allan J., Carbonell J. G., Doddington G., Yamron J., Yang Y. (1998) Topic detection and tracking pilot study final report. Available at: people.cs.pitt.edu/~chang/265/proj10/sisref/1.pdf (accessed: 31.10.2025).

Alsaedi N., Burnap P., Rana O. (2017) Can we predict a riot? Disruptive event detection using Twitter. *ACM Transactions on Internet Technology* 17 (2): 1–26.

AlSumait L., Barbará D., Domeniconi C. (2008) On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *8th IEEE international conference on data mining*. IEEE. Pp. 3–12.

Blei D. M., Lafferty J. D. (2006) Dynamic topic models. In *23rd International conference on machine learning*. ACM. Pp. 113–120.

Blekanov I. S., Tarasov N., Bodrunova S. S. (2022) Transformer-based abstractive summarization for Reddit and Twitter: Single posts vs. comment pools in three languages. *Future Internet 14* (3). Available at: https://www.mdpi.com/1999-5903/14/3/69/pdf (accessed: 31.10.2025).

Blekanov I. S., Tarasov N., Bodrunova S. S., Sergeev S. L. (2023) Mapping opinion cumulation: Topic modeling-based dynamic summarization of user discussions on social networks. In *International Conference on Human-Computer Interaction*. Cham: Springer. Pp. 25–40.

Bodrunova S. S. (2023) Kumulyativnaya deliberatsiya: novaya normativnost' v izuchenii publichnykh sfer onlayn [Cumulative deliberation: new normativity in the study of online public spheres]. *Vestnik Moskovskogo universiteta. Seriya 10. Zhurnalistika* 1: 87–122. (In Russian)

Bohman I. J., Rehg W. (1997) *Deliberative Democracy.* Cambridge: MIT Press.

Brigadir I., Greene D., Cunningham P. (2014) *Adaptive representations for tracking breaking news on Twitter*. arXiv:1403.2923.

Cai H., Tang Z., Yang Y., Huang Z. (2014) Eventeye: Monitoring evolving events from tweet streams. In *22nd ACM International conference on multimedia*. ACM. Pp. 747–748.

Chang Y., Tang J., Yin D., Yamada M., Liu Y. (2016) Timeline Summarization from Social Media with Life Cycle Models. In *IJCAI-16 Proceedings*. AAAI. Pp. 3698–3704.

Chen G., Xu N., Mao W. (2018) An encoder-memory-decoder framework for sub-event detection in social media. In *27th ACM international conference on information and knowledge management.* ACM. Pp. 1575–1578.

Dermouche M., Velcin J., Khouas L., Loudcher S. (2014) A joint model for topic-sentiment evolution over time. In *2014 IEEE International conference on data mining.* IEEE. Pp. 773–778.

Gao W., Peng M., Wang H., Zhang Y., Han W., Hu G., Xie Q. (2020) Generation of topic evolution graphs from short text streams. *Neurocomputing* 383: 282–294.

Golovnin O. K., Kurganov A. V. (2022) Avtomatizirovannaya sistema analiza kommentariyev pol'zovateley v sluzhbe mgnovennogo obmena soobshcheniyami Telegram [Automated system for analyzing user comments in the instant messaging service Telegram]. In *Proceedings of the International Scientific and Technical Conference "Advanced Information Technologies"*. Samara. Pp. 23–25. (In Russian)

Goyal P., Kaushik P., Gupta P., Vashisth D., Agarwal S., Goyal N. (2020) Multilevel event detection, storyline generation, and summarization for tweet streams. *IEEE Transactions on Computational Social Systems* 7(1): 8–23.

Grootendorst M. (2022) *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv:2203.05794.

Habermas J. (1996) *Between facts and norms: Contributions to a discourse theory of law and democracy.* Boston: MIT Press.

Hong L., Yin D., Guo J., Davison B. D. (2011) Tracking trends: incorporating term volume into temporal topic models. In *17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. Pp. 484–492.

Huang J., Peng M., Wang H., Cao J., Gao W., Zhang X. (2017) A probabilistic method for emerging topic tracking in microblog stream. *World Wide Web* 20: 325–350.

Jo Y., Hopcroft J. E., Lagoze C. (2011) The web of topics: discovering the topology of topic evolution in a corpus. In *20th International conference on world wide web*. ACM. Pp. 257–266.

Kawamae N. (2018) Topic chronicle forest for topic discovery and tracking. In *11th ACM International conference on web search and data mining*. ACM. Pp. 315–323.

Kim D., Oh A. (2011) Topic Chains for Understanding a News Corpus. In: Gelbukh, A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2011. Lecture Notes in Computer Science, vol 6609. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-19437-5_13

King A. A., Anderson F. D. (1971) Nixon, Agnew, and the 'silent majority': A case study in the rhetoric of polarization. *Western Speech* 35 (4): 243–255.

Koch T., Arendt F., Maximilian L. (2017) Media effects: Cumulation and duration. In P. Rössler, C. A. Hoffner, L. Zoonen (eds.) *The International Encyclopedia of Communication*. DOI: 10.1002/9781118783764.wbieme0217.

Kol'tsova O. Yu., Maslinskiy K. A. (2013) Vyyavlenie tematicheskoy struktury rossiyskoy blogosfery: avtomaticheskie metody analiza tekstov [Identification of the thematic structure of the Russian blogosphere: automatic text analysis methods]. *Sotsiologiya: metodologiya, metody, matematicheskoe modelirovanie* 36: 113–139. (In Russian)

Lee P., Lakshmanan L. V., Milios E. E. (2013) *Event evolution tracking from streaming social posts*. arXiv:1311.5978.

Lee P., Lakshmanan L. V., Milios E. E. (2014) Incremental cluster evolution tracking from highly dynamic network data. In *2014 IEEE 30th International conference on data engineering*. IEEE. Pp. 3–14.

Li J., Li S. (2013) Evolutionary hierarchical dirichlet process for timeline summarization. In *51st Annual Meeting of the Association for Computational Linguistics*. Volume 2: Short Papers. ACL. Pp. 556–560.

Lin C. X., Mei Q., Han J., Jiang Y., Danilevsky M. (2011) The joint inference of topic diffusion and evolution in social communities. In *2011 IEEE 11th International conference on data mining* IEEE. Pp. 378–387.

Lu X., Guo Y., Chen J., Wang F. (2022) Topic change point detection using a mixed bayesian model. *Data Mining and Knowledge Discovery* 36: 146–173.

Mitrofanova O. A., Adamova M. A., Bukreeva L. A., Golubev R. V., Gusyatskaya P. A., Zernova A. K., Litvinova A. A., Makeev K. V., Pavlikova V. S., Plyusnina E. P., Sologub P. Yu., Sukhan D. D., Troshina A. V., Utkina A. A. (2024) Intellektual'nyy analiz dannykh v korpuse tekstov po korpusnoy i komp'yuternoy lingvistike [Data mining in a corpus of texts on corpus and com-

putational linguistics]. *International Journal of Open Information Technologies* 12: 11–26. (In Russian)

Momeni E., Karunasekera S., Goyal P., Lerman K. (2018) Modeling evolution of topics in large-scale temporal text corpora. *International AAAI Conference on web and social media* 1 (12): 656–659.

Mouffe C. (2000) *The Democratic Paradox*. New York: Verso.

Noelle Neumann E. (1974) The spiral of silence a theory of public opinion. *Journal of Communication* 24 (2): 43–51.

Passali T., Gidiotis A., Chatzikyriakidis E., Tsoumakas G. (2021) Towards human-centered summarization: A case study on financial news. In *First workshop on bridging human-computer interaction and natural language processing* ACL. Pp. 21–27.

Pfetsch B. (2018) Dissonant and disconnected public spheres as challenge for political communication research. *Javnost – The Public* 25 (1–2): 59–65.

Saha A., Sindhwani V. (2012) Learning evolving and emerging topics in social media: A dynamic NMF approach with temporal regularization. In *5th ACM International conference on web search and data mining* ACM. Pp. 693–702.

Song J., Huang Y., Qi X., Li Y., Li F., Fu K., Huang T. (2016) Discovering hierarchical topic evolution in time stamped documents. *Journal of the Association for Information Science and Technology* 67 (4): 915–927.

Sorokina S. G. (2024) Intellektual'naya obrabotka tekstovoy informatsii: obzor avtomatizirovannykh metodov summarizatsii [Intelligent text processing: a review of automated summarization methods]. *Virtual'naya kommunikatsiya i sotsial'nye seti* 3: 203–222. (In Russian)

Srijith P. K., Hepple M., Bontcheva K., Preotiuc-Pietro D. (2017) Sub-story detection in Twitter with hierarchical Dirichlet processes. *Information Processing & Management* 53 (4): 989–1003.

Tang X., Yang C. C. (2011) Following the social media: Aspect evolution of online discussion. In *Social Computing, Behavioral-Cultural Modeling and Prediction: 4th International Conference Proceedings*. Springer. Vol. 4. Pp. 292–300.

Wang P., Zhang P., Zhou C., Li Z., Yang H. (2017) Hierarchical evolving Dirichlet processes for modeling nonlinear evolutionary traces in temporal data. *Data Mining and Knowledge Discovery* 31: 32–64.

Wang Z., Chen J., Chen J., Chen H. (2024) Identifying interdisciplinary topics and their evolution based on BERTopic. *Scientometrics* 129: 7359–7384. DOI: 10.1007/s11192-023-04776-5.

Wang Z., Shou L., Chen K., Chen G., Mehrotra S. (2015) On summarization and timeline generation for evolutionary tweet streams. *IEEE Transactions on Knowledge and Data Engineering* 27 (5): 1301–1315.

Yardi S., boyd d. (2010) Dynamic debates: An analysis of group polarization over time on Twitter. *Bulletin of Science, Technology & Society* 30 (5): 316–327.

Zhang Z., Fang M., Chen L., Namazi-Rad M. R. (2022) *Is neural topic modelling better than clustering? An empirical study on clustering with contextual embeddings for topics*. arXiv:2204.09874.

Zhou H., Yu H., Hu R., Hu J. (2017) A survey on trends of cross-media topic evolution map. *Knowledge-Based Systems* 124: 164–175.

Zhou Y., Kanhabua N., Cristea A. I. (2016) Real-time timeline summarisation for high-impact events in Twitter. In *22nd European conference on artificial intelligence*. IOS Press. Pp. 1158–1166.

Поступила в редакцию
23.06.2025