ОНЛАЙН-ИССЛЕДОВАНИЯ

А.А. Бызов
(*Москва*)

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ТЕКСТОВ В СОЦИАЛЬНЫХ НАУКАХ[1]

На протяжении практически всей истории социологии социологи стремились изучать неструктурированные органические тексты: материалы газет, дневники, мемуары, письма, документы, а с недавнего времени и сообщения, публикации и другие тексты на различных онлайн-платформах. В этой статье обсуждается то, как современные техники интеллектуального анализа текста (ИАТ) могут улучшить классические социологические подходы к анализу такого типа данных. Статья построена по следующему плану. Сначала обсуждаются примеры классического количественного контент-анализа и его ограничения, которые решаются с помощью ИАТ. Затем обсуждается, как ИАТ применяется в современных исследованиях социальных наук. На примере исследования с применением структурного тематического моделирования показывается, как ИАТ может применяться в исследованиях аннотаций научных статей для выявления встречающихся в этих статьях тем, их распространенности в разные годы и связей между этими темами. На другом примере исследования, в котором классифицировались сообщения в социальной сети *Twitter*, показывается, как такой тип нереактивных текстовых данных сопоставляется с результатами интернет-опросов и телефонных опросов. Наконец, в заключении статьи обсуждаются некоторые современные подходы к анализу текстов с применением глубинного обучения.

Александр Александрович Бызов – аналитик Института образования, аспирант школы социологических наук, Национальный исследовательсий университет «Высшая школа экономики». E-mail: debesergopotes12@gmail.com.

Публикация подготовлена в результате проведения работы (№19-04-055) в рамках Программы «Научный фонд Национального исследовательского университета “Высшая школа экономики” (НИУ ВШЭ)» в 2019 г. и в рамках государственной поддержки ведущих университетов Российской Федерации «Проект 5–100».

Ключевые слова: интеллектуальный анализ текстов, тематическое моделирование, классификация текстов, методология социологического исследования, методы социологического исследования.

На протяжении практически всей истории социологии социологи стремились изучать неструктурированные органические тексты[1]: материалы газет, дневники, мемуары, письма, документы[2], а с недавнего времени и сообщения, публикации и другие тексты на различных онлайн-платформах[3]. Этому интересу сейчас может способствовать беспрецедентная скорость производства текстовых данных и возможность получения доступа к их сбору (см., напр.: [7]). Например, по данным *Internet live stats*, за одну секунду отправляются 8713 постов в *Twitter*, 1607 постов в *Tumblr* и 2 850 886 электронных писем [8].

[1] Прилагательное «*неструктурированные*», часто используемое в описании текстов, не означает, что у них нет структуры. Очевидно, что тексты структурируются в соответствии с правилами языка. Скорее оно означает, что тексты не структурированы для использования в анализе данных без какой-либо обработки. См. подробнее: [1; 2]. Под *органическими,* или *нереактивными,* понимаются такие тексты, сбор которых не требовал «активного участия исследуемых субъектов в процессе исследования, т.е. не предполагающие эксплицитного осознания субъектами их роли или самого факта участия в исследовании, а также целей и возможных последствий последнего» [3, с. 25].

[2] Известен пример М. Вебера, который в 1910 г. на собрании Германского социологического общества предложил исследовательский проект по изучению газет, в рамках которого он описывал, в том числе, количественное и качественное исследования динамики материалов в газетах [4]. Подробнее ознакомиться с ранними попытками исследований органических текстов в социальных науках можно в книге К. Криппендорфа [5].

[3] Социологи Дж. Эванс и П. Асевес (Aceves) «широкими мазками» так описывают современные источники этих данных: «...текстовые следы разнятся от поведения на сайтах, социальных медиа, отправленных мгновенных сообщениях, интернет-торговли и до автоматически затранскрибированных YouTube видео, медицинских карточек, оцифрованных библиотек и муниципальных сервисов» [6, p. 22].

У работы с такого типа органическими текстами есть множество преимуществ и недостатков. Приведу лишь некоторые из них, начиная с преимуществ. Во-первых, органический, или нереактивный, характер этих данных позволяет избежать систематических ошибок измерения, связанных с контактом исследователя с изучаемыми субъектами (см., напр.: [3]). Во-вторых, цифровой характер хранения многих из текстов, а также открытость их хранения на некоторых онлайн-платформах[1] позволяют конструировать выборки с характеристиками больших данных[2]. Наконец, органические тексты, как и другие онлайн-данные, позволяют ответить на некоторые вопросы социальных наук, ответы на которые не были возможны без данных такого типа (см., напр.: [6; 12]).

Что касается недостатков, то, во-первых, органические тексты создаются не для целей исследования, поэтому их часто может не хватать для ответа на исследовательский вопрос. С этой проблемой также может быть связано и то, что поведение людей на разных онлайн-платформах неестественно и ограничено тем, какие возможности закладывали инженеры в ту или иную платформу [11]. Во-вторых, органические тексты в Интернете могут производиться не людьми, которые честно выражают свое мнение по тем или иным вопросам, а ботами, организациями или нанятыми людьми для выражения той или иной позиции [13]. В-третьих, онлайн-платформы, на которых люди оставляют цифровые следы в форме органических текстов, могут быть подвержены трем «сдвигам» (*drifts*): 1) могут измениться люди, которые пользуются этими платформами; 2) могут измениться способы, которыми

[1] Очевидно то, что далеко не все платформы предоставляют доступ к своим данным. Например, ко многим данным *Facebook* у исследователей практически нет доступа (см.: [9]).

[2] Например, исчерпывающие по объему выборки (n = все случаи) или высоко гранулированные выборки (можно посмотреть, о чем были все газетные статьи в определенный день) [10; 11].

люди пользуются этими платформами и 3) могут измениться сами платформы [11].

На сегодняшний в социальных науках устоялось несколько подходов к анализу органических текстов – это контент-анализ (количественный и качественный), аргументативный анализ, риторика, качественный анализ идей и идеологий, нарративный анализ, анализ метафор, критическая лингвистика и различные версии дискурс-анализа [14][1]. Во всех этих подходах в том или ином виде присутствует хотя бы одна из двух следующих стратегий восприятий текста.

Скорее *холистическое* восприятие, то есть текст интерпретируется или оценивается напрямую без какой-либо формы систематического извлечения элементов. Этот подход характеризуется интересом скорее к тому, что текст означает или что текст говорит об авторе, его попытке повлиять на аудиторию или о социальной структуре. К. Бенуа называет этот подход «текст-как-текст» и относит сюда литературный анализ, дискурс-анализ и т.д. [2][2].

Скорее *атомарное* восприятие, то есть текст сначала подвергается какой-то форме систематического извлечения элементов (коды, темы, термины, слова и т.д.). Этот подход характеризуется интересом скорее к тому, что в тексте сказано, какая в нем содержится информация. К. Бенуа называет этот подход «текст-как-данные» и относит сюда контент-анализы, статистические сводки и т.д. [2].

---

[1] Эта классификация – далеко не единственная среди тех, которые предлагаются для описания подходов к текстовому анализу в социальных науках. Например, Г. Игнатоу и Р. Михалча (Mihalcea) предлагают следующую классификацию: конверсационный анализ, анализ дискурсивных позиций, критический дискурс-анализ, контент-анализ, фукодианский анализ, анализ текста как социальной информации [15]. Другую классификацию предлагает К. Бенуа: дискурс-анализ, тематический анализ, контент-анализ, статистическая сводка, машинное обучение и дистрибутивная семантика [2]. А. Бриман предлагает чуть более краткую версию методов: количественный контент-анализ, качественный контент-анализ, семиотика, герменевтика и дискурс-анализ [16].

[2] Также см. о таком восприятии: [17].

В этой статье я бы хотел обсудить то, как современные техники интеллектуального анализа текста (ИАТ)[1] улучшают классические социологические подходы[2] к анализу текстов, в которых используется *атомарное* восприятие[3]. Статья построена по следующему плану. Сначала я описываю пример классического дизайна исследования, в котором применяется атомарное восприятие – количественный контент-анализ – и описываю несколько его ограничений, которые в какой-то степени решаются с помощью интеллектуального анализа текста. Затем я рассматриваю, что такое ИАТ, какие в нем есть техники и как они применяются в социальных науках.

Количественный контент-анализ в применении к текстам и интеллектуальный анализ текстов

Количественный контент-анализ – один из старейших исследовательских методов в арсенале социальных ученых [5, p. 10–24]. Есть множество определений этому методу. Б. Берельсон формулирует его следующим образом: «контент-анализ – это исследовательская техника, направленная на объективное, систематическое и количест-

[1] Интеллектуальный анализ текста – это зонтичное понятие, описывающее разные компьютерные инструменты и статистические техники квантификации текста [1, p. 1]. Подробнее об ИАТ см. дальше в тексте.

[2] В социальных науках использование методов интеллектуального анализа текстов обсуждается в рамках так называемой вычислительной социальной науки (*computational social science*) [18; 19; 20].

[3] Сосредоточение на подходе «текст-как-данные» не означает, что интеллектуальная обработка текстовой информации может быть полезной только в этом подходе. Например, есть несколько исследований, которые показывают пользу ИАТ в исследованиях с дискурс-анализом. См.: [21; 22; 23]. Помимо этого, развивается метод вычислительной обоснованной теории, в рамках которой в какой-то степени применяются оба подхода [17].

венное описание содержания коммуникации» [24, p. 18]. К. Криппендорф определяет контент-анализ как «исследовательскую технику, направленную на вынесение воспроизводимых и надежных выводов из текстов… о контекстах их использования» [5, p. 18]. К. Нойндорф определяет количественный контент-анализ как процедуру «получения сводного обзора сообщений, которая следует стандарту научного метода (включая внимание к объективности–интерсубъективности, априорности дизайна, надежности, валидности, распространяемости на генеральную совокупность, воспроизводимости и тестируемости гипотез на основе теорий) и не ограниченного типами измеряемых переменных или контекстом сообщений» [25, p. 39].

Процедура количественного контент-анализа включает в себя несколько шагов. Сначала исследователь должен поставить вопрос, обосновать и найти теоретическую рамку. Затем он проводит процедуры концептуализации и операционализации. После этого ему следует выстроить схему кодирования для кодировщиков или найти/составить словарь для компьютерного кодирования. На следующем этапе исследователь должен принять решение, использовать ли генеральную совокупность текстов или же построить какую-то форму репрезентативной выборки. Если он работает с кодировщиками, то тогда необходимо провести их тренировку и пилотажное исследование надежности кодирования между несколькими кодировщиками. После этого наступает процедура самого кодирования с помощью кодировщиков (которых должно быть не менее двух и которые должны оценивать не менее 10% общих текстов для того, чтобы оценить надежность кодировки между несколькими кодировщиками). Если исследователь работал с кодировщиками, то после того, как они закодировали данные, необходимо провести оценку финальной надежности. Наконец, закодированные данные приводятся к табличному виду, анализируются и представляются результаты этого анализа [25, p. 69].

Описание процедуры проведения количественного контент-анализа в его применении к текстовым данным может показать,

какие ограничения есть у классических социологических атомарных анализов текста с использованием кодировщиков. Во-первых, извлечение информации людьми – невероятно трудозатратно и требует длительного периода подготовки кодировщиков [18, p. 103]; более того, кодировщики не всегда могут воспроизвести даже собственное кодирование. Во-вторых, люди часто делают ошибки и могут вносить свои интерпретации, что приводит, например, к низкой надежности кодировки [18, p. 103; 26]. В-третьих, текстов становится так много, что ни одному человеку не под силу их прочитать или закодировать [18, p. 103], что заставляет, например, использовать вместо генеральной совокупности только выборки, что, в свою очередь, приводит к большому снижению гранулированности результатов анализа. Наконец, люди не приспособлены к тому, чтобы находить сложные паттерны, например, сетевую структуру, латентные свойства, характеристики, связанные со временем [18, p. 103].

Многие из этих проблем могут решаться с помощью автоматического контент-анализа, при котором составляется список паттернов (слова, словосочетания и т.д.), которые ищутся в текстовых данных[1]. Результаты такого типа анализа воспроизводимы как одним и тем же компьютером снова и снова, так и несколькими компьютерами, использующими один и тот же словарь. Более того, эти словари можно применять к большому количеству текстов и в них можно заложить простые формы анализа сетевой структуры, латентных свойств или характеристик времени. Основные ограничения словарного метода связаны с его жесткой привязкой к

[1] Например, если перед исследователем стоит задача по оценке эмоциональной тональности (положительная или отрицательная) текстовой информации различных новостных источников на русском языке, он может воспользоваться словарем оценочных слов и выражений русского языка «РуСентиЛекс» [27], в котором, например, слово «авторитарный» закодировано как «негативное», а слово «авторитет» как «позитивное». Так исследователь может посчитать среднюю тональность текстов каждой новости того или иного источника и сравнить их.

тем паттернам, которые задает исследователь. Поэтому, например, применение автоматического контент-анализа к текстовой информации может приводить к тому, что слова будут вырываться из контекста[1] [28; 29, p. 169–171].

Помимо словарных методов, которые широко используются учеными из социальных наук, есть и другие компьютерные методы анализа текста, способные обеспечить воспроизводимость результатов анализа данных на большом объеме текстов с возможностью находить сетевую структуру или латентные характеристики, но при этом не так привязанные к жестким паттернам, обозначенным исследователем заранее. Эти методы часто называют интеллектуальным анализом текста. Во многом развитие этих методов связано с достижениями в таких областях, как извлечение информации (*information retrieval*), интеллектуальный анализ данных (*data mining*) и компьютерная лингвистика [30; 31; 32]. Эти методы включают в себя семь отдельных, хоть и сильно связанных между собой практических областей: 1) поиск и воспроизведение информации, 2) кластеризация документов, 3) классификация документов, 4) интеллектуальный анализ данных из интернет-страниц[2], 5) извлечение информации, 6) обработка

[1] Приведу два примера из работы Р. Стайна об исследованиях эмоциональной окраски текстов. Во-первых, словарные методы могут не учесть сарказм: «пришлось снова и снова жать на эту классную кнопку, чтобы получить чашку кофе. Уйма веселья с утра. Не самый лучший способ начать день». Для словарного метода эти три слова «классная», «веселья», «лучший» будут свидетельствовать о положительной окраски текста. Другой пример связан с сообщениями, в которых положительно описывается одна часть явления и отрицательно – другая (например, хорошие актеры, плохой сюжет), что для словарного алгоритма будет означать нейтральную окраску сообщения [28].

[2] Может показаться неясным, почему интеллектуальный анализ данных из интернет-страниц выделяется отдельно. Авторы этой классификации считают, что данные, получаемые из Интернета, обладают некоторыми особенностями (полуструктурированность, наличие ссылок между страницами на сайтах), которые диктуют необходимость выделять эту область отдельно [32].

естественного языка и (7) извлечение концептов [32, p. 32]. Рассмотрим далее, как исследователи из социальных наук применяют интеллектуальный анализ текстов.

Применение интеллектуального анализа текста в социальных науках

В социальных науках методы ИАТ активно применяются в исследованиях социальных движений [33; 34; 35], культуры [7; 36], различные аспекты политического поведения и речи [37; 38; 39; 40; 41; 42; 43]. На данный момент существует множество описаний того, какие задачи исследователей из социальных наук помогают решить методы ИАТ. Например, Дж. Гриммер и Б. Стюарт считают, что эти методы могут решать две широкие категории задач: классификация и шкалирование (расположение текста на каком-либо измерении – например, насколько в тексте выражена «левая» или «правая» идеология) [44]. М. Шунвельде (Schoonvelde) с коллегами выделяют пять задач, которые обсуждаются в контексте их применения к задачам социальных наук: тематическое моделирование, шкалирование, анализ эмоциональной окраски текстов, анализ сложности текста и анализ личностных черт автора текста (см. подр.: [45]). Дж. Уилкерсон и А. Казас пишут о классификации, шкалировании, анализе вторичного использования текстов и применении таких методов обработки естественного языка, как распознание частей речи и распознание сущностей (США, Америка, Соединенные Штаты) [46]. В этой статье я чуть подробнее остановлюсь на классификации текстов в двух ее вариантах: нахождение тем в текстах (тематическое моделирование) и разметка текстов на основе существующей классификации. Такой выбор обусловлен в первую очередь популярностью этой задачи в социальных науках в целом и в социологии в частности. Например, Гриммер и Стюарт пишут, что «присваивание текстам категорий – это наиболее распространенный метод анализа контента в полити-

ческой науке» [44, p. 273]. Про распространенность задачи классификации текстов пишут и в исследованиях коммуникаций [47], а также в социологии [48]. Другие задачи, такие как шкалирование, анализ вторичного использования текстов или анализ личностных черт автора текста, характерны скорее для отдельных задач той или иной социальной науки (например, политологии или психологии)[1]. Прежде чем перейти к обсуждению задачи на классификацию текстов, хотелось бы прояснить два вопроса: 1) каковы базовые принципы проведения исследования с применением ИАТ и 2) какие предварительные шаги рекомендуется осуществлять перед непосредственно применением того или иного из рассматриваемых методов ИАТ.

Гриммер и Стюарт в своей знаменитой статье «Текст как данные: возможности и ограничения методов автоматического контент-анализа в анализе политических текстов» выделяют четыре следующих принципа при проведении количественного текстового анализа.

1. Все количественные модели языка неправильны, но некоторые из них – полезны.

2. Количественные методы анализа текстов дополняют людей, а не заменяют их.

3. Нет одного лучшего метода для автоматического текстового анализа.

4. Валидизируй, валидизируй, валидизируй [44].

[1] Что, конечно, ни в коем случае не отменяет значимость этих задач или их применимость в социальных науках в целом. Заинтересованного читателя я бы хотел отослать к следующим работам по этим направлениям: шкалирование см.: [49; 50; 51; 52], анализ вторичного использования текстов см.: [53; 54; 55], анализ личностных черт автора текста см.: [56; 57; 58]. При этом необходимо также понимать, что шкалирование может, на самом деле, решаться с помощью двух рассматриваемых в статье подходов, однако у этих подходов есть несколько особенностей, которые необходимо учитывать (например, нужно, чтобы тексты располагались на одном измерении, если нужно применить алгоритм машинного обучения без учителя; об этом см. дальше в тексте).

Хотелось бы прокомментировать первые два принципа. Первый принцип отсылает к тому, что сложность языка подразумевает, что все используемые методы не могут предоставить точное описание текста, а оценивать эти неправильные модели нужно на основе их возможностей выполнить какую-то полезную для исследователя задачу. Бенуа пишет, что получение инсайта в практике использования текста-как-данных возможно только в том случае, когда процедура подготовки текста к анализу исключает возможность понять текст, так как необходимо уничтожить структуру оригинального текста и превратить его стилизированные и упрощенные переменные (слова, словосочетания) в таблицу, которую ни один читатель напрямую уже не сможет интерпретировать [2]. Под вторым принципом Гриммер и Стюарт подразумевают то, что методы автоматической обработки не позволяют избавиться от необходимости хорошего дизайна (см. ниже количество решений, которые принимают исследователи) или чтения текстов, скорее они просто увеличивают наши возможности в анализе текстов. Хорошим примером может быть методический эксперимент по сравнению человеческого кодирования и кодирования через несколько автоматических алгоритмов [47], в результате которого можно сказать, что человеческое кодирование не исчезнет, а скорее может работать в тандеме с автоматическим. Более того, одно из ключевых преимуществ такого типа текстового анализа как раз в том, что он позволяет заметить паттерны, которые человек сам заметить не может.

К. Бенуа описывает процедуру предварительной обработки текстов (превращения текста в данные, то есть в матрицу) следующим образом: сначала каждый документ подвергается процедуре токенизации, потом исключается пунктуация и «стоп-слова», слова приводятся к одному регистру и подвергаются стеммингу/лемматизации. После этого каждый документ становится строкой в матрице, а каждое слово или другая единица наблюдения (например, два слова вместе, так называемые биграммы) во всех документах –

столбцами, а ячейки заполняются количеством наблюдений в каждом документе, так называемая матрица термин-документ или матрица документ-признак [2]. Рассмотрим каждую из процедур подробнее. Под токенизацией обычно понимают процесс выделения слов (токены) из документов [59][1]. Исключение пунктуации, стоп-слов[2], приведение к одному регистру, удаление слов, встречающихся в малом количестве документов, стемминг или лемматизация[3] – это классические способы снижения размерности матрицы [59; 61] (про проклятие размерности см. подр.: [62]). Здесь важно то, что каждый из этих шагов влечет за собой определенные последствия для анализа, поэтому все эти решения по предварительной обработке данных должны приниматься с учетом потенциальных их последствий для анализа [63; 64; 65][4].

Наконец, можно перейти к небольшому обсуждению двух задач на классификацию: тематическое моделирование и разметка текстов на основе существующей классификации. Вторая задача может нуждаться в пояснении. Под существующей классификацией понимается необходимость разметить большой массив текста по

[1] При этом необходимо понимать, что исследователи не обязательно работают со словами, иногда вместо «мешка слов» могут использоваться «мешки n-граммов», когда токеном становится не слово, а, например, пара, другое количество идущих вместе слов (биграммы или n-граммы) или сочетание слов (которые, кстати, называют униграммами) и, например, биграмм (см., напр.: [60]).

[2] Под стоп-словами обычно понимают слишком часто встречающиеся слова, которые при этом не очень дискриминативны [59, p. 672], например предлоги.

[3] Под *стеммингом* обычно понимают процесс превращения (*collapse together*) различных морфологических вариантов слова в один [59, p. 786]. Под *лемматизацией* понимают приведение слов к их базовой форме или словарной форме (*citation form*) [59, p. 623]. Оба эти процесса направлены на то, чтобы слова, например в разных числах или склонениях, воспринимались как одно слово, что, в свою очередь, приводит к снижению размерности матрицы термин-документ.

[4] Например М.Дж. Денни и А. Спирлинг показывают, что небольшие изменения в процессе предподготовки текста могут привести к содержательно различающимся результатам.

заданным исследователем параметрам (относится ли этот текст, например, к теме дистанционного обучения или нет).

Тематическое моделирование относится к широкому классу применения алгоритмов машинного обучения без учителя[1] к нереактивным текстовым данным, трансформированным в матрицу документ-термин. Тематические модели – это «статистические алгоритмы, направленные на идентификацию и измерение латентных (скрытых) тем внутри корпуса текстовых документов» [67, p. 1]. Тематические модели можно разделить на две группы: 1) те, что предполагают, что каждый документ может иметь только одну тему (*single-membership models*) и 2) те, что предполагают, что каждый документ может содержать несколько тем (*mixed-membership models*) [68]. Модели, которые предполагают, что каждый документ может иметь только одну тему, реализуются, например, с помощью кластерного анализа (*k*-средние, *k*-медианы и т.д.), однако куда большую популярность приобрели модели, предполагающие, что в каждом документе может быть множество тем. На данный момент существует множество таких тематических моделей: классическое латентное размещение Дирихле (LDA) [62; 69], скоррелированные тематические модели [70; 71], динамичные тематические модели [72; 73], иерархические тематические модели [38; 74; 75] и структурные тематические модели, которые показали себя особенно полезными для социальных наук [34; 68; 76]. Большая часть таких тематических моделей включает в себя итеративное включение/исключение часто повторяющихся слов, решение относительно количества тем, исследование сетевой структуры тем и именование тем, которые были определены на основе списка наиболее характерных для этих тем слов [61].

[1] Машинное обучение без учителя в широком смысле может быть понято как «класс методов, направленный на анализ входных данных (X) без отсылки к истинным выходным данным (Y)» [66, p. 28]. В такие методы входят: анализ главных компонент, факторный анализ, кластерный анализ, анализ латентных классов, анализ последовательностей, тематическое моделирование и поиск сообществ в анализе социальных сетей (SNA) и т.д. [66].

Приведу пример из исследования Н. Линдстедта, в котором использовалось структурное тематическое моделирование на материалах аннотаций научных статей по теме социальных движений, опубликованных в 11 журналах в период с 2005 по 2017 г. [34]. Структурное тематическое моделирование интересно тем, что оно позволяет учитывать групповые характеристики текста (гендер автора, год написания, географические регионы и т.д.) как ковариаты, обладающие эффектом на распространенность темы и/или на ее содержание [76]. По результатам использования этой модели Линдстедт классифицировал аннотации на 24 темы. Например, самой выраженной темой в этих аннотациях было вовлечение новых участников в социальное движение (*recruitment*), а самой невыраженной темой – рабочее движение [34, p. 312]. Это решение позволяет не только достаточно быстро охарактеризовать современные работы по социальным движениям, но и найти документы (аннотации), в которых наиболее выражены те или иные темы. Далее Линдстедт включает в модель в качестве ковариата среднее количество цитирований статей в год. Этот ковариат позволил ему заметить, что несмотря на то, что количество статей с темой «женские движения» падало в исследуемый период, а количество статей по теме «раса/этничность» росло, влиятельность этих тем, понимаемая как количество цитирований в год, оставалась идентичной [34, p. 313–314]. Еще один интересный результат его работы связан с построением графа корреляции между темами, который позволил автору заметить, какие темы связаны между собой и встречаются вместе, а какие встречаются редко. Например, Линдстедт показывает, что тема «мобилизация» не связана ни с одной из других 23 тем (сила связи > 0,01), а тема «раса/этничность» связана только с темой «гражданские/человеческие права», которая, в свою очередь, связана только с «политическими возможностями». Таким образом, Линдстедт смог из аннотаций статей выделить темы, которые в них встречаются, распространенность этих тем в разные годы и распространенность этих тем при

учете их влиятельности, а также то, как эти темы связаны между собой в аннотациях.

Разметка текстов на основе существующей классификации относится к широкому классу применения алгоритмов машинного обучения с учителем[1] к нереактивным текстовым данным, трансформированным в матрицу документ-термин. Разметка текстов может использоваться, например, для помощи в ручном кодировании [26; 48; 77; 78; 79]. Существует на сегодняшний день множество методов разметки текстов (наивный Байесовский классификатор, *k*-ближайшие соседи, деревья решений и т.д.) [80], однако все они включают в себя три базовых шага: «...(1) конструирование тренировочной выборки, (2) применение алгоритма машинного обучения с учителем, направленного на изучение отношений между признаками и категориями в тренировочной выборке, а затем применение результатов этого анализа на тестовой выборке, и (3) валидизация результатов модели и классификация оставшихся документов» [44, p. 275].

Интересный пример такой классификации приведен в исследовании, в котором пытались предсказать результаты голосования по Брекзиту на основе классификации общественного мнения через данные 23 млн сообщений в *Twitter* [78]. Авторы сконструировали тренировочную и тестовую выборки на основе данных хэштегов (опциональный способ пользователей *Twitter* помечать тему сообщения; например, фраза «*#ГолосуйЗаБрекзит*» в сообщении). Для того, чтобы отобрать сообщения тех, кто хочет покинуть Европу, использовалось два хэштега вместе: *#VoteLeave* и *#TakeControl*. Для отбора тех сообщений, в которых выражено

[1] Машинное обучение с учителем в широком смысле подразумевает «класс методов, использующих тренировочные данные пар входных данных (X) и выходных данных (Y) для того, чтобы обучиться параметрам, которые предсказывают Y через X на новых массивах данных» [66, p. 28]. В такие методы могут входить пенализированные регрессии, деревья классификаций и регрессий, методы ближайшего соседа и т.д. [66].

стремление остаться в Европе, также использовалось два хэштега: *#StrongerIn* и *#Remain*, *VoteRemain*, *LabourInForBritain* или *InTogether*. В итоге было отобрано 116 886 сообщений, из них в тренировочную выборку попало 78 300 сообщений, а в тестовую – 38 566 сообщений [78, p. 6–7]. В качестве алгоритма машинного обучения с учителем авторы использовали метод опорных векторов (SVM). Модель продемонстрировала высокое качество на тестовой выборке (97,05% точности для предсказания «остаться» и 97,12% для предсказания «покинуть»), а также высокую точность и полноту (от 95 до 100%). На основе этой модели авторы классифицировали оставшиеся 23 млн сообщений [78, p. 8–10]. Одним из интересных итогов этой статьи было сопоставление результатов классификации сообщений из *Twitter* с данными интернет-опросов и телефонных опросов. Данные из *Twitter* показали высокую корреляцию с результатами интернет-опросов и низкую корреляцию с результатами телефонных опросов [78, p. 11–12].

Заключение

В этой статье я стремился вкратце рассмотреть основные способы применения интеллектуального анализа текстовых данных в социальных науках. Я стремился показать, что хотя ИАТ может быть полезен и для *холистического*, и для *атомарного* восприятия текста, свою популярность эти методы приобрели именно в атомарном восприятии текста. На примере классического количественного контент-анализа было показано, какие проблемы ИАТ уже решает в социальных науках: воспроизводимость результатов, возможность работы с большими объемами выборок, возможность выйти за пределы четко определенных словарей. Наконец, я стремился обозначить основные принципы проведения исследования с предположением «мешок слов» в целом и на примере задачи классификации в двух ее вариантах: тематическое моделирование и разметка текстов на основе существующей классификации.

Основное ограничение моделей, использующих «мешок слов», – это то, что они игнорируют контекст использования тех или иных слов [67]. Эта проблема решается в контекстных моделях языка (*word embeddings*), где слова используются не по отдельности, а в векторе [81]. Эти модели показали лучшее качество, чем использование «мешка слов», – например, для построения тех же тем. Они уже показали интересные результаты в исследованиях стереотипов [82], культуры [36], политических предпочтений [83], эмоции в тексте [84] и протестных акций [85].

За скобками этой статьи осталась разработка теоретической рамки для исследований с применением ИАТ в социальных науках. Здесь я могу порекомендовать обратиться к работам Г. Игнатоу, который стремится показать, что количественные методы анализа текстовых данных должны быть основаны на реалистской конструктивистской онтологии (см.: [15; 86]).

ЛИТЕРАТУРА

1. Text Mining for Central Banks: Handbook / D. Bholat [et al.] // LSE Research Online [site]. Last update: 11.04.2020. URL: http://eprints.lse.ac.uk/62548/ (date of access: 15.12.2019).

2. *Benoit K.* Text as Data: An Overview. Version: 17.07.2019. URL: https://kenbenoit.net/pdfs/28%20Benoit%20Text%20as%20Data%20draft%202.pdf (date of access: 15.12.2019).

3. *Девятко И.* Инструментарий онлайн-исследований: попытка каталогизации // Онлайн исследования в России 3.0. Москва: Online Market Intelligence, 2012. С. 17–31.

4. *Lazarsfeld P.F., Oberschall A.R.* Max Weber and Empirical Social Research // American Sociological Review. 1965. Vol. 30. No. 2. P. 185–199.

5. *Krippendorff K.* Content Analysis: An Introduction to its Methodology. Thousands Oaks, CA: Sage Publications, 2004.

6. *Evans J.A., Aceves P.* Machine Translation: Mining Text for Social Theory // Annual Review of Sociology. 2016. No. 42. P. 21–50.

7. *Bail C.A.* The Cultural Environment: Measuring Culture with Big Data // Theory and Society. 2014. Vol. 43. No. 3–4. P. 465–482.

8. 1 Second – Internet Live Stats. URL: https://www.internetlivestats.com/one-second/ (date of access: 03.11.2019).

9. *Ledford H.* Facebook Gives Social Scientists Unprecedented Access to its User Data // Nature [site]. 2019. May 03. URL: https://www.nature.com/articles/d41586-019-01447-5 (date of access: 15.12.2019).

10. *Kitchin R.* Big Data, New Epistemologies and Paradigm Shifts // Big Data & Society. 2014. Vol. 1. No. 1. P. 1–12.

11. *Salganik M.* Bit by Bit: Social Research in the Digital Age. Princeton, NJ: Princeton University Press, 2019.

12. *Golder S.A., Macy M.W.* Digital Footprints: Opportunities and Challenges for Online Social Research // Annual Review of Sociology. 2014. No. 40. P. 129–152.

13. *Lazer D., Radford J.* Data ex Machina: Introduction to Big Data // Annual Review of Sociology. 2017. No. 43. P. 19–39.

14. *Boréus K., Bergström G.* Analyzing Text and Discourse: Eight Approaches for the Social Sciences. Washington, DC: Melbourne Sage, 2017.

15. *Ignatow G., Mihalcea R.* An Introduction to Text Mining: Research Design, Data Collection, and Analysis. Thousand Oaks, CA: Sage Publications, 2018.

16. *Bryman A.* Social Research Methods. Oxford: Oxford Univ. Press, 2016.

17. *Nelson L.K.* Computational Grounded Theory: A Methodological Framework // Sociological Methods & Research. 2020. Vol. 49. No. 1. P. 3–42.

18. *Cioffi-Revilla C.* Introduction to Computational Social Science: Principles and Applications. Fairfax, VA: Springer, 2017.

19. *Chen S.-H.* Big Data in Computational Social Sciences and Humanities. Cham: Springer, 2018.

20. Computational Social Science / D. Lazer [et al.] // Science. 2009. Vol. 323. No. 5915. P. 721–723.

21. A Useful Methodological Synergy? Combining Critical Discourse Analysis and Corpus Linguistics to Examine Discourses of Refugees and Asylum Seekers in The UK Press / P. Baker [et al.] // Discourse & Society. 2008. Vol. 19. No. 3. P. 273–306.

22. *Bednarek M., Caple H.* Why Do News Values Matter? Towards a New Methodological Framework for Analyzing News Discourse in Critical Discourse Analysis and Beyond // Discourse & Society. 2014. Vol. 25. No. 2. P. 135–158.

23. *Jo W.* Possibility of Discourse Analysis Using Topic Modeling // Journal of Asian Sociology. 2019. Vol. 48. No. 3. P. 321–342.

24. *Berelson B.* Content Analysis in Communication Research. Glencoe, IL: Free Press, 1952.

25. *Neuendorf K.A.* The Content Analysis Guidebook. Los Angeles, CA: Sage Publications, 2017.

26. *Mikhaylov S., Laver M., Benoit K.R.* Coder Reliability and Misclassification in the Human Coding of Party Manifestos // Political Analysis. 2012. Vol. 20. No. 1. P. 78–91.

27. *Лукашевич Н.В., Левчик А.В.* Создание лексикона оценочных слов русского языка РуСентилекс // OSTIS-2016: материалы VI междунар. науч.-

техн. конф. (Минск, 18–20 февраля 2016 года) / Отв. ред. В.В. Голенков. Минск: БГУИР, 2016. С. 377–382.

28. *Stine R.A.* Sentiment Analysis // Annual Review of Statistics and Its Application. 2019. Vol. 6. No. 1. P. 287–308.

29. Analyzing Media Messages: Using Quantitative Content Analysis in Research / D. Riff [et al.] London: Routledge, 2019.

30. *Feldman R., Sanger J.* The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge: Cambridge Univ. Press, 2007.

31. *Wachsmuth H.* Text Analysis Pipelines: Towards Ad-hoc Large-Scale Text Mining. Cham: Springer, 2015.

32. Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications / G. Miner [et al.] Amsterdam: Academic Press, 2012.

33. *Hanna A.* Computer-aided Content Analysis of Digitally Enabled Movements // Mobilization: An International Quarterly. 2013. Vol. 18. No. 4. P. 367–388.

34. *Lindstedt N.C.* Structural Topic Modeling for Social Scientists: A Brief Case Study with Social Movement Studies Literature, 2005–2017 // Social Currents. 2019. Vol. 6. No. 4. P. 307–318.

35. Big Data, Social Media, and Protest: Foundations for a Research Agenda / J.A. Tucker [et al.] // Computational Social Science: Discovery and Prediction. New York: Cambridge Univ. Press, 2016. P. 199–224.

36. *Kozlowski A.C., Taddy M., Evans J.A.* The Geometry of Culture: Analyzing Meaning through Word Embeddings // American Sociological Review. 2019. Vol. 84. No. 5. P. 905–949.

37. *Brady H.E.* The Challenge of Big Data and Data Science // Annual Review of Political Science. 2019. Vol. 22. No. 1. P. 297–323.

38. *Grimmer J.* A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases // Political Analysis. 2010. Vol. 18. No. 1. P. 1–35.

39. *Grimmer J.* Measuring Representational Style in the House: The Tea Party, Obama, and Legislators' Changing Expressed Priorities // Computational Social Science: Discovery and Prediction. New York: Cambridge Univ. Press, 2016. P. 225–245.

40. *Slapin J.B., Proksch S.-O.* A Scaling Model for Estimating Time-series Party Positions from Texts // American Journal of Political Science. 2008. Vol. 52. No. 3. P. 705–722.

41. *Young L., Soroka S.* Affective News: The Automated Coding of Sentiment in Political Texts // Political Communication. 2012. Vol. 29. No. 2. P. 205–231.

42. *Proksch S.-O., Slapin J.B.* Position Taking in European Parliament Speeches // British Journal of Political Science. 2010. Vol. 40. No. 3. P. 587–611.

43. *Roberts M.E.* Introduction to the Virtual Issue: Recent Innovations in Text Analysis for Social Science // Political Analysis. 2016. Vol. 24. No. 10. P. 1–5.

44. *Grimmer J., Stewart B.M.* Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts // Political Analysis. 2013. Vol. 21. No. 3. P. 267–297.

45. *Schoonvelde M., Schumacher G., Bakker B.N.* Friends with Text as Data Benefits: Assessing and Extending the Use of Automated Text Analysis in Political Science and Political Psychology // Journal of Social and Political Psychology. 2019. Vol. 7. No. 1. P. 124–143.

46. *Wilkerson J., Casas A.* Large-scale Computerized Text Analysis in Political Science: Opportunities and Challenges // Annual Review of Political Science. 2017. No. 20. P. 529–544.

47. Computational Communication Science: A Methodological Catalyzer for a Maturing Discipline / M. Hilbert [et al.] // International Journal of Communication. 2019. No. 13. P. 3913–3934.

48. The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods / L.K. Nelson [et al.] // Sociological Methods & Research. 2018. P. 1–36.

49. *Klüver H.* Measuring Interest Group Influence Using Quantitative Text Analysis // European Union Politics. 2009. Vol. 10. No. 4. P. 535–549.

50. *Baerg N., Lowe W.* A Textual Taylor Rule: Estimating Central Bank Preferences Combining Topic and Scaling Methods // Political Science Research and Methods. 2020. Vol. 8. No. 1. P. 106–122.

51. *Lowe W., Benoit K.* Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark // Political Analysis. 2013. Vol. 21. No. 3. P. 298–313.

52. *Lowe W.* Understanding Wordscores // Political Analysis. 2008. Vol. 16. No. 4. P. 356–371.

53. Text as Policy: Measuring Policy Similarity through Bill Text Reuse / F. Linder [et al.] // Policy Studies Journal. 2020. Vol. 48. P. 546–574..

54. *Allee T., Lugg A*. Who Wrote the Rules for the Trans-Pacific Partnership? // Research & Politics. 2016. Vol. 3. No. 3. P. 1–9.

55. *Wilkerson J., Smith D., Stramp N.* Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach // American Journal of Political Science. 2015. Vol. 59. No. 4. P. 943–956.

56. Automatic Personality Assessment through Social Media Language / G. Park [et al.] // Journal of Personality and Social Psychology. 2015. Vol. 108. No. 6. P. 934–952.

57. Personality, Gender, and Age in the Language of Social Media: The Open-vocabulary Approach / H.A. Schwartz [et al.] // PLOS ONE. 2013. Vol. 8. No. 9. P. 1–16.

58. *Schwartz H.A., Ungar L.H.* Data-Driven Content Analysis of Social Media: A Systematic Overview of Automated Methods // The ANNALS of the American Academy of Political and Social Science. 2015. Vol. 659. No. 1. P. 78–94.

59. *Jurafsky D., Martin J.H.* Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing // Upper Saddle River, NJ: Prentice Hall, 2008.

60. *Bekkerman R., Allan J.* Using Bigrams in Text Categorization. 27.12.2003. URL: http://ciir.cs.umass.edu/pubfiles/ir-408.pdf (date of access: 15.12.2019).

61. A Review of Best Practice Recommendations for Text Analysis in R (and a UserFriendly App) / G.C. Banks [et al.] // Journal of Business and Psychology. 2018. Vol. 33. No. 4. P. 445–459.

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

63. *Schofield A., Mimno D.* Comparing Apples to Apple: The Effects of Stemmers on Topic Models // Transactions of the Association for Computational Linguistics. 2016. No. 4. P. 287–300.

64. *Denny M.J., Spirling A.* Text Preprocessing for Unsupervised Learning: Why it Matters, When it Misleads, and What to Do about it // Political Analysis. 2018. Vol. 26. No. 2. P. 168–189.

65. *Schofield A., Magnusson M., Mimno D.* Pulling out the Stops: Rethinking Stopword Removal for Topic Models // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Vol. 2: Short Papers. Valencia: Association for Computational Linguistics, 2017. P. 432–436.

66. *Molina M., Garip F.* Machine Learning for Sociology // Annual Review of Sociology. 2019. No. 45. P.27–45.

67. *Wesslen R.* Computer-assisted Text Analysis for Social Science: Topic Models and Beyond. 03.04.2018. URL: https://arxiv.org/pdf/1803.11045 (date of access: 15.12.2019).

68. Structural Topic Models for Open-ended Survey Responses / M.E. Roberts [et al.] // American Journal of Political Science. 2014. Vol. 58. No. 4. P. 1064–1082.

69. *Blei D.M.* Probabilistic Topic Models // Communications of the ACM. 2012. Vol. 55. No. 4. P. 77–84.

70. *Blei D.M., Lafferty J.D.* A Correlated Topic Model of Science // The Annals of Applied Statistics. 2007. Vol. 1. No. 1. P.17–35.

71. Efficient Correlated Topic Modeling with Topic Embedding / J. He [et al.] // Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Canada: ACM, 2017. P. 225–233.

72. *Blei D.M., Lafferty J.D.* Dynamic Topic Models // Proceedings of the 23rd International Conference on Machine Learning. New York: ACM, 2006. P. 113–120.

73. Scaling up Dynamic Topic Models / A. Bhadury [et al.] // Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences. Montreal, 2016. P. 381–390.

74. Hierarchical Topic Modeling with Automatic Knowledge Mining / Y. Xu [et al.] // Expert Systems with Applications. 2018. No. 103. P. 106–117.

75. Scalable Training of Hierarchical Topic Models / J. Chen [et al.] // Proceedings of the VLDB Endowment. 2018. Vol. 11. No. 7. P. 826–839.

76. *Roberts M., Stewart B., Tingley D.* Structural Topic Models. URL: https://www.structuraltopicmodel.com/ (date of access: 15.12.2019).

77. *Loftis M.W., Mortensen P.B.* Collaborating with the Machines: A Hybrid Method for Classifying Policy Documents // Policy Studies Journal. 2020. No. 48. P. 184–206.

78. Predicting the Brexit Vote by Tracking and Classifying Public Opinion Using Twitter Data / J. Amador [et al.] // Statistics, Politics and Policy. 2017. Vol. 8. No. 1. P. 85–104.

79. *Watanabe K.* Newsmap // Digital Journalism. 2018. Vol. 6. No. 3. P. 294–309.

80. *Anandarajan M., Hill C., Nelson T.* Classification Analysis: Machine Learning Applied to Text // Practical Text Analytics: Maximizing the Value of Text Data. Switzerland: Springer, 2019. P. 131–149.

81. Distributed Representations of Words and Phrases and their Compositionality / T. Mikolov [et al.] // Advances in Neural Information Processing Systems. 2013. P. 3111–3119.

82. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes / N. Garg [et al.] // Proceedings of the National Academy of Sciences. 2018. Vol. 115. No. 16. P. E3635–E3644.

83. *Gurciullo S., Mikhaylov S.J.* Detecting Policy Preferences and Dynamics in the Un General Debate with Neural Word Embeddings // 2017 International Conference on the Frontiers and Advances in Data Science (FADS). Xi'an, China: IEEE, 2017. P. 74–79.

84. *Seyeditabari A., Zadrozny W.* Can Word Embeddings Help Find Latent Emotions in Text? Preliminary Results // The Thirtieth International Flairs Conference. Marco Island, USA, 2017. P. 206–209.

85. *Zhang H., Pan J.* CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media // Sociological Methodology. 2019. Vol. 49. No. 1. P. 1–57.

86. *Ignatow G.* Theoretical Foundations for Digital Text Analysis // Journal for the Theory of Social Behaviour. 2016. Vol. 46. No. 1. P. 104–120.

Byzov Alexander,
National Research University Higher School of Economics (NRU HSE), Moscow, debesergopotes12@gmail.com

Text mining in social sciences

Throughout most of their history, sociologists have sought to study unstructured organic texts: newspaper materials, diaries, memoirs, letters, documents, and, more recently, messages, publications, and other writings on various online platforms. This article discusses how modern techniques of text mining can improve classical sociological approaches to the analysis of this type of data. The material is structured according to the following plan. First, examples of traditional quantitative content analysis and its limitations are discussed that could be solved with the help of text mining. Then I demonstrate how text mining is applied in contemporary social science research with two examples. First, I review a paper in which structural topic modeling was used to investigate topics of scholars' articles on social movements, the prevalence of discovered topics throughout the years, and the links between these topics. Second, I present the results of another study in which automatically classified Twitter messages were used compared to online and phone surveys. Finally, I conclude with a discussion of some of the current approaches to text analysis using deep learning and theoretical issues related to the application of text mining
Keywords: text mining, topic modeling, text classification, methodology and methods

References

1. Bholat D. [et al.] Text Mining for Central Banks: Handbook, *LSE Research Online [site]*. Last update: 11.04.2020. URL: http://eprints.lse.ac.uk/62548/ (date of access: 15.12.2019).
2. Benoit K. *Text as Data: An Overview*. Version: 17.07.2019. URL: https://kenbenoit.net/pdfs/28%20Benoit%20Text%20as%20Data%20draft%202.pdf (date of access: 15.12.2019).
3. Devyatko I. "Online research toolkit: an attempt at cataloging" (in Russian), in: *Onlajn issledovaniya v Rossii 3.0*. Moskva: Online Market Intelligence, 2012. P. 17–31.
4. Lazarsfeld P.F., Oberschall A.R. Max Weber and Empirical Social Research, *American Sociological Review*, 1965, 30 (2), 185–199.

5. Krippendorff K. *Content analysis: An Introduction to Its Methodology*. Thousands Oaks, California: Sage Publications, 2004.
6. Evans J.A., Aceves P. Machine translation: Mining Text for Social Theory, *Annual Review of Sociology,* 2016, 42, 21–50.
7. Bail C.A. The Cultural Environment: Measuring Culture with Big Data, *Theory and Society,* 2014, 43 (3–4), 465–482.
8. *1 Second – Internet Live Stats*. URL: https://www.internetlivestats.com/one-second/ (date of access: 03.11.2019).
9. Ledford H. *Facebook gives social scientists unprecedented access to its user data*. URL: https://www.nature.com/articles/d41586-019-01447-5 (date of access: 15.12.2019).
10. Kitchin R. Big Data, New Epistemologies and Paradigm Shifts, *Big Data & Society*, 2014, 1 (1), 1–12.
11. Salganik M. *Bit by Bit: Social Research In The Digital Age*. Princeton, New Jersey: Princeton University Press, 2019.
12. Golder S.A., Macy M.W. Digital footprints: Opportunities and challenges for online social research, *Annual Review of Sociology,* 2014, 40, 129–152.
13. Lazer D., Radford J. Data ex Machina: Introduction to Big Data, *Annual Review of Sociology,* 2017, 43, 19–39.
14. Boréus K., Bergström G. *Analyzing Text and Discourse: Eight Approaches for The Social Sciences*. Washington DC: Melbourne Sage, 2017.
15. Ignatow G., Mihalcea R. *An Introduction to Text Mining: Research Design, Data Collection, and Analysis*. Thousand Oaks, California: Sage Publications, 2018.
16. Bryman A. *Social Research Methods.* Oxford: Oxford University Press, 2016.
17. Nelson L.K. Computational Grounded Theory: A Methodological Framework, *Sociological Methods & Research,* 2020, 49 (1), 3–42.
18. Cioffi-Revilla C. *Introduction to Computational Social Science: Principles and Applications*. Fairfax, VA: Springer, 2017.
19. Chen S.-H. *Big Data in Computational Social Sciences and Humanities*. Cham: Springer, 2018.
20. D. Lazer [et al.] Computational Social Science, *Science*, 2009, 323 (5915), 721–723.

21. P. Baker [et al.] A Useful Methodological Synergy? Combining Critical Discourse Analysis and Corpus Linguistics to Examine Discourses of Refugees and Asylum Seekers in The UK Press, *Discourse & Society*, 2008, 19 (3), 273–306.
22. Bednarek M., Caple H. Why Do News Values Matter? Towards a New Methodological Framework for Analysing News Discourse in Critical Discourse Analysis and Beyond, *Discourse & Society,* 2014, 25 (2), 135–158.
23. Jo W. Possibility of Discourse Analysis using Topic Modeling, *Journal of Asian Sociology,* 2019, 48 (3), 321–342.
24. Berelson B. *Content Analysis in Communication Research*. Glencoe, Ill.: Free Press, 1952.
25. Neuendorf K.A. *The Content Analysis Guidebook*. Los Angeles, California: Sage Publications, 2017.
26. Mikhaylov S., Laver M., Benoit K.R. Coder Reliability and Misclassification in The Human Coding Of Party Manifestos, *Political Analysis,* 2012, 20 (1), 78–91.
27. Lukashevich N.V., Levchik A.V. "Creation of a lexicon of evaluative words of the Russian language RuSentylex" (in Russian), in: *OSTIS-2016: materialy VI mezhdunar. nauch.-tekhn. konf.* (Minsk, February, 18–20, 2016). Minsk: BGUIR, 2016. P. 377–382.
28. Stine R.A. Sentiment Analysis, *Annual Review of Statistics and Its Application,* 2019, 6 (1), 287–308.
29. D. Riff [et al.] *Analyzing Media Messages: Using Quantitative Content Analysis in Research.* London: Routledge, 2019.
30. Feldman R., Sanger J. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Cambridge: Cambridge University Press, 2007.
31. Wachsmuth H. *Text Analysis Pipelines: Towards Ad-Hoc Large-Scale Text Mining.* Cham: Springer, 2015.
32. G. Miner [et al.] *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*. Amsterdam: Academic Press, 2012.
33. Hanna A. Computer-Aided Content Analysis of Digitally Enabled Movements, *Mobilization: An International Quarterly*, 2013, 18 (4), 367–388.

34. Lindstedt N.C. Structural Topic Modeling for Social Scientists: A Brief Case Study with Social Movement Studies Literature, 2005–2017, *Social Currents,* 2019, 6 (4), 307–318.
35. Tucker J.A. [et al.] "Big Data, Social Media, and Protest: Foundations for a Research Agenda", in: *Computational Social Science: Discovery and Prediction*. New York: Cambridge University Press, 2016. P. 199–224.
36. Kozlowski A.C., Taddy M., Evans J.A. The Geometry of Culture: Analyzing Meaning Through Word Embeddings, *American Sociological Review*, 2019, 84 (5), 905–949.
37. Brady H.E. The Challenge of Big Data and Data Science, *Annual Review of Political Science*, 2019, 22 (1), 297–323.
38. Grimmer J. A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases, *Political Analysis,* 2010, 18 (1), 1–35.
39. Grimmer J. "Measuring Representational Style in the House: The Tea Party, Obama, and Legislators' Changing Expressed Priorities", in: *Computational Social Science: Discovery and Prediction*. New York: Cambridge University Press, 2016. P. 225–245.
40. Slapin J.B., Proksch S.-O. A Scaling Model for Estimating Time-Series Party Positions from Texts, *American Journal of Political Science*, 2008, 52 (3), 705–722.
41. Young L., Soroka S. Affective News: The Automated Coding of Sentiment in Political Texts, *Political Communication*, 2012, 29 (2), 205–231.
42. Proksch S.-O., Slapin J.B. Position Taking in European Parliament Speeches, *British Journal of Political Science*, 2010, 40 (3), 587–611.
43. Roberts M.E. Introduction to the Virtual Issue: Recent Innovations in Text Analysis for Social Science, *Political Analysis*, 2016, 24 (10), 1–5.
44. Grimmer J., Stewart B.M. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts, *Political Analysis*, 2013, 21 (3), 267–297.
45. Schoonvelde M., Schumacher G., Bakker B.N. Friends with Text as Data Benefits: Assessing and Extending the Use of Automated Text Analysis in Political Science and Political Psychology, *Journal of Social and Political Psychology*, 2019, 7 (1), 124–143.
46. Wilkerson J., Casas A. Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges, *Annual Review of Political Science*, 2017, 20, 529–544.

47. Hilbert M. [et al.] Computational Communication Science: A Methodological Catalyzer for a Maturing Discipline, *International Journal of Communication*, 2019, 13, 3913–3934.
48. Nelson L.K. [et al.] The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods, *Sociological Methods & Research*, 2018, 1–36.
49. Klüver H. Measuring Interest Group Influence Using Quantitative Text Analysis, *European Union Politics*, 2009, 10 (4), 535–549.
50. Baerg N., Lowe W. A Textual Taylor Rule: Estimating Central Bank Preferences Combining Topic and Scaling Methods, *Political Science Research and Methods*, 2020, 8 (1), 106–122.
51. Lowe W., Benoit K. Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark, *Political Analysis,* 2013, 21 (3), 298–313.
52. Lowe W. Understanding Wordscores, *Political Analysis*, 2008, 16 (4). P. 356–371.
53. Linder F. [et al.] Text as Policy: Measuring Policy Similarity through Bill Text Reuse, *Policy Studies Journal,* 2020, 48, 546–574.
54. Allee T., Lugg A. Who wrote the rules for the Trans-Pacific Partnership? *Research & Politics*, 2016, 3 (3), 1–9.
55. Wilkerson J., Smith D., Stramp N. Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach, *American Journal of Political Science,* 2015, 59 (4), 943–956.
56. Park G. [et al.] Automatic Personality Assessment Through Social Media Language, *Journal of Personality and Social Psychology,* 2015, 108 (6), 934–952.
57. Schwartz H.A. [et al.] Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach, *PLOS ONE,* 2013, 8 (9), 1–16.
58. Schwartz H.A., Ungar L.H. Data-Driven Content Analysis of Social Media: A Systematic Overview of Automated Methods, *The ANNALS of the American Academy of Political and Social Science*, 2015, 659 (1), 78–94.
59. Jurafsky D., Martin J.H. *Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing*. Upper Saddle River, NJ: Prentice Hall, 2008.

60. Bekkerman R., Allan J. *Using Bigrams in Text Categorization*. 27.12.2003. URL: http://ciir.cs.umass.edu/pubfiles/ir-408.pdf (date of access: 15.12.2019).
61. Banks G.C. [et al.] A Review of Best Practice Recommendations for Text Analysis in R (and a UserFriendly App), *Journal of Business and Psychology,* 2018, 33 (4), 445–459.
62. Koltsova O.Y., Maslinsky K.A. Identifying the Thematic Structure of the Russian Blogosphere: Automatic Text Analysis Methods (in Russian) // *Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling),* 2013, 36, 113–139.
63. Schofield A., Mimno D. Comparing Apples to Apple: The Effects of Stemmers on Topic Models, *Transactions of the Association for Computational Linguistics,* 2016, 4, 287–300.
64. Denny M.J., Spirling A. Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, And What to Do About It, *Political Analysis,* 2018, 26 (2), 168–189.
65. Schofield A., Magnusson M., Mimno D. "Pulling Out the Stops: Rethinking Stopword Removal for Topic Models", in: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Vol. 2: Short Papers. Valencia, Spain: Association for Computational Linguistics, 2017. P. 432–436.
66. Molina M., Garip F. Machine learning for sociology, *Annual Review of Sociology,* 2019, 45, 27–45.
67. Wesslen R. *Computer-Assisted Text Analysis for Social Science: Topic Models and Beyond*. 03.04.2018. URL: https://arxiv.org/pdf/1803.11045 (date of access: 15.12.2019).
68. M.E. Roberts [et al.] Structural Topic Models for Open-Ended Survey Responses, *American Journal of Political Science,* 2014, 58 (4), 1064–1082.
69. Blei D.M. Probabilistic Topic Models, *Communications of the ACM,* 2012, 55 (4), 77–84.
70. Blei D.M., Lafferty J.D. A Correlated Topic Model of Science, *The Annals of Applied Statistics,* 2007, 1 (1), 17–35.
71. He J. [et al.] "Efficient Correlated Topic Modeling with Topic Embedding", in: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax, Canada: ACM, 2017. P. 225–233.

72. Blei D.M., Lafferty J.D. “Dynamic Topic Models”, in: *Proceedings of the 23rd International Conference on Machine Learning*. New York, NY, USA: ACM, 2006. P. 113–120.
73. Bhadury A. [et al.] “Scaling Up Dynamic Topic Models”, in: *Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences*. Montreal, Canada. 2016. P. 381–390.
74. Xu Y. [et al.] Hierarchical Topic Modeling with Automatic Knowledge Mining, *Expert Systems with Applications*, 2018, 103, 106–117.
75. Chen J. [et al.] Scalable Training of Hierarchical Topic Models, *Proceedings of the VLDB Endowment*, 2018, 11 (7), 826–839.
76. Roberts M., Stewart B., Tingley D. *Structural Topic Models*. URL: https://www.structuraltopicmodel.com/ (date of access: 15.12.2019).
77. Loftis M.W., Mortensen P.B. Collaborating with the Machines: A Hybrid Method for Classifying Policy Documents, *Policy Studies Journal*, 2020, 48, 184–206.
78. Amador J. [et al.] Predicting the Brexit Vote by Tracking and Classifying Public Opinion Using Twitter Data, *Statistics, Politics and Policy*, 2017, 8 (1), 85–104.
79. Watanabe K. Newsmap, *Digital Journalism,* 2018, 6 (3), 294–309.
80. Anandarajan M., Hill C., Nelson T. “Classification Analysis: Machine Learning Applied to Text”, in: *Practical Text Analytics: Maximizing the Value of Text Data*. Switzerland: Springer, 2019. P. 131–149.
81. Mikolov T. [et al.] Distributed Representations of Words and Phrases and Their Compositionality, *Advances in Neural Information Processing Systems,* 2013, 3111–3119.
82. Garg N. [et al.] Word Embeddings Quantify 100 Years of Gender And Ethnic Stereotypes, *Proceedings of the National Academy of Sciences*, 2018, 115 (16), E3635–E3644.
83. Gurciullo S., Mikhaylov S.J. “Detecting Policy Preferences and Dynamics In The Un General Debate With Neural Word Embeddings”, in: *2017 International Conference on the Frontiers and Advances in Data Science (FADS)*. IEEE, Xi’an, China. 2017. P. 74–79.
84. Seyeditabari A., Zadrozny W. “Can Word Embeddings Help Find Latent Emotions in Text? Preliminary Results”, in: *The Thirtieth International Flairs Conference*. Marco Island, USA. 2017. P. 206–209.

85. Zhang H., Pan J. CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media, *Sociological Methodology,* 2019, 49 (1), 1–57.
86. Ignatow G. Theoretical Foundations for Digital Text Analysis, *Journal for the Theory of Social Behaviour,* 2016, 46 (1), 104–120.