ТЕОРИЯ И МЕТОДОЛОГИЯ

Сергей Петрович Сидоров

доктор физико-математических наук,
профессор Саратовского национального исследовательского
государственного университета имени Н.Г. Чернышевского,
Саратов, Россия;
e-mail: sidorovsp@yahoo.com

Софья Владимировна Тихонова

доктор философских наук,
профессор Саратовского национального исследовательского
государственного университета имени Н.Г. Чернышевского,
Саратов, Россия;
e-mail: segedasv@yandex.ru

Инструментальные методы анализа медиапространства в цифровой гуманитаристике

УДК: 168.522:001.895
DOI: 10.24412/2079-0910-2023-3-118-131

Статья посвящена анализу места и роли инструментальных математических методов в методологии цифровой гуманитаристики. Авторы исследуют их потенциал с точки зрения преодоления фрагментарности цифровой методологии за счет широты применимости метода. «Достаточная» широта инструментальных методов, способная усилить положительную конвергенцию цифровой методологии, обеспечивается их применимостью к самым разным аспектам новостных потоков медиапространства. Под новостным потоком понимается совокупность новостей, генерируемых: а) информационными агентствами, б) предварительными материалами первичных источников, в) социальными сетями. Авторы рассматривают инструментальные средства анализа как разновидность систем машинного обучения, приме-

няемых для определения трендов медиапространства. Их применимость ориентирована на количественную оценку текстовых сообщений (характер, влияние, релевантность, новизна), а также различные по степени сложности формы сентимент-анализа, позволяющие отразить контекст новостного сообщения, положительный, отрицательный или нейтральный. Для этого используется или расчет индекса DISAG расхождения в оценке сообщения, или такие модели машинного обучения, как наивный байесовский классификатор, логистическая регрессия, композиции деревьев решений, полносвязные нейросети, сверточные нейросети, рекуррентные нейросети. Особое внимание уделено вспомогательным базам данных — словарям, лексикону и грамматике, а также библиотекам подпрограмм, предназначенным для выполнения задач, связанных с анализом текстов, и агрегаторам новостного потока. Авторы приходят к выводу о том, что глубокий анализ количественных характеристик тех или иных новостных потоков или взаимодействий пользователей социальных сетей позволяет решать типовые задачи в основных областях цифровой гуманитаристики, тем самым способствуя унификации ее методологии.

Ключевые слова: цифровая гуманитаристика, цифровая методология, медиапространство, новостной поток, инструментальные методы анализа.

Благодарность

Исследование выполнено при финансовой поддержке Российского научного фонда (РНФ) в рамках научного проекта № 22-18-00153 «Образ СССР в исторической памяти: исследование медиастратегий воспроизводства представлений о прошлом в России и зарубежных странах», https://rscf.ru/project/22-18-00153/.

Введение

Цифровая гуманитаристика — один из самых актуальных трендов в науках о человеке и обществе в нынешнем столетии. Потенциал цифровых исследований в этой сфере выглядит настолько масштабным, что, возможно, мы стоим на пороге весьма существенных изменений научных дисциплин, традиционно интегрированных в культуру печатной книги. В ближайшей перспективе возможен как рост получаемых результатов (приращение нового знания), так и увеличение «плотности покрытия» предметов гуманитарного изучения (освоение белых пятен, усиление эмпирического качества обоснования гипотез). Но самое интересное происходит в области методологии цифровой гуманитаристики. Собственно, определение дисциплинарного статуса этого направления связано именно с рефлексией специфики его методологических оснований. Пока доминирующей остается позиция, согласно которой цифровая гуманитаристика — это скорее новое междисциплинарное поле, чем самостоятельная дисциплина. Е.В. Самостиенко, например, связывает цифровую гуманитаристику с концепцией «зон обмена» П. Галисона и рассматривает ее как зону обмена, «включающую в себя большое количество автономных зон и создающую нечто наподобие рассеянной когнитивной лаборатории, содержащей не только идеи, но и информационно-технологический базис, набор коммуникационных практик и инструментов, с помощью которых эта сфера оказывается вписана в более широкую социокультурную и технологическую инфраструктуру» [*Самостиенко*, 2018, с. 38]. Вслед за ней Р.И. Мамина и Е.Е. Елькина видят в цифровой

гуманитаристике конвергентные модели и практики глобального сетевого проекта, появление которого связано с «трансформацией объекта и предмета исследования, ослаблением критериев объективности научного познания, методологическим и инструментальным характером междисциплинарности, преимущественно прикладным характером исследований» [*Мамина, Елькина*, 2020, с. 34]. Тем не менее Е.К. Погорский подчеркивает трансдисциплинарный характер цифровых гуманитарных наук [*Погорский*, 2014], а трансдисциплинарность сегодня трактуется как одна из ключевых эпистемических характеристик современного научного знания. Ее задача — связать воедино не только акторов академической науки, но и гражданских экспертов в процессе интерграции различных методологических проектов в рабочий инструмент. Эти интегральные «сплавы» характерны как для фундаментальных областей, так и для локальных стратегий. Таким образом, хотя цифровая гуманитаристика продолжает оставаться методологически рыхлой дисциплиной, в ней вполне могут появляться и закономерно появляются достаточно широкие по применимости методы, способные связать в единое целое если не всю цифровую гуманитаристику, то достаточно крупные ее ареалы. Если до сих пор основным подходом изучения развития цифровой гуманитаристики был подход институциональный, связанный с анализом сложившихся цифроориентированных исследовательских центров [*Mozhaeva*, 2015], их миссий и проектов, то в начавшемся десятилетии у нас появляются основания фиксировать именно методологическую динамику цифровой гуманитаристики. Таким образом, нам необходимо сосредоточиться на поиске, обосновании и рассмотрении «достаточно широких» методов цифровой гуманитаристики. В данной статье мы остановимся на анализе группы математических методов, связанных с информационными технологиями и инструментальными средствами.

Специфика «цифровой» методологии: от атрибуции до калибровки

М. Террас, подводя итоги первого десятилетия существования цифровой гуманитаристики (точкой отсчета автор считает 2004 г.; отметим, что эта позиция не является общепринятой) и обосновывая перспективы следующего десятилетия, предложила использовать две кривые прогноза распространения инноваций для того, чтобы установить, что является собственно цифровой гуманитаристикой с точки зрения методологии [*Terras*, 2016]. Первая кривая — колоколообразная кривая Роджерса, разделенная на четыре сегмента. Первый сегмент — ранние инновации, начинается рост; второй сегмент — раннее меньшинство, распространение инновации достигает пика; третий сегмент — начинается спад, в котором технологии проникают в позднее большинство; четвертый сегмент — кривая приходит к нулю, захватывая позднейших пользователей. Цифровая гуманитаристика, по мнению Террас, предполагает использование цифровых технологий преимущественно на их инновационном этапе. Логика этого утверждения такова — там, где цифровые технологии рутинизированы и используются повсеместно, они адаптированы к общим культурным контекстам. Сегодня все исследователи работают с медиатекстами цифровых библиотек, эта работа подчиняется цифровым принципам функционирования такого рода сервисов, но она слабо затрагивает используемую читателем методологию исследования, которая, несмотря, например, на поиск текстов по

ключевым словам, продолжает оставаться качественной методологией аналитики печатного текста. Максимальная «концентрация» цифры достигается тогда, когда исследователь или проектирует ее сам, или адаптирует под свои задачи еще не получившим массового хождения способом.

Вторая кривая — цикл ажиотажа компании *Gartner*, в которой ажиотаж вокруг новой технологии проходит пять этапов — за технологическим прорывом следует пик завышенных ожиданий, когда люди ждут, что новое решение окажется панацеей от большинства проблем. Завышенные ожидания, подкрепленные опытными данными о границах применимости инновации, обрушиваются во впадину разочарования, где инновация выглядит почти так же малопривлекательно, как и на старте. И только новые усилия по ее совершенствованию могут привести ее по «склону просвещения» на «плато продуктивности», где инновация начинает давать устойчивые результаты, которые, однако, никогда не достигают высоты прогнозов пика завышенных ожиданий. По мнению Террас, писавшей свою статью в 2014 г., цифровая гуманитаристика находится на вершине пика ожиданий, где и политики, и общественность, и само академическое сообщество ждут от нее небывалых свершений. Ей предстоит еще упасть во впадину разочарований и затем медленно и неуклонно начать свое восхождение к плато надежных массовых результатов.

Возможно, это удивительно, но в 2023 г. (почти прошла прогнозируемая Террас новая декада!) ситуация с цифровой гуманитаристикой не изменилась — она по-прежнему на пике завышенных ожиданий. Почему так происходит? Именно потому, что ее место на кривой Роджерса обозначено верно. Здесь формируется маргинальный для традиционной методологии статус исследователей, обеспечивающий им научное лидерство [*Шиповалова*, 2018]. Поскольку сердце «цифры» там, где возможны эксперименты с еще не стандартизированной технологией, мы наблюдаем стабильную ситуацию, когда исследовательская задача ставится в одной технологической перспективе, а продолжается в другой, так и не завершившись в исходной. Пока исследователи осваивают один «сервис», появляется новый, более удобный, или с большим количеством опций, или с упрощенным доступом. Потребность в работе со старым отпадает прежде, чем он получит достаточно широкое распространение в академической среде. Часто цифровую гуманитаристику представляют как чрезвычайно фрагментированную область, в которой в разных сегментах совершенно разные цифровые технологии внедряются в научный метод. Нельзя сказать, что у этого подхода нет оснований. Но нужно не забывать о том, что разные технологически опосредованные методы обеспечивают различное качество результата. Поэтому всегда будут появляться достаточно устойчивые цифровые методы, которые могут давать сильные результаты и иметь значение для достаточно широких групп исследований. Такие методы если и не получат доминирующей позиции в цифровой гуманитаристике, то, по крайней мере, могут задавать в ней тренды. Они и относятся к области нашего интереса.

Но, прежде чем мы перейдем к их рассмотрению, необходимо одно методологическое замечание. Спектр цифровых технологий, применяемых в качестве рабочих инструментов в гуманитарных исследованиях, весьма разнообразен, как и их функциональные роли. Не случайно Э. Арнольд настаивает на том, что необходимо разделять четыре различных значения, в которых вообще сегодня употребляется понятие цифровой гуманитаристики (Digital Humanities): во-первых, это цифровые гуманитарные науки как исследовательская служба; во-вторых, цифровые

гуманитарные науки как метод исследования; в-третьих, исследования цифровых сервисов для гуманитарных наук, в-четвертых, исследования цифровых методов в гуманитарных науках [*Arnold*, 2020]. В первом случае технология является не частью исследовательского процесса, она играет роль вспомогательной инфраструктуры для методологии, которая продолжает оставаться «классической». Во втором случае сам метод устроен таким образом, что представляет собой последовательность компьютерно-опосредованных процедур, использование машинного языка или базы данных составляет его суть. В третьем случае исследуются конкретные технологии, применяемые в научных изысканиях; само такое исследование является разновидностью прикладной информатики. В четвертом случае речь идет об исследованиях самих цифровых методов (Арнольд в качестве примера приводит исследования алгоритмов автоматической или полуавтоматической лемматизации корпусов текстов). Все четыре варианта тесно связаны между собой, но тем не менее очевидно, что перед нами детализация методологии как таковой — цифровизация инфраструктуры метода, цифровизация самого метода, цифровизация его отдельных процедур и этапов. Соответственно, искомый «достаточно широкий» цифровой метод должен относиться ко второму значению Э. Арнольда. При этом его широта будет определена не столько характером технологий, сколько спецификой его объекта. Когда объект достаточно обширен, потенциал связанного с ним метода очевидно возрастает. Одним из самых широких объектов цифровой гуманитаристики на сегодняшней день является медиапространство. В подавляющем большинстве случаев при анализе современного общества, его культурных и антропологических характеристик исследователи фокусируются именно на нем.

Сам концепт медиапространства закрепляется в социогуманитарном дискурсе с середины 1980-х гг. благодаря работам Роберта Сталтса, посвященным довольно экзотическому для того времени опыту удаленной работы. Весьма быстро это понятие стали применять не только к телерешениям для офисов, но и ко всем технически опосредованным формам взаимодействия пространственно рассредоточенных людей. Сегодня о медиапространстве чаще всего говорят как о «вместилище» информационных процессов, хотя в рамках пространственного подхода его могут рассматривать как систему отношений, ядром которой являются традиционные СМИ, перешедшие в цифровую форму, и так называемые медиа [*Елисеева*, 2019]. Иначе говоря, концепт медиапространства позволяет абстрагироваться от технологической архитектуры сети Интернет, сосредотачиваясь на информационных аспектах ее функционирования в том виде, в котором они затрагивают массового пользователя. Также он удобен тогда, когда речь идет о специфических коммуникационных возможностях новых интернет-сервисов в сравнении со старыми, а также о тех способах их применения, которые «изобретают» пользователи при их освоении. Поэтому обращение к медиапространству работает на всех уровнях социогуманитарного анализа — и тогда, когда речь идет о системных глобальных процессах внедрения медиа и тиражирования их содержания, и тогда, когда предметом анализа оказываются локальные пользовательские практики в антропологической перспективе. Медиапространство как методологическая категория достаточно масштабно по своему объему и содержанию, чтобы затрагивать интересы почти каждого гуманитария.

В последние годы набирает популярность метафора калибровки исследовательской оптики, с помощью которой фиксируют ревизию концептуального ядра

классических подходов, процедуры методологического синтеза или адаптацию уже сложившихся методов к нетрадиционным для них предметным областям. Полагаем, что калибровка исследовательской области в случае цифровой гуманитаристики вполне осуществима как уточнение пригодности метода к анализу типовых процессов предметной области. В медиапространстве к таковым можно отнести новостные потоки. С точки зрения структуры они относятся к динамичным, перманентно обновляемым элементам, в противоположность элементам статичным (например, сайтам).

Новостные потоки обеспечивают концентрацию внимания на удаленных от пользователей событиях; с их помощью соотносятся личный опыт и медийная картина мира через изменчивую систему клишированных образов. В соответствии с теорией глобального новостного потока, фокусирующейся на изучении традиционных СМИ, число и характер информационных сообщений о том или ином государстве в национальной медиасфере зависит от географических и экономических факторов, а также интенсивности межстрановых взаимодействий, распределение сообщений по разным типам СМИ зависит от специфики национальной системы массмедиа [*Казун*, 2018, с. 99]. Очевидно, что новые медиа подчиняются этим факторам в «ослабленном» формате; ключевыми в их работе являются алгоритмы генерирования и цензуры новостных лент. Однако, хотя медиапространство включает в себя различные с точки зрения институциональной природы медиа, к анализу их новостного содержания возможен универсальный подход.

Инструментальные средства анализа информационного потока

В последнее время возник устойчивый интерес к инструментальным средствам анализа информационного потока, генерируемого новостными агентствами, социальными сетями, блогерами и т. п. С одной стороны, завершившийся перенос традиционных медиа в интернет-пространство, а также широкое привлечение пользователей к участию в работе социальных сетей создали базу для автоматизированного поиска и анализа информации с помощью специализированных ботов и интернет-плагинов. С другой стороны, активное развитие методов сбора, агрегации и анализа текстовых данных (text mining), а также создание предназначенных для решения этих задач инструментальных средств предоставляет исследователям широкие возможности для изучения медиапространства.

Инструментальные и программные средства анализа медиапространства можно отнести к системам машинного обучения [*Liu*, 2020], так как они в своей работе используют методы искусственного интеллекта и интеллектуального анализа данных, включая алгоритмы обработки естественного языка, семантического и синтаксического анализа, а также методы математической статистики. Такие инструментальные средства позволяют в режиме реального времени определять тенденции роста интереса к той или иной теме, находить наиболее важные события или тренды.

Перейдем к описанию функционирования инструментальных средств и методов анализа медиапространства и новостной аналитики.

В процессе работы обозначенных инструментов происходит количественная оценка различных характеристик текстовых сообщений, среди которых можно выделить следующие:

1) *характер новости* (необходимо установить, является ли упоминание интересующего нас объекта в сообщении позитивным или негативным, или какое влияние — положительное или отрицательное — оказывается новостью на тот или иной анализируемый объект);
2) *влияние новостного сообщения* (характеризует силу влияния новости на масштаб вызванных ею изменений);
3) *релевантность* (показывает, насколько событие, описанное в новостном сообщении, относится к интересующему исследователя объекту);
4) *новизна* (показывает, насколько новым и информативным является данное сообщение).

Инструментальные средства анализа медиапространства нацелены на более глубокое понимание явлений и процессов, происходящих в нем, на основе количественной оценки вышеуказанных характеристик новостных сообщений.

Очевидно, что в современном мире уровень интенсивности генерации сообщений различными новостными агентствами или пользователями социальных сетей столь высок, что человек не в состоянии своими силами обработать этот информационный поток. События, потенциально способные изменить ситуацию, могут быть потеряны или просмотрены в огромном потоке сообщений. Именно поэтому автоматизированные средства анализа могут служить эффективным помощником при исследовании новостного потока.

Оценка характеристик новостей в виде количественных значений дает возможность их применения в математических моделях медиапотока. Зачастую процесс анализа новостей носит автоматический характер и включает следующие шаги:

1) отбор новостных сообщений в режиме реального времени из различных ресурсов;
2) предварительный анализ текстовых сообщений;
3) оценку ожиданий, вызванных публикацией новости, на основе текущей ситуации;
4) создание и применение количественных моделей.

Охарактеризуем эти шаги более подробно.

Новости могут быть извлечены из потока, генерируемого различными источниками:

1. *Новостные ресурсы информационных агентств.* Традиционные медиа до сравнительно недавнего времени распространяли свои сообщения, используя печатные ресурсы, радио, телевидение. Заметим, что это затрудняло получение общей картины новостного потока. Однако современные медиа и информационные агентства перенесли публикацию своих новостей и сообщений в сеть Интернет, что повлияло на процесс отслеживания, сбора и анализа единиц новостного потока. Более того, наличие тегов и индексирования новостей сделало возможной их автоматизированную обработку в режиме реального времени.

2. *Предварительные новости или материалы из первичных источников.* Предварительные новости представляют собой сырой, необработанный материал, который журналисты и медиа могут применять при написании текста новостного сообщения. Такой материал часто получается на основе анализа первичных источников, например, текстов принимаемых указов, законов, отчетов ЦБ или Счетной палаты, судебных документов, отчетов различных правительственных агентств, корпоративных ресурсов, компаний, анонсов, индустриальной и макроэкономической статистики.

3. *Социальные сети и информационные площадки (блоги, форумы, социальные сети, видеохосты и т. п.)*. Очевидно, что качество размещенных в социальных сетях сообщений должно тщательно проверяться из-за большого количества фейковых новостей, и существенный массив таких сообщений не является достоверным. Однако они дают исследователю возможность изучать и анализировать общее настроение совокупности однотипных сообщений, строить характеристики настроений социума на основе анализа лайков или дизлайков, оценок интереса к той или иной теме сообщений и т. п., а также использовать результаты моделирования для анализа трендов.

Сентимент-анализ (анализ настроений, или интеллектуальный анализ мнений) определяется как задача поиска мнений авторов о конкретных объектах. Общая архитектура общей системы анализа настроений выглядит следующим образом. Прежде чем приступить к классификации текстов и определению их настроения, необходимо провести их предобработку. Под текстом понимается одна строка из алфавитных и неалфавитных символов. Заметим, что обрабатывать его в таком виде неудобно, поэтому на начальном этапе необходимо выделить числовые признаки, для чего данные приводятся к удобному виду и нормализуются. Для заданной коллекции текстовых документов предобработка проводится следующим образом: осуществляется токенизация, затем все слова приводятся к нижнему регистру, далее удаляются стоп-слова и знаки пунктуации, происходит фильтрация слов по частоте/длине/регулярному выражению, и наконец осуществляется процесс лемматизации (стемминг).

Новостная аналитика измеряет релевантность, характер, новизну и весомость новости. Обработанные новости превращаются из текстовой информации в статистические данные, на основе которых анализируются взаимоотношения различных новостей (их взаимная корреляция). Потому время выхода новости при осуществлении такого анализа является важным.

Методы сентимент-анализа сообщений

Одна из важных задач, которые подлежат решению при разработке системы автоматизированного анализа новостного потока, состоит в получении количественной оценки сообщений. Другими словами, необходимо на основании содержания новостного сообщения (текста новости) найти некоторую количественную оценку, которая отражает контекст упоминания объекта в новости (положительное, нейтральное или негативное), что соответствует ожиданиям относительно объекта, которые связаны с этой новостью или вызваны ее появлением [*Liu*, 2020].

Самая простая формулировка этой проблемы состоит в оценке смысла новости на основе бинарной классификации (положительная/негативная). Более сложные формулировки могут включать использование шкалы положительности/негативности, что дает более дифференцируемую оценку ожиданий, связанных с сообщением.

Чтобы осуществить эту классификацию, необходимо проанализировать контекст и текст новости, ее эмоциональное содержание, т. е. интерпретацию новости читателями в терминах положительного или негативного восприятия. Количественная оценка эмоционального наполнения сообщения может проводиться с

использованием экспертов и/или психологических словарей. Однако это не исключает конфликта интерпретаций между различными экспертами.

Для оценки ожиданий, связанных с новостью, можно использовать следующие два достаточно простых подхода. Первый из них вычисляет некоторый индекс, базирующийся на поведении пользователей социальных сетей [*Lavrenko et al.*, 2000], а именно вычисляется индекс DISAG расхождения в оценке сообщения

$$DISAG = \left| 1 - \left| \frac{B - S}{B + S} \right| \right|$$

где B — число лайков или положительных комментариев, а S — число дизлайков или негативных комментариев к сообщению социальной сети. Тогда значение DISAG = 0 означает, что среди пользователей нет разногласия, в то время как DISAG = 1 означает полное разногласие.

Второй подход к определению ожиданий относительно этой новости состоит в использовании методов машинного обучения и методов обработки естественного языка, в результате чего рассчитывается индекс ожиданий.

В частности, для классификации текстов используются многие модели машинного обучения, среди которых: наивный байесовский классификатор, логистическая регрессия, композиции деревьев решений, полносвязные нейросети, сверточные нейросети, а также рекуррентные нейросети.

Для оценки индекса ожиданий, связанных с новостным сообщением, применяются следующие методы [*Liu*, 2020].

1. *Примитивный классификатор,* самый простой из алгоритмов, считает количество положительных и отрицательных коннотаций слов. Если разность между этими величинами больше (или меньше) некоторого порогового значения, то новость считается положительной (или негативной), в иных случаях — нейтральной.

2. *Метод классификации сообщений на основе вычисления векторного расстояния* основан на том, что используется многомерное пространство, осями которого являются слова лексикона. Задается также множество обучающих примеров (сообщений), для которых заранее известна их принадлежность к одному из классов (позитивное, негативное или нейтральное). В процессе своей работы алгоритм находит скалярное произведение между новым сообщением и всеми сообщениями из множества обучающих примеров и определяет новое сообщение в тот класс, к которому это сообщение наиболее близко в смысле векторного расстояния.

3. *Дискриминантный классификатор* основан на предварительно установленных весовых значениях для каждого слова лексикона. Далее классификатор считает значение настроения сообщения как сумму весов слов лексикона, входящих в это сообщение (взвешенный подсчет). Для оценки веса для слова лексикона применяется обучающее множество сообщений, для каждого из которых известно, к какому из классов (нейтральное, отрицательное, положительное) оно принадлежит.

4. *Классификатор фраз «прилагательное — наречие»* основан на том факте, что фразы, использующие прилагательные и наречия, акцентируют ожидания и имеют больший вес. Этот метод считает слова из лексикона в новостном сообщении, но учитывает только те из них, которые идут совместно с прилагательными или наречиями.

5. Другим распространенным методом является *байесовский классификатор,* в котором новостное сообщение приписывается тому классу, вероятность принадлежать которому больше.

В настоящее время во многих языках программирования разработаны библиотеки подпрограмм, предназначенные для выполнения задач, связанных с анализом текстов (text mining). В рамках данной статьи невозможно рассказать обо всех из них, поэтому упомянем лишь библиотеки *tm* (язык *R*) и *Spacy* (язык *Python*):

— Библиотека *tm* для работы на языке программирования *R* обладает рядом функций, предназначенных для проведения интеллектуального анализа текстов. Эта библиотека содержит методы импорта данных, обработки корпуса текстов, предварительной обработки текстовых фрагментов, управления метаданными и создания матриц терминов и документов [*Feinerer et al.*, 2008].

— *spaCy* — это бесплатная библиотека с открытым исходным кодом, написанным на языке *Python*, с множеством встроенных возможностей для обработки естественного языка (Natural Language Processing — NLP). Эта библиотека является одним из самых популярных инструментов, предназначенных для решения задач обработки и анализа текстовых данных. Пакет подпрограмм содержит функции обработки естественного языка, включая инструменты предварительной обработки и очистки текстовых данных. Кроме того, он содержит средства токенизации, удаления стоп-слов, нормализации слов, векторизации текста [*Neumann et al.*, 2019]. Важной частью функциональных возможностей является использование классификаторов машинного обучения для прогнозирования настроений.

Еще одной важной тенденцией последнего времени является появление агрегаторов новостного потока, создание полномасштабных баз новостей и публикаций за большой промежуток времени. Такие агрегаторы являются результатом работы коллективов крупных медиаресурсов и медиакомпаний и могут предоставлять ограниченно бесплатный доступ для исследователей.

Одним из таких амбициозных проектов является *GDELT* (Глобальная база данных событий, языка и тональности), созданный Джорджтаунским университетом на основе использования разработок *Google* [*Leetaru, Schrodt*, 2013]. *GDELT* отслеживает и собирает в режиме реального времени все онлайн-новости, автоматически переводит их на английский язык и кодирует каждую новостную статью по теме, настроению и тональности, местоположению и объектам (организациям и лицам). Этот процесс использует алгоритмы анализа текстов (text mining), поддерживаемые *Google Cloud*; при этом генерируется более 1 триллиона различных записей данных в год.

Заключение

Таким образом, в настоящее время исследователи, специализирующиеся на изучении проблем цифровой гуманитаристики, имеют в своем арсенале достаточно широкий набор математических методов и инструментальных программных средств, которые могут служить мощным средством для проведения более глубокого анализа процессов, происходящих в медиапространстве, на основе оценки

количественных характеристик соответствующего новостного потока или взаимодействия пользователей социальных сетей. Эти инструменты очевидно применимы для традиционных задач теории журналистики *media studies*, прямо ориентированных на новостной цифровой контент. Вместе с тем они пригодны для решения любых задач, связанных с исследованием содержания «общественного сознания», «общественного мнения» или «дискурсов», т. е. концептов, методологически развиваемых философскими, социологическими, лингвистическими и культурологическими дисциплинами. С помощью анализа новостных потоков могут изучаться коллективные представления о прошлом и истории, релевантные исследовательским задачам в русле *memory studies*. Значим этот инструментарий и при анализе политических идентичностей. Кроме того, рассмотренные инструментальные средства пригодны для решения задач социологии науки, поскольку сегодня сама наука существует как сеть, объективирующаяся в цифровых форматах, а основные социальные сети, специализированные под научно-исследовательские задачи (*Academia.edu* и *ResearchGate*) оснащены инструментами генерации новостных потоков. Рассмотренные нами инструментальные математические методы обладают широким эвристическим потенциалом, позволяющим применять их к самым различным аспектам функционирования медиапространства, консолидируя тем самым исследовательские программы цифровой гуманитаристики.

Литература

Доу Ш. Психология финансовых рынков: Кейнс, Мински и поведенческие финансы / Пер. с англ. А. Маловой // Вопросы экономики. 2010. № 1. С. 99–113.

Елисеева М.А. Медиапространство: социально-философский анализ // Известия Саратовского университета. Новая серия. Сер.: «Философия. Психология. Педагогика». 2019. Т. 19. № 1. С. 4–7. DOI: 10.18500/1819-7671-2019-19-1-4-7.

Казун А.Д. Глобальный новостной поток (О каких странах говорят российские СМИ и почему?) // Полития: Анализ. Хроника. Прогноз (Журнал политической философии и социологии политики). 2018. Т. 91. № 4. С. 90–105.

Мамина Р.И., Елькина Е.Е. Digital Humanities: новая наука или конвергентные модели и практики глобального сетевого проекта? // Дискурс. 2020. Т. 6. № 4. С. 22–38. DOI: 10.32603/2412-8562-2020-6-4-22-38.

Можаева Г.В. Digital Humanities: цифровой поворот в гуманитарных науках // Гуманитарная информатика. 2015. № 9. С. 8–23. DOI: 10.17223/23046082/9/1.

Погорский Э.К. Особенности цифровых гуманитарных наук [Электронный ресурс]. Режим доступа: http://www.zpu-journal.ru/e-zpu/2014/5/Pogorskiy_Digital-Humanities/ (дата обращения: 10.07.2023).

Самостиенко Е.В. Digital Humanities в русскоязычном контексте: траектория институционализации и механизмы формирования автономных зон // Вестник Вятского государственного университета. 2018. № 4. С. 37–45.

Шиповалова Л.В. Маргинальность и лидерство в науке // Социология науки и технологий. 2018. Т. 9. № 4. С. 39–51. DOI 10.24411/2079-0910-2018-10019.

Arnold E. Digital Humanities: Is it Research or is it Service? // Digital Humanities München. 2020. 26 Juli.

Liu B. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge. Cambridge University Press, 2020. 448 p.

Feinerer H., Hornik K., Meyer D. Text Mining Infrastructure in R // Journal of Statistical Software. 2008. Vol. 25. No. 5. P. 1–54.

Lavrenko V., Schmill M., Lawrie D., Ogilvie P., Jensen D., Allan J. Language Models for Financial News Recommendation // Conference on Information and Knowledge Management. Proceedings of the Ninth International Conference on Information and Knowledge Management. McLean, Virginia, United States, 2000. P. 389–396.

Leetar K., Schrodt P.A. Gdelt: Global Data on Events, Location and Tone, 1979–2012 // Technical Report. KOF Working Papers, 2013.

Neumann M., King D., Beltagy I., Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing // Proceedings of the 18th BioNLP Workshop and Shared Task. Florence: Association for Computational Linguistics, 2019. P. 319–327.

Terras M. A Decade in Digital Humanities // Journal of Siberian Federal University. Humanities & Social Sciences. 2016. Vol. 9. No. 7. P. 1637–1650. DOI: 10.17516/1997-1370-2016-9-7-1637-1650.

Instrumental Methods of Media Space Analysis in Digital Humanities

Sergei P. Sidorov

Saratov State University,
Saratov, Russia;
e-mail: sidorovsp@yahoo.com
ORCID: 0000-0003-4047-8239

Sophia V. Tikhonova

Saratov State University,
Saratov Russia;
e-mail: segedasv@yandex.ru
ORCID: 0000-0003-2487-3925

The article deals with the analysis of the place and the role of instrumental mathematical methods in the methodology of the digital humanities. The authors explore the potential of those methods in terms of overcoming the fragmentation of the digital methodology due to the breadth of applicability of the method. The "sufficient" breadth of instrumental methods, which can strengthen the positive convergence of digital methodology, is ensured by their applicability to various aspects of news flows of the media space. A news stream is a collection of news generated by a) news agencies, b) preliminary materials from primary sources, c) social networks. The authors consider analytical tools as a kind of machine learning systems used to determine trends in the media space. Their applicability is focused on the quantitative assessment of text messages (nature, influence, relevance, novelty), as well as the forms of sentiment analysis of varying degrees of complexity, allowing to reflect the context of a news message, positive, negative or neutral. To do this, either the calculation of the DISAG index of the discrepancy in the evaluation of the message is used, or such machine learning models as the naive Bayesian classifier, the logistic regression, the compositions of decision trees, the fully connected neural networks, the convolutional neural networks, the recurrent neural networks. The authors pay special attention to auxiliary databases — dictionaries, lexicon and grammar, as well as libraries of subroutines designed to perform tasks related to text analysis, and aggregators of the news stream. The authors come to the conclusion that a deep analysis of the quantitative characteristics of certain news

streams or interactions of the users of social networks allows solving typical tasks in the main areas of digital humanities, thereby contributing to the unification of its methodology.

Keywords: digital humanities, digital methodology, media space, news flow, instrumental methods of analysis.

Acknowledgment

The research was carried out with support from the Russian Science Foundation according to the research grant No. 22-18-00153 “The image of the USSR in historical memory: a study of media strategies for reproducing ideas about the past in Russia and foreign countries”, https://rscf.ru/project/22-18-00153 /.

References

Arnold, E. (2020). Digital Humanities: Is it Research or is it Service? *Digital Humanities München*, 26 Juli.

Dow, Sh. (2010). Psikhologiya finansovykh rynkov: Keyns, Minski i povedencheskiye finansy [Psychology of financial markets: Keynes, Minsky and behavioral finance], *Voprosy ekonomiki*, no. 1, 99–113 (in Russian).

Eliseeva, M.A. (2019). Media Space: Socio-Philosophical Analysis, *Izvestiya Saratovskogo universiteta. Novaya seriya. Ser.: Filosofia. Psykhologiya. Pedagogika*, *19* (1), 4–7 (in Russian). DOI: 10.18500/1819-7671-2019-19-1-4-7.

Feinerer, H., Hornik, K., Meyer, D. (2008). Text Mining Infrastructure in R, *Journal of Statistical Software, 25* (5), 1–54.

Kazun, A. (2018). Global’nyy novostnoy potok (O kakikh stranakh govoryat rossiyskiye SMI i pochemu?) [Global news flow (What countries Russian media talk about and why)], *The Journal of Political Theory, Political Philosophy and Sociology of Politics Politeia*, *91* (4), 90–105 (in Russian). DOI: 10.30570/2078-5089-2018-91-4-90-105.

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., Allan, J. (2000). Language Models for Financial News Recommendation, in *Conference on Information and Knowledge Management. Proceedings of the Ninth International Conference on Information and Knowledge Management* (pp. 389–396), McLean, Virginia, United States.

Leetaru, K., Schrodt, P.A. (2013). Gdelt: Global Data on Events, Location and Tone, 1979–2012, *Technical Report, KOF Working Papers*.

Liu, B. (2020). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, Cambridge, UK: Cambridge University Press.

Mamina, R.I., Yelkina, E.E. (2020). Digital Humanities: Is it a New Science or a Set of Models and Practices of the Global Network Project?, *Discourse*, *6* (4), 22–38 (in Russian). DOI: 10.32603/2412-8562-2020-6-4-22-38.

Mozhaeva, G.V. (2015). Digital Humanities: Tsifrovoy povorot v gumanitarnykh naukakh [Digital Humanities: Digital turn in the humanities], *Gumanitarnaya Informatika*, no. 9 , 8–23 (in Russian). DOI: 10.17223/23046082/9/1.

Neumann, M., King, D., Beltagy, I., Ammar, W. (2019). ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing, *Proceedings of the 18th BioNLP Workshop and Shared Task* (pp. 319–327), Florence: Association for Computational Linguistics.

Pogorskiy, E.K. (2014). Osobennosti tsifrovykh gumanitarnykh nauk [Features of digital humanities]. Available at: http://www.zpu-journal.ru/e-zpu/2014/5/Pogorskiy_Digital-Humanities/ (date accessed: 10.07.2023) (in Russian).

Samostienko, E.V. (2018). Digital Humanities v russkoyazychnom kontekste: trayektoriya institutsionalizatsii i mekhanizmy formirovaniya avtonomnykh zon [Digital Humanities in the Russian-speaking context: the trajectory of institutionalization mechanisms of formation of autonomous zones], *Vestnik Vyatkinskogo gosudarstvennogo universiteta*, no. 4, 37–45 (in Russian).

Shipovalova, L.V. (2018). Marginal'nost' i liderstvo v nauke [Marginality and leadership in science], *Sotsiologiya nauki i tekhnologiy*, *9* (4), 39–51 (in Russian). DOI: 10.24411/2079-0910-2018-10019.

Terras, M. (2016). A Decade in Digital Humanities, *Journal of Siberian Federal University. Humanities & Social Sciences*, *9* (7), 1637–1650. DOI: 10.17516/1997-1370-2016-9-7-1637-1650.