На правах рукописи

Просянюк Дарья Вячеславовна

Методы тематической классификации текста (на примере образа Российской Федерации в New York Times)

Специальность: 22.00.01 –

Теория, методология и история социологии

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата социологических наук

Москва - 2014

Работа выполнена в Федеральном государственном автономном учреждении высшего профессионального образования «Национальный исследовательский университет «Высшая школа экономики»

Научный руководитель: кандидат социологических наук, доцент
Градосельская Галина Витальевна
доцент кафедры методов сбора и анализа социологической информации факультета социологии Национального исследовательского университета «Высшая школа экономики»

Официальные оппоненты: доктор социологических наук
Жаворонков Александр Васильевич
Ведущий научный сотрудник
Центра методологии социологических исследований Института социологии Российской академии наук

кандидат социологических наук
Крутий Ирина Андреевна
руководитель управления маркетинговыми интернет-коммуникациями «Современной гуманитарной академии»

Ведущая организация: **ФГОБУ ВПО МГИМО** Московский государственный институт международных отношений (Университет) МИД РФ

Защита состоится «3» марта 2015 года в 17.00 часов на заседании Совета по защите диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук Д 212.198.09 на базе Российского государственного гуманитарного университета по адресу: 125993, ГСП-3, Москва, Миусская площадь, дом 6, корп. 5, ауд. 406.

С диссертацией можно ознакомиться в научной библиотеке РГГУ по адресу: 125993, ГСП-3, Москва, Миусская площадь, д. 6. и на официальном сайте организации по адресу www.rsuh.ru.

Автореферат разослан «25» января 2015 года.

Ученый секретарь Буланова М.Б.
доктор социологических наук, профессор

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность

Стремительное распространение технологий производства, обработки, трансляции и хранения информации в текстовом виде, лавинообразный рост и широкая доступность данных в электронном виде, а также повышение роли информации как ресурса и основы принятия решений обусловили запрос на разработку автоматизированных средств обработки и анализа текстовых данных. В настоящее время мы являемся свидетелями интервенции формализованных методов анализа текстовых данных[1], что обусловливает завышенные ожидания к возможностям автоматизированных средств и их неадекватное использование. Основными причинами завышенных ожиданий, на наш взгляд, является агрессивная маркетинговая политика корпораций-разработчиков специализированного программного обеспечения, акцентирующая внимание на возможности практически полного исключения человека при сборе, обработке и анализе информации; низкий уровень осведомленности пользователей об алгоритмах и ограничениях методологий, лежащих в основе того или иного программного продукта; а также исключительно небольшое количество междисциплинарных научных исследований, направленных на решение задач определения «границ» и условий применения формализованных методов анализа текстовых данных в гуманитарных науках (которые являются одним из их основных «потребителей»).

Вместе с тем, подавляющее большинство современных гуманитарных исследований, содержащих этап обработки текстовой информации (в том числе ответы на открытые вопросы анкеты, транскрипты интервью и фокус-групп, тексты новостей и пр.), продолжают использовать традиционные методы ана-

[1] Формализованные методы анализа текстовых данных развивались обособленно, чаще в технических дисциплинах, таких как искусственный интеллект, нейросетевое моделирование, лингвистическое обеспечение систем автоматизированного проектирования и программирования и пр.

лиза, основанные на эвристических алгоритмах[2]: кодировании, априорной категоризации и пр. Основными причинами устойчивого применения традиционных методов анализа текстовых данных, на наш взгляд, является определенная степень инерционности методической составляющей исследований; неизученность, и, как следствие, отсутствие доказательств надежности и валидности формализованных методов при решении конкретных задач социального анализа; а также отсутствие исследований, посвященных верификации возможностей и условий интеграции различных направлений методов анализа текстовых данных.

Таким образом, в настоящее время наблюдается значительный разрыв между потенциальными возможностями формализованных методов анализа текстовых данных и фактическим использованием их потенциала. Применение формализованных методов для анализа больших массивов текстовых данных для решения задач социального анализа является скорее новаторством, чем нормой. Возможно, по причине того, что применение методов не стало нормой, они используются не всегда корректно и адекватно поставленным задачам.

Несмотря на взрывной рост количества методов и алгоритмов формализованного тематического анализа, крайне малочисленны исследования, дающие представления и конкретные руководства эмпирическому исследователю-гуманитарию об их специфике, достоинствах и недостатках. Узконаправлены и немногочисленны исследования, сфокусированные на сравнительной оценке применимости отдельных направлений методов анализа текстовых данных в конкретных исследовательских ситуациях, определяющие роль эвристических алгоритмов в процессе анализа. Следствием является отсутствие пошагового алгоритма анализа корпуса текстовых данных, основанного и направленного на решение конкретной задачи социального анализа, необходимого в эмпирических исследованиях. Сказанное позволяет считать, что работа, направленная на

[2] Под эвристическими алгоритмами понимается способ анализа данных и решения задач, не имеющий строгого обоснования, но дающий приемлемые решения в большинстве практически значимых задач.

изучение специфики и ограничений методов формализованного анализа текстовых данных и разработку стратегий их интеграции с эвристическими методами является актуальной. Подобная схема поможет систематизировать и адаптировать основные наработки точных наук в области анализа текстовых данных, продемонстрирует области единоличного «господства» каждого из направлений анализа, поспособствует очерчиванию круга типовых задач, потенциально интересных для решения формализованными методами. Также подобное руководство может быть тиражировано и адаптировано для решения широкого круга научных и практических задач.

Разработанность проблемы

Мы исходим из предположения, что определение типов и конкретных параметров методов тематической классификации текста зависит от задач исследования. Поэтому круг проанализированных в диссертации работ содержит публикации, посвященные как современным методам и алгоритмам тематической классификации текста, так и работы, связанные с содержательным фокусом исследования. В качестве такового был выбран образ Российской Федерации в одном из наиболее влиятельных американских и мировых изданий – «Нью-Йорк таймс». Интерес и актуальность изучения данного объекта обусловливаются важностью в информационном обществе образа страны для адекватного диалога между странами на различных уровнях.

В спектре современных методов анализа текста в гуманитарных науках можно выделить два основных подхода к тематической классификации текста – формализованный и эвристический (неформализованный, слабо формализованный).

Начало развития формализованных методов анализа текста в гуманитарных дисциплинах принято связывать с возникновением метода контент-анализа. Работы, нацеленные на количественное измерение параметров содержания текстов, тематическую классификацию газет появляются на рубеже

XIX—XX веков. В этом русле работали Г. Спид[3], М. Уилли[4], С. Кингсбери, Х. Харт и Л. Кларк, Дж. Вудворд. Методику анализа средств массовой информации, предложенную в своей работе М. Уилли, использовал советский исследователь общественного мнения и прессы В.А. Кузьмичев[5].

Стремительное распространение средств массовой информации, а также повышение актуальности изучения пропагандистских материалов обусловили необходимость разработки метода, позволяющего выявлять социальные цели текстов на основании количественного анализа эксплицированного содержания. Теоретической основой послужила классическая модель массовой коммуникации Г. Д. Лассвелла (*кто, что, по какому каналу, кому* говорит и *с каким эффектом*). На конференции по исследованию междисциплинарных средств массовой коммуникации в Чикаго в августе 1941 г. был предложен термин для нового метода – контент-анализ. Суть анализа в данный период заключалась в анализе знаков и утверждений с целью проверки их влияния на аудиторию; результатом анализа была частота определенных символов, их интенсивность и оценка отправителя. Среди видных исследователей следует назвать Б. Берельсона и П. Сальтера[6], Н. Лейтеса, И. Пула, И. Яниса, Р. Фаднера, А. Каплана, Дж. Голдсена, А. Геллера.

Разработке и апробированию методики статистического измерения интенсивности отношения коммуникатора к определенным объектам в тексте на основании лингвистически зафиксированных единиц посвящены работы Ч. Осгуда, Дж. Нанелли и С. Сапортой[7], Г. ван ден Верг и К. ван дер Виер.

В советской школе освоение контент-анализа как социологической методики проходило посредством анализа писем системы Гостелерадио, анализа текстов массовых газет, мониторинга телевизионных информационных про-

[3] Speed G. Do Newspapers Now Give the News? // The Forum. 1893. Vol. XV. P. 705-711
[4] Willey M. The Country Newspapers. Chapel Hill. N.C.: University of North Carolina Press. 1926
[5] Кузьмичев В.А. Печатная агитация и пропаганда. М., Л., 1930.
[6] Berelson B. and Salter P. Majority and Minority Americans // Public Opinion Quarterly. 1946. Vol.10. No.2. P. 168 - 190
[7] Osgood Ch., S. Saporta, J. Nunnally. Evaluative Assertion Analysis // Litera. 1956. Vol.3. P. 47-102

грамм. Данное направление представлено работами В. Шляпентоха, Б.А. Грушина, А.В. Жаворонкова, Л. Н. Федотовой[8], И.А. Красавченко, А.В. Баранова и многих других.

Развитие технологий производства и трансляции информации, широкая доступность данных в электронном виде определили потребность в разработке формализованных методов анализа больших массивов текстовых данных. Взрывной рост количества методов и алгоритмов формализованного анализа текстов приходится на рубеж XX—XXI веков. На сегодняшний день наиболее широкое распространение получил подход «мешок слов» (bag of words). Основной гипотезой данного подхода является предположение о том, что порядок следования слов в тексте не имеет значения для анализа, текст рассматривается как неупорядоченная совокупность слов (вектор, состоящий из частот слов). В упрощенном варианте каждое слово имеет равный «вес», порядок документов в корпусе (также как и порядок слов в тексте) не имеет значения, слова, встречающиеся часто в большом количестве текстов (так называемые стоп-слова) исключаются из анализа, разные формы слов считаются одним словом. Одно из первых упоминаний данного похода мы находим у З. Харриса[9].

В современной компьютерной лингвистике формализованные подходу к анализу текста, основанные на подходе «мешок слов», разрабатываются Д. Журавски и Дж. Мартином[10], К. Маннингом, П. Рагхаваном и Х. Шютце[11] и другими.

Разработке метода тематического моделирования, нацеленного на выявление латентных тем в корпусе текстов, посвящены работы Д. Блэя[12],

[8] Федотова Л.Н. Телемосты СССР-США: комплексное социологическое исследование. М., 1990. С. 16-17, 32-33.
[9] Harris Z. S. Distributional structure // Word. 1954. No. 10. P. 146-162.
[10] Jurafsky D., Martin J. Speech and natural language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Prentice Hall: Upper Saddle River, NJ, 2000.
[11] Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск.: Пер. с англ. М: ООО «И.Д. Вильямс, 2011.
[12] Blei. D. Probabilistic topic models // Communications of the ACM. 2012. Vol. 55. No. 4. P. 77–84

Д. Мимно[13], А. Дауда[14], М. Джордана, А. Энджи, Дж. Ли, Л. Жоу, Ф. Мухамеда, Е. Завитсаноса, Г. Палиураса, Г. Вуроса, Дж. Цанга, У. Сонга, С. Занга, Ш. Лью, К. В. Воронцова и А. А. Потапенко[15], С.В. Царькова и многих других.

Метод анализа тональности, призванный выявить эмоциональную «окраску» текста, разрабатывается в работах Б. Лью[16], Б. Панга, Л. Ли и С. Вайтинатан и других.

Очевидное преимущество данного подхода к анализу текста состоит в возможности обработки больших корпусов текстов. В целом, в современных условиях основным ограничением являются технические возможности компьютеров. Вторым преимуществом является объективность кодирования – на этапе обработки данных полностью исключено человеческое влияние, а, следовательно, риск субъективности и неустойчивости результатов. В качестве недостатков данного подхода следует отметить учет исключительно прямого значения слов, неразличение жанров, скрытых смыслов, коннотаций и пр. Также необходимо указать на технические сложности. Особенно явно проблемы проявляются при работе с русским языков, сложность которого обусловливает проблемы на этапе нормализации[17] (особенно лемматизации), учет синонимии, анафорических связей[18] и пр.). Одним из недостатков данного подхода является определение темы как совокупности слов в тексте, в то время как зачастую

[13] Mimno D., Blei D. Bayesian Checking for Topic Models // Empirical Methods in Natural Language Processing, 2011. P. 227-237.

[14] Daud A. Using Time Topic Modeling for Semantics-Based Dynamic Research Interest Finding // Knowledge-Based. 2012. Vol. 26, P. 154–163.

[15] Воронцов К. В., Потапенко А. А. Регуляризация, робастность и разреженность вероятностных тематических моделей // Компьютерные исследования и моделирование. 2012. Т. 4. № 12. Р. 693–706.

[16] Bing L. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers. 2012.

[17] Нормализация - приведение всех словоформ одного слова к единой основе. Целью нормализации является уменьшение количества уникальных слов, то есть снижение размерности текста. Существует два вида первого этапа нормализации текста: лемматизация (lemmatization) и стэмминг (stemming). Первый предполагает приведение каждого слова в словарную форму (лемму) (существительное – именительный падеж, единственное число, глагол – неопределенная форма и пр.), второй – приведение слова к его основе (морфеме) путем «обрезания» (окончаний, суффиксов и пр.), чтобы оставшаяся часть была одинаковой для всех грамматических форм. Лемматизация является более эффективной, так как использует словари и опирается на контекст, стэмминг намного более грубый алгоритм, но более быстрый. В нашей работе при проведении эмпирического анализа применялась лемматизация.

[18] Анафорические связи в тексте – отношения между частями текста (между словами, словосочетаниями, высказываниями), при которых в смысл одного слова (словосочетания, высказывания) входит отсылка к другому слову (словосочетанию, высказыванию).

семантика, наиболее точно описывающая проблему текста, не эксплицирована. Данный недостаток призван компенсировать альтернативный метод - эвристический тематический анализ.

Неформализованный, эвристический тематический анализ рассматривает текст как совокупность смыслов. Всякий текст трактуется как авторское описание и представление проблемы, реализуемое с помощью целенаправленного конструирования социальных смыслов. Исследователя интересует, скорее, выявление и толкование смыслов, явно и неявно транслируемых автором, интерпретация проблем, реконструкция позиций и типов аргументации, интерпретация авторского видения социальной реальности. Эвристический тематический анализ восходит к теории аргументации[19], основан на индуктивном подходе, который, в первую очередь, имеет описательный характер и поисковые задачи[20].

Эвристический тематический анализ требует активного участия и интерпретации со стороны исследователя. Он выходит за рамки подсчета слов или фраз и сосредоточивается на выявлении и описании явных и неявных идей в текстах, то есть тематической структуры текста. При проведении анализа разрабатываются коды - маркеры тем, используемые в дальнейшем анализе. В целом можно отметить наличие двух точек зрения на сущность тематического анализа. Ряд исследователей (Г. Гест, К. МакКуин, Е. Нэйми, В. Браун и В. Кларк[21]) полагают, что тематический анализ является интегральным методом: он включает в себя процедуры, заимствованные у обоснованной теории, дискурс-анализа и других методов. Метод перенимает преимущества у других методов из теоретического и методологического арсенала и адаптирует к прикладным исследованиям (автор настоящей работы придерживается данной точ-

[19] Attride-Stirling J. Thematic networks: an analytic tool for qualitative research // Qualitative Research. 2001. No. 1, P. 385-405.
[20] Guest G., MacQueen K., Namey E. Applied thematic analysis. Thousand Oaks, California: Sage. 2012.
[21] Braun V., Clarke V. Using thematic analysis in psychology // Qualitative Research in Psychology. 2006. Vol. 3. No. 2.

ки зрения). С другой стороны, существует точка зрения[22], что тематический анализ не является самодостаточным методом анализа данных, а, скорее, инструментом, который используется другими методами. В любом случае, данных подход достаточно распространен в гуманитарных науках (см., например, работы Д. Сингер и М. Хантер[23], Х. Рубин и И. Рубин[24], Г. Тейлор и Дж. Ашер[25]).

Исследования, посвященные проблеме формирования и восприятия образа Российской Федерации, были объединены в несколько направлений. Они сфокусированы на политической[26], географической[27], политико-психологической[28], маркетинговой[29], историко-культурной[30] или социологической[31] детерминанте в подходе к изучению образа Российской Федерации. Ни в одной из дисциплин исследователи не уделяют внимания обоснованию применения формализованных методов для эмпирического анализа образа Российской Федерации.

Проблема исследования заключается в несоответствии между потенциальными возможностями формализованных методов анализа текстовых данных и обоснованностью их использования. Отсутствует оценка целесообразности и эффективности интеграции формализованных методов с эвристическими методами анализа текстовых данных.

---

[22] Boyatzis R.E. Transforming qualitative information: thematic analysis and code development. Thousand Oaks, CA: Sage. 1998.
[23] Singer D., Hunter M. The experience of premature menopause: a thematic discourse analysis // Journal of Reproductive and Infant Psychology. 1999. Vol.17. No. 63. P. 63-81.
[24] Rubin H.J., Rubin, I.S. Qualitative interviewing: the art of hearing data. Thousand Oaks, CA: Sage. 1995.
[25] Taylor G.W., Ussher J.M. Making sense of S&M: a discourse analytic account // Sexualities. 2001. Vol. 4. No. 293. P. 293-314.
[26] Галумов Э.А. Имидж.против имиджа. М.: Известия, 2005.
[27] Замятин Д.Н. Метагеография: Пространство образов и образы пространства. М.: Аграф, 2004.
[28] Образы государств, наций и лидеров / Под ред. Е.Б. Шестопал. М.: Аспект Пресс, 2008.
[29] Панкрухин А.П. Маркетинг территорий. 2-е изд., дополн. Спб.: Питер, 2006.
[30] Федоров А.В. Трансформация образа России на западном экране: от эпохи идеологической конфронтации (1946–1991) до современного этапа (1992–2010). М.: Изд-во МОО «Информация для всех», 2010.
[31] «Рычащий медведь» на «диком Востоке» (Образы современной России в работах американских авторов: 1992-2007) / Сост. Э. Я. Баталов, В. Ю. Журавлева, К. В. Хозинская. М.: Российская политическая энциклопедия (РОССПЭН), 2009.

Теоретический объект исследования — смешанная (mixed), или интегративная, методология тематического анализа больших текстовых массивов.

Предмет исследования — конфигурация формализованных и эвристических методов на разных этапах реализации интегральной стратегии тематической классификации текста.

Конфигурация методов рассматривалась на примере текстового массива, репрезентирующего образ Российской Федерации. **Эмпирическим объектом** исследования явился корпус статей о Российской Федерации, опубликованных в «Нью-Йорк таймс» в период с августа 2011 г. по июль 2012 г.

Цель исследования – оценить относительную эффективность формализованных и эвристических методов на разных этапах реализации интегральной стратегии тематической классификации текста. В соответствии с поставленной целью, в работе последовательно решаются **следующие задачи**:

1) Дать систематическое описание основных методов тематической классификации текста;

2) Систематизировать применение основных подходов и методов тематической классификации текста применительно к кейсу исследований образа России;

3) Разработать и апробировать алгоритм тематической классификации текста в рамках стратегии смешивания формализованных и эвристических методов анализа текстов на примере репрезентаций образа России в «Нью-Йорк таймс» в период 2011–12 гг.;

4) Сравнить оценки свойств тематической структуры массива текстов, полученные альтернативными методами: формализованный тематический анализ (кластерный анализ, тематическое моделирование) vs. эвристический тематический анализ;

5) Сравнить оценки тональности массива текстов, полученные альтернативными методами: метод оценки тональности, основанный на обучении с учителем vs. эвристическое кодирование.

Методологические и теоретические основания исследования

Методология исследований с использованием смешанных методов описана в работах Дж. Брюэра и А. Хантера[32], Дж. Красвела[33], Дж. Грина, В. Карачелли, В. Грэхама[34], Р. Джонсона и Л. Кристенсена[35], И. Ньюмана и К. Бентц[36], А. Ташакорри и К. Тэдди[37]. Исследования Е. Кример и М. Гостон[38] демонстрируют возможности смешивания формализованных и эвристических методов при применении контент-анализа.

В более узком смысле теоретико-методологическую базу исследования составляют работы, посвященные основным подходам и алгоритмам методов тематической классификации текста, описывающие основные принципы и этапы применения методов тематической классификации текста. Формализованное направление представлено работами К. Криппендорфа[39], Р. Поппинга[40], К. Робертса[41], Дж. Гриммера[42], Б. Лью, Д. Журавски и Дж. Мартина, Д. Блэя, А. Дауда. Представление методов эвристического блока основано на работах Г. Геста, К. МакКуин, Е. Нэйми, В. Браун и В. Кларк.

---

[32] Brewer J., Hunter A. Multimethod research: A synthesis of styles. Newbury Park, CA: Sage, 1989.

[33] Creswell J. Research design: Qualitative, quantitative, and mixed approaches. Thousand Oaks, CA: Sage, 2003.

[34] Greene J., Caracelli V., Graham W. Toward a conceptual framework for mixed-method evaluation designs // Educational Evaluation and Policy Analysis. 1989. Vol. 11. P. 255-274.

[35] Johnson R., Christensen L. Educational research: Quantitative, qualitative, and mixed approaches. Boston, MA: Allyn and Bacon, 2004.

[36] Newman I., Benz C. Qualitative-quantitative research methodology: Exploring the interactive continuum. Carbondale, IL: Southern Illinois University Press, 1998.

[37] Tashakkori A., Teddlie C. (Eds.). Handbook of mixed methods in social and behavioral research. Thousand Oaks, CA: Sage, 2003.

[38] Creamer E., Ghoston M. Using a Mixed Methods Content Analysis to Analyze Mission Statements From Colleges of Engineering // Journal of Mixed Methods Research, 2013. P. 15-28.

[39] Krippendorff K. Content Analysis: An Introduction to Its Methodology. 2nd edition, Thousand Oaks, CA: Sage 2004.

[40] Popping R. Computer-assisted text analysis. London: SAGE Publications, 2000.

[41] Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences From Texts and Transcripts (Routledge Communication Series) ed. by Carl W. Roberts. Routledge, 1997.

[42] Grimmer J. 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.

Различение структурных элементов новостного сообщения как обособленных единиц анализа основано на подходе Т. Ван Дейка к новостям как особому типу дискурса[43].

Применительно к кейсу исследований образа России были проанализированы работы Э.А. Галумова, Д.Н. Замятина, Е.Б. Шестопал, Т.Н. Пищевой, Н.С. Виноградовой, А.Д. Недовой[44], С. Анхольта[45], А.П. Панкрухина, А.В. Федорова, В.И. Журавлевой[46].

Эмпирическая база исследования

Эмпирическую базу исследования составляет корпус статей «Нью-Йорк таймс» о России за период август 2011 – июль 2012 г. В данный промежуток времени уровень информационного внимания к событиям в России был достаточно высок, поскольку проходили думские и президентские выборы, а также был назначен новый состав кабинета министров. «Нью-Йорк таймс» был выбран потому, что данное издание традиционно является одной из предпочитаемых элитой газет, одной из наиболее цитируемых политиками, повестка «Нью-Йорк таймс» имеет значительное влияние на общественное мнение. Также данная газета содержит больший объем иностранных новостей, чем другие крупные американские газеты. Кроме того, что нетипично для американской прессы, «Нью-Йорк таймс» уделяет значительный объем печатных площадей иностранным корреспондентам и поэтому считается одной из наиболее независимых газет в США в сборе информации.

Отбор статей для анализа проходил в несколько этапов и был основан на различении релевантной[47] и пертинентной информации[48]. Финальный корпус

[43] Ван Дейк Т. А. Язык, познание, коммуникация. Б.: БГК им. И. А. Бодуэна де Куртенэ, 2000.
[44] Пищева Т.Н., Виноградова Н.С., Недова А.Д. Образ России под углом зрения политических коммуникаций // ПОЛИС. 2010. № 4, С. 107 – 121.
[45] Anholt S. Forward // Journal of Brand Management. 2002. Vol. 29. No. 4. P. 229-239.
[46] Журавлева В.И. Понимание России в США: образы и мифы. 1881-1914. М.: РГГУ, 2012.
[47] Релевантность информации – степень соответствия результатов поиска задаче, поставленной в запросе.
[48] Пертинентность информации – степень соответствия результатов поиска информационной потребности пользователя/исследователя.

статей для анализа составил 411 статей «Нью-Йорк таймс», посвященных России.

Научная новизна исследования заключается в следующем:

1. Описаны и систематизированы методы тематической классификации текста в рамках двух основных направлений: формализованного и эвристического. В рамках формализованного, подхода выделены два направления анализа: с известными априори категориями (кластеризация, метод анализа тональности, контент-анализ) и неизвестными категориями (тематическое моделирование). Проведенная систематизация демонстрирует методные альтернативы для решения типовых задач социального анализа, а также предлагает возможные стратегии алгоритмизации в рамках каждого из подходов.

2. Разработан, обоснован и апробирован алгоритм тематической классификации текста в рамках стратегии смешивания методов. Алгоритм включает поэтапное применение формализованных и эвристических методов тематической классификации текста: многоступенчатый отбор данных (основанный на различении релевантной и пертинентной информации), определение единиц анализа, контент-анализ, определение тональности заголовков, классификация заголовков, описание кластеров заголовков; контент-анализ, классификация, выделение основных тем текстов статей; описание и анализ каждой темы, индуктивное выведение интегрального образа.

3. Обосновано выделение контекстуальных факторов, учет которых необходим для изучения любого тематически выделенного корпуса текстов, и дано их модельное описание. В частности, обобщены подходы к изучению образа России в СМИ. По эпистемологическим и методологическим основаниям выделено шесть направлений в исследовании факторов формирования образа страны: политическое, географическое, психологическое, маркетинговое, историко-культурное и социологическое.

4. На примере сравнения и оценки качества результатов тематического анализа, проведенного альтернативными методами, показано, что применение эвристических процедур кодирования значительно улучшает качество полученных результатов. В качестве альтернативных способов решения задачи тематического анализа рассматривались следующие методы: кластерный анализ, тематическое моделирование, эвристический.

5. В качестве дополнительного результата эмпирического исследования показано, что при описании событий в России авторы «Нью-Йорк таймс» апеллируют к традиционным ценностям американского общества. По результатам проведения контент-анализа продемонстрировано, что в большинстве статей упоминается ценность «демократия и свободное предпринимательство».

Основные положения, выносимые на защиту

1. В работе с данными с многозначной операционализацией стратегия смешивания методов позволяет повысить качество (точность, правдоподобность, дифференцированность) результатов анализа.

2. Применение эвристического кодирования кратно повышает качество формализованного отбора в условиях использования простого поискового запроса.

3. По сравнению с формализованным методом анализа (реализованного методами кластерного анализа[49], тематического моделирования[50]) эвристическое кодирование дает более дифференцированную тематическую структуру заголовков статей.

4. Применение стратегии смешивания методов, то есть последовательное применение формализованных и эвристических методов, позволило перейти от неправдоподобно различных профилей к правдоподобно сходным

[49] Алгоритм двукластерного решения (bisecting k-means), косинусная мера. Использовалось программное обеспечение TLab.
[50] Алгоритм латентного размещения Дирихле. Использовалось программное обеспечение TLab.

профилям тематической структуры, полученных на основе анализа различных сегментов одних и тех же текстов.

5. Применение эвристического кодирования кратно повышает качество формализованной оценки тональности текста, реализованной методом обучения с учителем.

6. Тематическое моделирование имеет преимущество перед кластерным анализом в способности обнаруживать специфические смыслы, «невидимые» для кластерного анализа.

Теоретическая и практическая значимость работы

Полученные автором теоретические и методические результаты могут быть использованы представителями различных отраслей знания в теоретических и эмпирических исследованиях.

Во-первых, работа развивает методологию анализа текстовых данных. Находясь, по существу, в междисциплинарной зоне гуманитарных и точных наук, работа демонстрирует и подчеркивает взаимодополняющую, но не конкурирующую природу формализованных и эвристических методов анализа текстовых данных. В работе представлен, поэтапно описан и апробирован алгоритм отбора источников, выделения единиц анализа и обработки корпусов текстовых данных, основанный на интеграции различных методов анализа, который может быть использован в качестве методических рекомендаций при проведении эмпирических исследований.

Во-вторых, классификация подходов к анализу образа страны и предложенный способ определения и изучения образа страны в СМИ могут быть использованы для дальнейшего, более комплексного и глубокого изучения образа России. На основании полученных результатов могут быть сформулированы конкретные рекомендации по планированию и проведению кампаний по улучшению образа России, координации действий всех заинтересованных сторон: государства, СМИ, общественных объединений, бизнеса и пр.

Наконец, в педагогической сфере результаты диссертационного исследования могут быть использованы в рамках курсов по методологии анализа социологических данных, научно-исследовательских семинаров, а также могут стать основой специального учебного курса по методам анализа текстовых данных.

Апробация результатов

Основные положения диссертации были апробированы в научных публикациях автора, а также в докладах на X Конференции Европейской Социологической Ассоциации «Social Relations in turbulent times» (Женева, 2011), VIII Конференции по применению сетевого анализа (Цюрих, 2011), 6-й научно-практической Конференция памяти А.О. Крыштановского «Современная социология - современной России» (Москва, 2012), научном семинаре научно-учебной группы «Сетевые методы и модели в анализе текстовой информации» (Москва, 2012, 2013), VII Конференции памяти Юрия Левады «Современное российское общество и социология (Москва, 2013), Всероссийской научно-практической конференции Института социологии РАН «Модернизация отечественной системы управления: анализ тенденций и прогноз развития» (Москва, 2013). Диссертация была обсуждена на заседании кафедры методов сбора и анализа социологической информации факультета социологии НИУ ВШЭ.

Полученные в диссертации результаты встроены в процесс преподавания семинарских занятий по курсу «Социальные сети» (1 курс магистратуры, специализация «Прикладные методы социального анализа рынков»). По результатам исследования опубликованы 3 статьи в изданиях, рекомендованных ВАК Министерства образования и науки РФ.

Структура работы

Диссертация состоит из введения, трех глав, включающих 12 параграфов, заключения, библиографического списка и приложений. Общий объем работы

– 181 страница, в том числе, 2 приложения на 2 страницах, 17 страниц библиографии, 17 таблиц и 12 рисунков.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во Введении обосновывается актуальность проблемы исследования, определяется объект и предмет, формулируется цель и задачи диссертации, определяется ее научная новизна и практическая значимость, дается характеристика теоретико-методологических оснований, излагаются положения, выносимые на защиту.

Глава 1 «Подходы к проведению тематического анализа текстов» посвящена аналитическому обзору двух направлений тематического анализа – формализованному и эвристическому. Аналитическая роль автора проявляется в систематизации подходов, предложении интеграции двух подходов к анализу любого тематически выделенного корпуса текстов в рамках стратегии смешивания методов (mixed methods research).

Параграф 1 «Способы отбора источников» описывает возможные варианты формирования корпуса источников исследования. При работе с большими корпусами текстов начальный этап – отбор источников - исключительно важен, поскольку результат, полученный на данном шаге, предопределяет дальнейший ход исследования и валидность полученных результатов. Основные способы отбора текстовых источников соотносятся с типами выборок: случайный отбор, систематический отбор, стратифицированный отбор, вероятностный отбор, кластерный отбор, снежный ком, целевой отбор, сплошной отбор, удобный отбор.

Параграф 2 «Формализованный подход: кластерный анализ, тематическое моделирование» посвящен обзору формализованных методов анализа текста, решающих задачу выделения тематической структуры текстов. В начале параграфа приведена авторская классификация основных понятий, используе-

мых при проведении анализа текста – слово/словосочетание, концепт, категория и словарь.

Тематический анализ – это метод выявления, описания и анализа определенных образцов (patterns), тем в тексте. Браун В. и Кларк В. описывают тему как важную идею, содержащуюся в данных, имеющую непосредственное отношение к исследовательскому вопросу. Основной задачей формализованных методов является классификация - распределение текстов по категориям. Категории могут быть известны заранее – тогда задача состоит в распределении текстов по известным категориям, либо могут быть неизвестны до начала этапа анализа данных. В таком случае задача состоит в поиске латентных категорий, их описании и распределении текстов по выделенным категориям.

Формализованные методы тематического анализа, рассмотренные в работе, основаны на модели «мешка слов» (bag of words). Предварительной процедурой любого метода подготовительный этап (preprocessing)[51]. Результатом подготовительного этапа является представление каждого текста *i (i=1…N)* в виде вектора, состоящего из частот встречаемости в тексте каждого уникального слова *M:* $W_i = W_{i1}, W_{i2}, \ldots, W_{iM}$. Каждое значение W_{im} является частотой встречаемости *m-го* слова в *i-м* документе. Корпус текстов представляется в виде матрицы, содержащей векторное представление всех текстов корпуса[52] [53].

Методы классификации по известным категориям разделяются на методы, основанные на словарях и методы обучения с учителем. Первые используют частоту ключевых слов для отнесения документа в определенную категорию или для измерения степени, в которой документ принадлежит к той или иной категории, вторые решают проблему как задачу классификации текстов, где

[51] Этап включает нормализацию текста, удаление из текста «шумов» (слов, не несущих смысловой нагрузки, таких как предлоги, междометия и пр., а также редко встречающихся слов), приведение слов к одному регистру, удаление знаков пунктуации.

[52] Обычно ее называют матрицей термин-документ (term–documentmatrix или document-termmatrix).

[53] Возможно альтернативное представление – вместо частоты встречаемости для каждого слова рассчитывается коэффициент tf*idf (частота термина, умноженная на обратную частоту документов, где присутствует термин, в корпусе). Подобный подход позволяет дифференцировать слова по сравнительной значимости в корпусе.

классификаторы построены с использованием одного из методов машинного обучения и обучения на наборах данных. В работе из первого класса методов использован метод кластеризации, метод оценки тональности текста[54]. Второй класс методов представлен в работе методом тематического моделирования[55].

В *параграфе 3 «Эвристический подход: тематический анализ»* описаны онтологические корни эвристического тематического анализа, алгоритм его проведения, а также возможные проблемные точки реализации его на практике.

Эвристический тематический анализ восходит к теории аргументации, ориентируется на индуктивной подход, который имеет описательный характер и поисковые ориентации, требует активного участия и интерпретации со стороны исследователя. В целом можно отметить наличие двух точек зрения на сущность тематического анализа: согласно первой, тематический анализ является интегральным методом, заимствующим то, что считает наиболее полезным у других методов из теоретического и методологического лагеря, и адаптирующим к прикладным исследованиям. Согласно противоположному мнению, тематический анализ не является самодостаточным методом анализа данных, а, скорее, инструментом, который используется другими методами. Автор данной работы придерживается первой точки зрения. Схема проведения эвристического тематического анализа включает следующие этапы: знакомство с данными, создание исходных кодов, поиск тем, обзор (интерпретация, описание) тем, описание и лейбелинг (называние) тем, подготовка доклада.

Параграф 4 «Стратегия смешивания методов (mixed methods research)» посвящен описанию сравнительно молодого похода, представляющего собой особый тип исследований, где объединяются формализованные и эвристиче-

[54] Область исследований, которая анализирует мнения людей, настроения, оценки, отношения и эмоции по отношению к различным объектам, таким как товары, услуги, организации, частные лица, проблемы, события, темы и их атрибуты.

[55] Основная идея данного подхода состоит в том, что каждый текст рассматривается как совокупность распределений тем, наличествующих в корпусе. Каждая тема, в свою очередь, определяется вероятностным распределением на множестве слов. Наиболее распространенным является алгоритм латентного размещения Дирихле (Latent Dirichlet Allocation), где в качестве распределения используется функция Дирихле.

ские исследовательские техники, методы, подходы. Актуальность данного подхода обусловлена всевозрастающей степенью междисциплинарности, сложности и динамичности современных исследований, которые требуют развития соответствующей методологии. Целью смешивания методов является не замена данных подходов, а, скорее, использование их сильных сторон и минимизация недостатков. Философской платформой стратегии смешивания методов является классический прагматизм (представленный, в первую очередь, идеями Ч.С. Пирса, У. Джеймса и Дж. Дьюи).

В параграфе описаны преимущества и недостатки обоих подходов, предложены способы их объединения в рамках различных типов программ исследования. Большинство исследований, использующих смешанные методы, могут быть разделены на два типа: смешанные модели, смешиваемые методы (mixed models, mixing methods) (смешивание качественных и количественных подходов в рамках одного или нескольких этапов исследовательского процесса) и смешанные методы (mixed methods) (включение количественного и качественного этапов в целом в исследование).

Далее в параграфе описан алгоритм проведения исследования с использованием смешанных методов, включающий восемь этапов: определение вопроса исследования; определение релевантности использования стратегии смешивания методов задачам исследования и целей ее использования; определения типа стратегии: смешанные модели или смешанные методы; сбор данных; анализ данных; интерпретация данных; проверка результатов (legitimatethe data); подготовка итогового отчета.

Глава 2 «Теоретико-методологические основания изучения образа Российской Федерации в средствах массовой информации» посвящена комплексному описанию контекстуальных факторов, учет которых необходим для изучения любого тематически выделенного корпуса.

В параграфе 1 «Образ Российской Федерации как междисциплинарное понятие» рассмотрены основные направления изучения образа России в современной науке: политическое, политико-географическое, политико-психологическое, маркетинговое, историко-культурное и политико-социологическое, проанализированы их особенности и ограничения. На основании рассмотренных подходов и собственных наработок сформулировано авторское рабочее определение образа России в СМИ. Образ России – это интегральный конструкт, состоящий из совокупности транслируемых масс-медиа характеристик России. Эти характеристики вписаны в контекст миропонимания, задаваемого мировоззренческим паттерном американского общества, который проанализирован в следующем параграфе диссертации.

Каждый элемент образа соотносится с темой, в контексте которой описывается Россия. В конце параграфа представлена сводная таблица основных подходов к изучению образа России: понимание сущности образа, факторов его формирования, источников информации, методов изучения, выделены ограничения подходов.

Параграф 2 «Ключевые ценности американского общества» посвящен описанию ключевых ценностей американского общества как одной из детерминант формирования повестки СМИ и конструирования образа России. Одним из условий успешного воздействия на реципиента информации является обращение к его интересам, ценностям и идеалам. Ценностная ориентация средств массовой коммуникации связана также с процессом «gatekeeping»[56] - именно

[56] Gatekeeping или «эффект привратника» - процесс формирования повестки СМИ журналистами и редакторами путем «забраковывания» одних событий и выбора других. Подробнее см., например, Черных А. Социология массовых коммуникаций. Учебное пособие. Издательский дом ГУ ВШЭ: Москва, 2008.

ценностные фильтры становятся основанием для выбора событий, достойных попасть в повестку. В контексте нашей работы правомерно говорить о ключевых ценностях американского общества как одном из факторов, влияющих на формирование повестки «Нью-Йорк таймс». Хотя культура США не единообразна, американский социолог Р. Уильямс выделил 10 ценностей, которые широко распространены и, по признанию многих, составляют ядро американского общества: равные возможности, достижения и успех, материальный комфорт, активность и труд, практичность и эффективность, прогресс, наука, демократия и свободное предпринимательство, свобода, расизм и групповое превосходство. Позже были добавлены индивидуализм, гуманизм, образование, чрезмерная религиозность, романтическая любовь как основу брака, экология (эко-образ жизни), безопасность, здоровье и благополучие и технологии 21 века. Данная теоретическая рамка является основой изучения ценностей американского общества в материалах о России в последней главе диссертации.

В параграфе 3 «Новости как дискурс» в качестве теоретической основы изучения текста новости описывается общая форма дискурса новости, предложенная Т. Ван Дейком. Автор отмечает, что понятие «дискурс» является весьма расплывчатым и трудным для определения. Подобная комплексность ведет в многозначности и необходимости его определения в нескольких смыслах. Одним из смыслов является обозначение определенного жанра, например «новостной дискурс», «образовательный дискурс», «юридический дискурс»[57].

Для описания общей формы дискурса новостей Ван Дейк вводит понятия суперструктуры[58] и макроструктуры[59] текста новости. Данное означает, что

[57] Van Dijk T. Ideology: A Multidisciplinary Approach. London: Sage, 1998.

[58] Общие схемы построения текста новостей, содержащие основные категории и правила их следования. Явными категориями таких схем являются заголовок и вводка (обычно первое предложение или абзац новостного теста). Совокупность данных категорий образует краткое содержание новостной статьи.

[59] Макроструктуры текста разделяются на семантические и прагматические. Под семантической макроструктурой понимается тематическое содержание текста, его глобальная связность. Другими словами, это семантическое содержание категорий, входящий в суперструктурную схему. Прагматическая макроструктура – это последовательность речевых актов, связанная особыми макроправилами. Макроструктуры и суперструктуры представляют макросемантику и макросинтаксис текста соответственно.

текст новости содержит 2 взаимосвязанных элемента – композиционный и семантический. Каждый из них обладает собственными социально-коммуникативными функциями, строением, лексико-семантическими особенностями. Это обусловливает необходимость разделения текста новости на отдельные единицы анализа, подбор адекватных методов анализа для каждой. Данные положения фундируют выделение в эмпирической части работы заголовка и тела статьи в качестве самодостаточных единиц анализа.

Глава 3 «Описание и апробация алгоритма тематической классификации текста в рамках стратегии смешивания методов» описывает алгоритм проведения эмпирического анализа и результаты его апробации.

В параграфе 1 «Описание алгоритма проведения исследования» описана поэтапная схема реализации исследования: способ отбора источников, определения единиц анализа, последовательность применения и объединения методов анализа.

Целью исследования является выявление и описание структуры образа нашей страны, создаваемого СМИ США. В информационном обществе СМИ являются одним из важнейших общественных институтов, значимым каналом трансляции информации широким слоям населения. Данный институт функционирует не в социальном «вакууме», наравне с другими институтами СМИ подвержены влиянию государственных режимов, идеологий, общественных ценностей[60]. Модель журналистики США принадлежит так называемому либеральному (североатлантическому) полюсу. Его характеризуют предельная коммерциализированность, нацеленность на публику, посредническая роль СМИ между политическими элитами и гражданами[61].

На **первом этапе** для решения задачи исследования был выбран «Нью-Йорк таймс», традиционно одна из предпочитаемых элитой газет, одна из

[60] Siebert F., Peterson T., Schramm W. Four Theories of the Press. Urbana, University of Ilinois Press, 1963.
[61] Hallin D. C., Mancini, P. Comparing media systems: Three models of media and politics. Cambridge: Cambridge University Press, 2004.

наиболее цитируемых политиками, повестка которой имеет значительное влияние на общественное мнение. Согласно рабочему определению, образ понимается как интегральный конструкт, состоящий из совокупности характеристик России. Каждый элемент образа соотносится с темой, в контексте которой описывается Россия. Для индуктивного выделения и описания интегрального образа России необходимо определение тематической структуры, профилей изучаемых текстов, последующий анализ каждой из тем, сфокусированный на характеристиках России.

В ходе исследования нам необходимо также решить ряд методических задач. Во-первых, необходимо рассмотреть статью как отдельную единицу анализа, проанализировать ее структурные элементы, их социально-коммуникативные функции, определить адекватные методы анализа. Во-вторых, важной методической проблемой, возникающей у прикладного исследователя, является необходимость различения релевантной и пертинентной информации. Ошибки на данном этапе исследования могут привести к чрезмерно общим или очевидным выводам, а при пессимистическом сценарии – к ошибочным заключениям. Наконец, важно провести и сравнить возможности формализованного и неформализованного тематического анализа, определить роль эвристического кодирования. Современная интервенция формализованных методов обработки текста дает исследователям-эмпирикам множество поводов для беспокойства – от их технической сложности и трудоемкости до бессодержательности результатов, получаемых в случаях отсутствия эвристических процедур и этапа верификации.

На **втором этапе** с помощью многоступенчатого отбора был сформирован корпус статей: на первом этапе была сформирована сплошная выборка за период август 2011 – июль 2012 по ключевому слову «Russia»[62]. На следующих этапах, на основании различения релевантной и пертинентной информации

[62] Автор руководствовался предположением, что статья, посвященная России, содержит ключевое слово «Russia» как минимум один раз. Поиск осуществлялся с помощью информационной базы данных LexisNexis.

была сформирована итоговая выборка, объем составил 411 статей (20% релевантного корпуса, 80% исходной выборки оказалось «шумом» исследования). Полученный результат свидетельствует о необходимости при решении задач социального анализа верификации корпуса текстов исследования, отобранных с помощью применения формализованных методов. Показано, что эвристическое кодирование кратно повышает качество результатов.

На **третьем этапе,** на основании различения состава и функций частей новостного сообщения, корпус текстов статей был разделен на обособленные единицы анализа (заголовок, текст статьи). Следующие этапы исследования реализовывались параллельно для каждой из единиц анализа.

На **четвертом этапе** с целью определения интенционально сконструированных авторами макроструктур текстов новостей, был проведен контент-анализ заголовков. В результате были выявлены речевые показатели, маркирующие статью, как посвященную России, определены наиболее часто встречаемые из них. Данные показатели присутствуют в заголовках 85% корпуса (95% из них содержит слова «Russia» («Russian»), «Putin», «Moscow»)[63].

На **пятом этапе** для выявления тематической структуры заголовков был проведен тематический анализ. В качестве альтернатив были использованы кластерный анализ[64] и эвристический тематический анализ[65], использовалась апостериорная категоризация. Первым способом удалось выявить 4 темы, вторым – 21 тему, 4 наиболее часто встречаемые темы: экономическая политика, внешняя политика, выборы, протесты совпали. По итогам оценки качества результатов кластеризации[66] коэффициенты точности/полноты не превышают 55% (Таблица 1). Наилучшие результаты качества (по показателям точности и

[63] Результат свидетельствует также о том, что заголовки инвариантны относительно событий в России.
[64] Формализованный тематический анализ проводился методом кластерного анализа, алгоритм двукластерного решения (bisecting k-means), косинусная мера. Использовалось программное обеспечение TLab.
[65] Эвристический тематический анализ проводился согласно этапам, описанным в §3 главы 1.
[66] Для оценки качества кластеризации использовался метод внешнего сравнения, показатели точности (precision) и полноты (recall). Точность - это доля релевантных документов в корпусе. Полнота - это доля найденных релевантных документов среди всех релевантных.

полноты) получены в кластере «Внешняя политика»[67]. Верификация кластеров, выделенных формальным методом, показала гетерогенность их состава и непригодность для интерпретации. Также верификация позволила увеличить показатели качества для кластеров «Выборы» и «Протесты» путем взаимного изменения названий кластеров (Таблица 1, строки «после замены»).

Таблица 1. Показатели качества кластеризации

Кластер	Точность	Полнота
Экономическая политика	45%	17%
Внешняя политика	33%	30%
Выборы	14%	15%
Протесты	13%	16%
Выборы (после замены)	23%	55%
Протесты (после замены)	25%	40%

На **шестом этапе** был проведен анализ тональности заголовков. Тональность заголовка определялась альтернативными методами: методом обучения с учителем[68] и эвристическим кодированием. Результаты показали, что заголовки новостных материалов, посвященных России, в «Нью-Йорк таймс» носят, в большинстве случаев, негативную эмоциональную окраску, единственной темой, освещаемой исключительно положительно, является российская культура.

Результаты применения методов совпали в 54,6% случаях (Таблица 2).

Таблица 2. Сравнение результатов анализа тональности заголовков статей альтернативными способами

	Программное обеспечение, %		
Кодировщик, %		+	-
	+	3,9	8,3
	-	37,1	50,7

На наш взгляд, данный результат свидетельствует о невозможности применения автоматического метода определения тональности, основанного на обучении с учителем, без последующей верификации результатов.

[67] Результат объясняется лексическим однообразностью заголовков статей данной темы.

[68] Для автоматического определения тональности использовалось программное обеспечение Tweenator, основанное на методе обучения с учителем, уровень анализа - предложение.

На **седьмом этапе** был проведен формализованный[69] и эвристический тематический анализ текстов статей, использовалась апостериорная категоризация. По результатам кластеризации удалось выделить 2 наиболее часто встречаемые темы: внутренняя и внешняя политика, по результатам тематического моделирования – 8 тем (выборы, руководство, ресурсы, полиция, Путин, СССР, Сирия, культура). Кластеры в обоих случаях не интерпретируемы. Только с помощью тематического моделирования удалось выявить тему «культура», которая не фигурировала в предыдущих результатах. Данная тема оказалась очень важна для содержательных результатов исследования, поэтому результат свидетельствует о преимуществе тематического моделирования перед кластеризацией.

На **восьмом этапе** на основании выделенной тематической структуры были описаны элементы образа России.

Параграф 2 «Образ России как интегральное понятие» посвящен описанию интегрального образа России, созданного «Нью-Йорк таймс» в период думских и президентских выборов в России. Образ России в период август 2011-июль 2012 г. состоит из следующих элементов: характеристика внутренней политика, характеристика внешней политики, характеристика экономической политики, характеристика культуры. В тексте диссертации основные характеристики этих сторон жизни современной России, формирующих ее интегральный образ, проинтерпретированы в контексте паттерна ключевых ценностей американского образа жизни. Это позволяет перейти в следующем параграфе к роли этих ценностей в формировании образа «другого».

В параграфе 3 «Ценности американского общества в статьях «Нью-Йорк таймс» о России» показана роль традиционных ценностей американского общества в представлении материалов о России. Параграф был добавлен по

[69] Анализ текстов статей проводился альтернативными методами – кластерный анализ (параметры совпадают с параметрами при анализе заголовков) и тематическое моделирование (алгоритм латентного размещения Дирихле). Использовалось программное обеспечение TLab.

результатам эмпирического анализа и может служить примером, демонстрирующим один из возможных механизмов влияния на массовую аудиторию.

Для успешного коммуникативного воздействия на аудиторию необходимо опираться на разделяемые ею интересы, ценности, стереотипы. По результатам анализа корпуса текстов исследования у автора сформировалось предположение, что при описании событий в России авторы издания, как правило, апеллируют к ценностям американского общества[70].

Для ответа на вопрос об отражении ценностей американского общества в статьях «Нью-Йорк таймс» о России был проведен контент-анализ корпуса статей исследования. Было показано, что при описании каждого из элементов образа России автор статьи, как правило, апеллирует к той или иной ценности американского общества. Самой часто встречаемой ценностью является «демократия и свободное предпринимательство» (упоминается в 71,6% корпуса). В параграфе подробно описано распределение ценностей по отношению к темам. Основная проблема подобного взгляда на Россию, по мнению автора, заключается в том, что в нем происходит подмена целей и средств их достижения. Иными словами, терминальные ценности подменяются инструментальными, и их достижение провозглашается необходимым условием процветания российского общества

В заключении кратко обобщаются результаты диссертационного исследования, приводятся основные выводы и обозначается круг проблем для дальнейших исследований. К основным достижениям работы относится разработка и апробация алгоритма анализа корпуса текстовых данных в рамках стратегии смешивания методов, компенсирующего ограничения формализованного и

[70] На протяжении десятилетий многие ученые, политики и журналисты пытались обобщить и сформулировать основополагающие ценности американского общества. Признавая его исключительную разнородность, большинство исследователей сходится во мнении, что безоговорочно разделяемой практически всеми американцами является вера в либеральную демократию как наилучший государственный строй. Автор в качестве теоретической схемы ключевых ценностей американского общества использовал набор ценностей, выделенный Р. Уильямсом, описанный в §2 гл. 2. Преимущество данной теоретической схемы для эмпирического анализа состоит в конечности списка ценностей и операциональности определений.

эвристического подходов к анализу текста, определение места и роли эвристического кодирования на каждом этапе анализа. С содержательной точки зрения значимость представляет описание интегрального образа России в «Нью-Йорк таймс».

Работы, опубликованные автором в ведущих рецензируемых научных журналах и журналах, рекомендованных ВАКом Министерства образования и науки России:

Просянюк Д.В. Теоретико-методологические основания изучения образа России // Человек. Сообщество. Управление. 2012. № 4 . С. 32-47.

Просянюк Д.В. Образ России в призме социально-проектных и информационных технологий // Власть. 2014. № 1. С. 50-54.

Просянюк Д.В. Роль СМИ в формировании образа России // Проблемы теории и практики управления. 2014. № 3. С. 109-115.

Другие публикации:

Просянюк Д.В. Содержательные основания выделения границ Интернет-сетей // В кн.: Современная социология — современной России: Сборник статей VI международной научно-практической конференции памяти А.О. Крыштановского / Науч. ред.: А.Б. Гофман, Г.В. Градосельская, И.Ф. Девятко, Д.Х. Ибрагимова, И.М. Козина, Л.Я. Косалс, В.А. Мансуров, В.Г. Николаев, О.А. Оберемко, Н.Е. Покровский, Ю.Н. Толстова, А.Ю. Чепуренко, Е.Р. Ярская-Смирнова. М.: Издательский дом НИУ ВШЭ, 2012. С. 561-581.

Просянюк Д.В. Дискурс-анализ электронных СМИ с применением сетевого подхода (пример обсуждения вступления РФ во Всемирную торговую организацию) // В кн.: Социологические методы в современной исследовательской практике: Сборник статей, посвященный памяти первого декана факультета социологии НИУ ВШЭ А.О. Крыштановского [Электронный ресурс] / Отв. ред.: О.А. Оберемко. М.: Издательский дом НИУ ВШЭ, 2011. С. 352-357.