МЕТОДЫ И МЕТОДОЛОГИЯ

DOI: 10.14515/monitoring.2017.5.05

Правильная ссылка на статью:
Дудина В. И., Юдина Д. И. Извлекая мнения из сети Интернет: могут ли методы анализа текстов заменить опросы общественного мнения? // Мониторинг общественного мнения: Экономические и социальные перемены. 2017. № 5. С. 63—78. DOI: 10.14515/monitoring.2017.5.05.
For citation:
Dudina V. I., Iudina D. I. Mining opinions on the Internet: can the text analysis methods replace public opinion polls?. *Monitoring of Public Opinion: Economic and Social Changes.* 2017. № 5. P. 63—78. DOI: 10.14515/monitoring.2017.5.05.

В. И. Дудина, Д. И. Юдина

ИЗВЛЕКАЯ МНЕНИЯ ИЗ СЕТИ ИНТЕРНЕТ: МОГУТ ЛИ МЕТОДЫ АНАЛИЗА ТЕКСТОВ ЗАМЕНИТЬ ОПРОСЫ ОБЩЕСТВЕННОГО МНЕНИЯ?

ИЗВЛЕКАЯ МНЕНИЯ ИЗ СЕТИ ИНТЕРНЕТ: МОГУТ ЛИ МЕТОДЫ АНАЛИЗА ТЕКСТОВ ЗАМЕНИТЬ ОПРОСЫ ОБЩЕСТВЕННОГО МНЕНИЯ?

MINING OPINIONS ON THE INTERNET: CAN THE TEXT ANALYSIS METHODS REPLACE PUBLIC OPINION POLLS?

ДУДИНА Виктория Ивановна — кандидат социологических наук, доцент кафедры прикладной и отраслевой социологии Санкт-Петербургского государственного университета, Санкт-Петербург, Россия.
E-MAIL: viktoria_dudina@mail.ru
ORCID: 0000-0003-2099-2345

Victoria I. DUDINA[1] — Cand. Sci (Soc.), Associate Professor
E-MAIL: viktoria_dudina@mail.ru
ORCID: 0000-0003-2099-2345

ЮДИНА Дарья Игоревна — социолог Центра социологических и Интернет-исследований Санкт-Петербургского государственного университета, Санкт-Петербург, Россия.
E-MAIL: dartisimus@gmail.com
ORCID: 0000-0002-6603-0697

Daria I. IUDINA[1] — Sociologist
E-MAIL: dartisimus@gmail.com

[1] St Petersburg State University, St Petersburg, Russia

Аннотация. Пользовательский контент интернет-ресурсов становится альтер-

Abstract. User-generated content becomes an alternative or supplementary

нативным или дополнительным источником изучения общественного мнения. Тем не менее остается не до конца проясненным вопрос, могут ли эти данные при современных методах анализа полностью или частично заменить массовые опросы. Цель статьи — показать на примере анализа больших текстовых данных методологические возможности для такой замены методами тематического моделирования и алгоритма по извлечению мнений. В статье проводится сравнение анализа комментариев пользователей видеохостинга Youtube к фильму «Чайка» Фонда борьбы с коррупцией с результатами опроса общественного мнения по поводу отношения к фильму и представленным в нем фактам. Результаты исследования показывают, что анализ мнений в Интернете пока не может полностью заместить массовые опросы, но предоставляет более широкий контекст для интерпретации мнений и их более детальной оценки, а также может быть использован для улучшения структуры анкеты и качества самих вопросов.

Ключевые слова: тематическое моделирование, извлечение мнений, общественное мнение, анализ текста, Интернет

source for public opinion studies. Nevertheless, the question whether modern methods of data processing can completely or partially replace opinion surveys remains unclear. The purpose of this article is to show methodological possibilities of topic modeling and opinion mining based on large text dataset analysis. The article provides a comparison between the analysis of YouTube user comments and the results of the public opinion survey devoted to the Chaika documentary produced by the Anti-Corruption Foundation. The study reveals that the analysis of opinions collected from the Internet cannot completely replace opinion surveys; however, it provides a broader context for interpreting the opinions and their more detailed assessments. Additionally, it can be used to improve the questionnaire structure and the quality of questions.

Keywords: topic modeling, opinion mining, public opinion, text analysis, Internet resources

Под влиянием развития цифровых технологий изменяется логика исследовательского процесса во многих областях науки. В социологии вместо использования опросных данных появляется возможность исследовать так называемые цифровые следы — от данных, регистрирующих простейшие факты поведения, до развернутых комментариев по определенным темам, которые можно найти в социальных медиа. Одним из заманчивых обещаний развития новых технологий и социальных медиа становится перспектива безопросного сбора данных о поведении и мнениях людей. Возможность измерять общественное мнение, не задавая вопросов, в перспективе может способствовать существенному снижению стоимости полевого этапа работы, улучшению качества аналитики и оперативному

отслеживанию тенденций в общественном мнении в режиме реального времени. Исследуя общественное мнение с помощью Twitter, Э. Коди с коллегами предложили понятие «unsolicited public opinion» (незапрошенное общественное мнение) [Cody et al., 2015; Cody et al., 2016]. И хотя авторы этого терминологического новшества являются математиками и специалистами в области компьютерных наук, а не социологами, тем не менее, подобное определение полезно и для социологов, поскольку позволяет отделить феномен общественного мнения, которое стихийно формируется и «извлекается» из сети как цифровой след, от общественного мнения, конструируемого в результате массового опроса.

Хотя измерение общественного мнения через анализ информации из социальных сетей, блогов и других социальных медиа постепенно становится все более распространенной практикой полстерских служб и компаний[1], методологических работ, описывающих не просто процесс сбора и анализа такой информации, но и сравнение полученных данных с результатами опросных методов, на данный момент очень мало. В статье [O'Connor et al., 2010] данные опросов различных полстерских организаций о доверии потребителей и политических установках за 2008 и 2009 гг. сравниваются с тональной оценкой аспектов, связанных с именем кандидата, полученной на основе анализа миллиарда англоязычных сообщений в Twitter за эти же временные интервалы. В нескольких случаях обнаруживается достаточно высокая корреляция. В статье [Cody et al., 2016] демонстрируется, что результаты тонального анализа текстов из Twitter хорошо коррелируют с рядом традиционных показателей и обладают определенной предсказательной силой. Авторы некоторых исследований сопоставляют результаты анализа социальных медиа не с данными опросов, а с реальным поведением на основании, например, результатов выборов. В статье [DiGrazia et al., 2013] ставится вопрос о том, насколько по данным из социальных медиа можно судить об офлайн-поведении и, в частности, могут ли данные социальных медиа рассматриваться в качестве количественного индикатора политического поведения. Авторы исследования выявляют статистически значимую ассоциацию между упоминанием имени кандидата в Twitter и данными Федеральной избирательной комиссии США. В статье [Tumasjan et al., 2010] ставится задача на примере федеральных выборов в Германии выяснить, насколько адекватно обсуждения в социальных медиа отражают офлайновые политические установки. На основании анализа тематической и психолингвистической ориентированности 100 тысяч сообщений в Twitter, содержащих отсылки к конкретной политической партии или политику, авторы делают вывод, что количество упоминаний партии коррелирует с ее результатом на выборах, а совместное упоминание нескольких партий отражает реальные политические связи и коалиции. Как видно из приведенных примеров, значительная часть таких работ связана с анализом текстов о кандидатах на политические должности в предвыборный период, а сравнение с результатами опросов происходит по простой схеме: количество упоминаний

[1] См. Медиалогия [Электронный ресурс]. URL: http://www.mlg.ru/about/ (дата обращения: 19.07.2017); Анализ информации из соцсетей [Электронный ресурс] // Ipsos Comcon. URL: https://www.ipsos.com/ipsos-comcon/ru-ru/analiz-informacii-iz-socsetei?language_content_entity=ru-ru (дата обращения: 19.07.2017).

и их тональная оценка из текстов социальных медиа сравниваются с оценкой и количеством собирающихся голосовать из опросов или с результатами выборов.

Несмотря на немногочисленность работ, авторы которых сравнивают результаты анализа мнений из социальных медиа и социологических опросов, сама методология извлечения мнений из неструктурированных текстов довольно разнообразна. Существует несколько обзоров, классифицирующих подходы к анализу текста, например [Pang, Lee, 2008; Tang et al., 2009; Ravi, Ravi, 2015]. В настоящее время можно выделить два основных подхода к решению задачи извлечения мнений: машинное обучение и лингвистические методы. Примерами использования машинного обучения являются тематическое моделирование [Lin et al., 2012; Rill et al., 2014] и различные методы обучения с учителем [Cambria et al., 2015; Coussement, Van den Poel, 2009]. Лингвистический подход включает в себя использование словарей [Tumasjan et al., 2010] и различных лингвистических парсеров [Gimpel et al., 2011; Poria et al., 2014]. Разработаны также гибридные методы, использующие оба этих подхода [Prabowo, Thelwall, 2009; Jiang et al., 2011]. К. Рави и В. Рави [Ravi, Ravi, 2015] выделили основные проблемы, с которыми сталкиваются исследователи при получении мнений из онлайн-источников. Помимо непосредственной задачи определения мнений возникает проблема способов разрешения неопределенности в случаях, когда смысл слов и фраз отличается от буквально понимаемого, например, когда в них присутствует сарказм или ирония [Juusto et al., 2014; Reyes, Rosso, 2012]. Другой трудностью, предполагающей необходимость дополнительной чистки данных, является активность спамеров и ботов [Hu et al., 2011; Forelle et al., 2015]. Отдельно можно выделить проблему дефицита или отсутствия инструментов для анализа мнений в других языках, кроме английского.

Таким образом, несмотря на то, что методы извлечения мнений находятся в стадии разработки и сталкиваются с различными сложностями, те перспективы, которые предлагает эта область для развития изучения общественного мнения, делает задачу оценки качества данных, получаемых в результате извлечения мнений из комментариев пользователей сети Интернет, очень актуальной. Традиционные опросы общественного мнения обладают хорошо разработанной методологией, а данные, получаемые с их помощью, репрезентативны и хорошо структурированы. Однако высокая затратность и интертность таких методов, влияние разного рода смещающих факторов, связанных, в частности, с организацией полевой работы, заставляют социологов развивать методы безопросного извлечения мнений. С развитием социальных медиа для исследователей открывается возможность изучения общественного мнения безопросным путем через разработку методов анализа больших массивов текстовых данных [Cody et al., 2016: 1] как замену и/или дополнение для традиционных опросов общественного мнения. В этом контексте задача сравнения данных, получаемых из социальных медиа с данными, получаемыми в ходе опросов общественного мнения, приобретает особую актуальность.

Исходя из предположения, что результаты, получаемые при анализе обсуждений на интернет-ресурсах, могут при определенных условиях стать альтернативой или дополнением для традиционных опросов, в исследовании была поставлена цель — сравнить результаты извлечения мнений из комментариев пользовате-

лей сети Интернет с результатами массового опроса по сходной проблематике. При этом ставилась задача обратиться к интернет-ресурсу, отличному от Twitter, который широко используется в подобных исследованиях, и рассмотреть кейс, связанный с реакцией общественного мнения на какое-то резонансное событие. Дополнительная задача данного исследования состояла в том, чтобы предложить варианты восполнения дефицита инструментария анализа мнений для русского языка. Важным критерием выбора кейса для анализа была актуальность и резонансность проблематики, которая обусловливала бы и наличие достаточного количества комментариев в сети Интернет, и опроса общественного мнения по данной проблеме, примерено совпадающего по времени проведения. В качестве резонансного события, послужившего кейсом для анализа реакции общественного мнения, в данном исследовании был выбран фильм «Чайка» Фонда борьбы с коррупцией, размещенный на видеохостинге Youtube[2]. К 28 декабря 2015 г. на этом интернет-ресурсе было оставлено около 19 тыс. постов и комментариев пользователей по поводу фильма. Также в декабре 2015 г. Левада-Центр провел опрос по поводу данного фильма «по репрезентативной всероссийской выборке городского и сельского населения среди 1600 чел. в возрасте 18 лет и старше в 137 населенных пунктах 48 регионов страны»[3] с целью выяснить реакцию общественного мнения на этот резонансный фильм.

Сбор и частично анализ данных из сети Интернет осуществлялся доступными статистическими и математико-лингвистическими методами, позволяющими хотя бы частично заменить работу аналитика. Такой выбор инструментария был обусловлен задачей обработки больших объемов неструктурированных текстов, в виде которых, как правило, представлены комментарии и обсуждения в социальных медиа.

Для достижения поставленной цели было скачано 2907 самых популярных, то есть получивших наибольшее число «лайков», комментариев из 19 тыс. комментариев, оставленных пользователями к моменту закачки данных (28.12.2015). Данная выборка текстов составила приблизительно 15% от всего объема совокупности постов и комментариев, размещенных пользователями Youtube с 1.12.2015 по 28.12.2015[4]. Подробно сбор и препроцессинг этих данных описаны в статье [Юдина, Дудина, 2016]. Полученный массив комментариев не анализировался на предмет наличия в нем подозрительный бот-активности, так как на платформе Youtube пока отсутствуют доступные инструменты такого анализа. Анализ полученных текстов был проведен при помощи тематического моделирования и анализато-

[2] Фонд борьбы с коррупцией. «Чайка». [Электронный ресурс]. URL: https://www.youtube.com/watch?v=eXYQbgvzxdM (дата обращения: 25.31.2015)

[3] Результаты опроса о реакции общества на фильм «Фонда Борьбы с Коррупцией» [Электронный ресурс] // Левада-Центр. URL: http://www.levada.ru/2015/12/23/reaktsiya-obshhestva-na-film-fonda-borby-s-korruptsiej/ (дата обращения: 18.05.2016)

[4] Поскольку API Youtube не позволяет закачивать более ста комментариев (см. YouTube Data API [Электронный ресурс]. URL: https://developers.google.com/youtube/v3/docs/comments/list#try-it (дата обращения: 25.31.2015), был использован специальный скрипт, скачивающий комментарии путем автоматического повторения действий, которые надо было бы предпринимать для ручной закачки (см. Bouman E. Youtube Comment Downloader [Электронный ресурс]. URL: https://github.com/egbertbouman/youtube-comment-downloader (дата обращения: 28.31.2015). Тем не менее этой мини-программе не удалось скачать все комментарии, видимо, потому что сервер Youtube настроен прерывать подобные «подозрительные» запросы.

ра, используемого для извлечения мнений. Оба подхода: выявление тематической структуры и извлечение мнений (opinion mining),— являются на сегодняшний день наиболее популярными способами работы с большими объемами текстовых данных.

Тематическое моделирование: оценка соответствия структуры анкеты темам, обсуждаемым пользователями интернет-ресурса

Для работы с тематическим моделированием была выбрана библиотека «topicmodels» [Grün, Hornik, 2011] языка R. Анализ результатов тематического моделирования проводился последовательно с двумя моделями, получившими название «базовая» и «расширенная». Для построения базовой тематической модели был выбран метод на основе обычного размещения Дирихле с автоматически высчитываемыми начальными значениями гиперпараметров α и β и выборкой Гиббса в качестве алгоритма оптимизации параметров модели. Для нахождения оптимального числа тем использовался показатель внутренней валидности — среднее гармоническое правдоподобие. Базовая тематическая модель, получение которой более подробно описано в статье [Юдина, Дудина, 2016], позволила выделить в общем объеме скачанных комментариев к фильму «Чайка» десять основных тем обсуждения: влияние нефтяных цен на уровень жизни; доказательства связи генпрокурора с бандой Цапка и преступлениями; ответственность президента за существование коррупции; последствия расследования для генпрокурора и авторов фильма; роль и образ Навального; роль США и России в войне на Украине; сомнения по поводу доказательств, представленных в фильме; сравнение государственного устройства и жизни в западных странах и в России; обсуждение основных фактов, представленных в фильме; отношение государства к народу. Поскольку особенностью опросов общественного мнения является возможность оценить количественное распространение мнений в исследуемой совокупности, то для сравнения результатов тематического моделирования с результатами опроса недостаточно просто идентифицировать темы, но необходимо каким-то образом квантифицировать результаты тематического моделирования. В данном исследовании такая квантификация была осуществлена через расчет относительной доли каждой темы в общем объеме дискуссии. Для этого вероятность принадлежности каждого слова к теме умножалась на частоту этого слова в корпусе текстов, а затем полученные результаты суммировались по каждой теме. Доля каждой темы представлена на рис. 1. Хотя в данном случае мы основывались на предположении, что чем выше удельный вес темы, тем больший резонанс она вызывает у участников обсуждения, в таком виде результаты, полученные при помощи тематической модели, достаточно сложно сравнивать с результатами опроса. В отличие от опросных данных, обсуждение в социальных сетях не репрезентативно относительно всего населения и в нем отсутствует стандартизация, накладываемая структурой анкеты.

Одним из возможных способов анализа при данных условиях было выбрано сравнение доли тем с распределением ответов на некоторые вопросы анкеты Левада-Центра и определение того, насколько включенные в анкету вопросы и варианты ответов соответствуют «нереактивным», то есть не наведенным анкетными вопросами, мнениям людей, оставивших комментарии к видео.

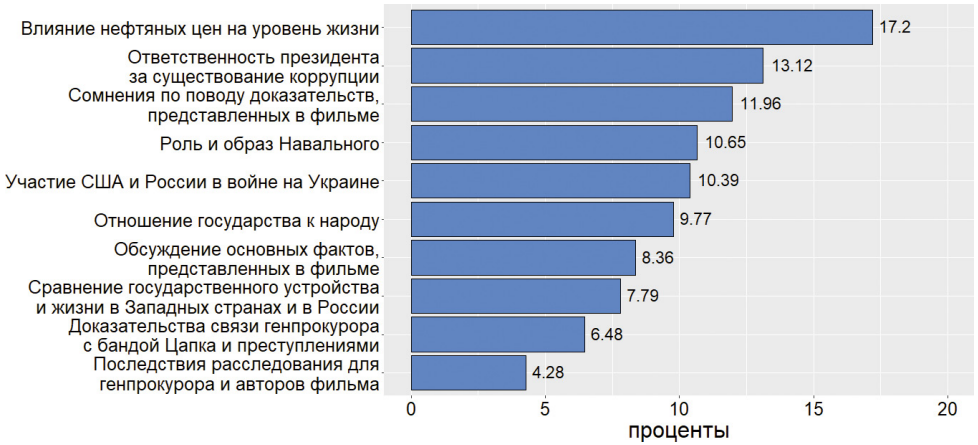

Рисунок 1. Доля тем в комментариях к фильму «Чайка» в тематической модели

В результатах опроса Левада-Центра по поводу фильма «Чайка», наиболее популярным ответом на вопрос «С каким из суждений об этих обвинениях в адрес семьи генерального прокурора России Юрия Чайки Вы бы скорее согласились?» стал вариант «Это похоже на правду, хотя достоверность этих обвинений оценить сложно» (39 % от числа тех респондентов, кто хотя бы слышал о фильме «Чайка»). Если сравнить этот вариант ответа с распределением доли тем, то видно, что созвучная ему тема «Сомнения по поводу доказательств, представленных в фильме», также популярна среди остальных тем (11,96 % — третья по популярности тема). Это же сходство касается и ответа на вопрос «Как Вы считаете, коррупционные схемы и связи с преступными группировками, в которые, согласно расследованию Фонда по борьбе с коррупцией, втянуты сыновья Юрия Чайки, это единичный случай или типичное явление?», в котором 82 % респондентов, считающих выводы фильма правдоподобными, выбрали вариант «Типичное явление, проявление разложения российской власти». Такой ответ в значительной степени соответствует второй по популярности теме (13,12 %) «Ответственность президента за существование коррупции». Примечательно, что тема получилась персонифицированной, и это говорит о том, что в восприятии участников дискуссии российская власть ассоциируется только с одним человеком — с президентом. Этот аспект не «ухватывается» опросом, но выявляется при анализе комментариев.

Можно сказать, что структура анкеты во многом совпадает со структурой тем, выявляемых при анализе комментариев пользователей. Почти все темы, которые относятся к обсуждению и оценке самого фильма, присутствующие в анкете, имеют место и в дискуссии пользователей. В то же время анализ дискуссии позволяет выявить более широкий спектр тематик, которые не учитывались структурой анкеты. Кроме вопроса об оценке коррупции *(«Как Вы считаете, коррупционные схемы и связи с преступными группировками, в которые, согласно расследованию Фонда по борьбе с коррупцией, втянуты сыновья Юрия Чайки, это единичный случай или типичное явление?)*, темы, составляющие контекст обсуждения, то есть напрямую

не связанные с фактами, представленными в фильме, в анкете отсутствуют. Это касается таких аспектов, выявленных при анализе дискуссии пользователей, как обсуждение личности Навального, отношение к событиям на Украине, оценка отношения государства к народу, сравнение России и Запада. Безусловно, в масштабных опросах на актуальные темы допустимый объем анкеты ограничивает тематику задаваемых вопросов и требует от социологов придерживаться при создании анкеты определенной логики. Любое событие (в нашем случае, фильм) воспринимается и оценивается людьми в определенном контексте и вызывает определенные коннотации, без понимания которых часто бывает сложно интерпретировать полученные распределения ответов на анкетные вопросы. Проведенное сравнение показывает, что анализ сетевых дискуссий, предшествующий массовому опросу или дополняющий его, содержит потенциал для улучшения опросного инструментария как за счет включения вопросов, которые более релевантны представлениям опрашиваемых, так и благодаря обеспечению контекста для интерпретации результатов опроса.

Извлечение мнений: компоненты алгоритма и область применения

Алгоритмы по извлечению мнений из текста, в отличие от тематических моделей, анализируют связи слов в предложении и/или в тексте в целом. Этот анализ необходим для того, чтобы определять слова, связанные с теми словами, по поводу которых необходимо получить мнение. Более продвинутые анализаторы также используют определение синтаксической структуры предложений для того, чтобы определить контекст, эмоциональную нагрузку слов.

В качестве примера того, как можно анализировать текст, используя один из таких алгоритмов, рассмотрим способ извлечения мнений о конкретных объектах внимания в комментариях к фильму «Чайка». Такими объектами были выбраны слова «Путин» и «Навальный», потому что эти персоны получили со стороны участников обсуждения наибольшее внимание, сопоставимое или даже превышающее внимание непосредственно к главным «героям» фильма — прокурору Чайке и его сыновьям. Частота упоминаний важна в данном случае, поскольку дает возможность собрать больше разнообразных мнений о каждом из изучаемых объектов.

Для получения мнений мы воспользовались двумя инструментами: специальным парсером (синоним слову анализатор), отбирающим слова, связанные с исследуемыми, и алгоритмом для тонального анализа, с помощью которого полученным словам дается эмоциональная оценка. Наиболее точные инструменты в этом анализе созданы на основе синтаксических и контекстных анализаторов. Для определения связи со словом и его эмоциональной характеристики в них используются алгоритмы, учитывающие связь слов в предложении, а также присутствие других слов и выражений, формирующих определенный контекст, необходимый для точного определения эмоциональной нагрузки слова. К сожалению, для русского языка доступных программных продуктов такого уровня, позволяющих анализировать значительное количество предложений в единицу времени, найти не удалось, поэтому были использованы более простые средства. Они дали не безошибочные, но достаточные для сравнения и интерпретации результаты.

Лингвистический анализатор, который удалось найти для задачи определения слов, связанных с целевыми словами (понятиями, категориями), использует мор-

фологическую связь между словами, находящимися в предложении рядом друг с другом, либо в соответствии с определенной схемой[5]. Такая схема называется грамматикой, а сами целевые понятия представлены словарем, содержащим в себе список слов, соответствующих этому понятию. Грамматика представляет собой последовательность из понятия/понятий из словаря и различных обозначений частей речи, их форм или регулярных выражений. Анализатор сравнивает текст с грамматикой и отбирает слова и выражения, соответствующие паттерну (грамматике). Данный алгоритм основан на теории контекстно-свободных грамматик (context-free grammar) из компьютерной лингвистики [Berstel, Boasson, 1990].

Грамматики для нашей задачи разрабатывались по следующей схеме: справа или слева от целевого понятия слово, представляющее собой глагол, существительное, прилагательное или причастие (последние два — в одном числе и падеже с целевым понятием); если перед словом стоит частица «не», то она также учитывается. Частица учитывается для тонального анализа, так как этот этап реализуется при помощи словаря, а частица «не» меняет эмоциональную нагрузку слова на противоположную. Можно было бы создать более сложные грамматики, например, те, которые учитывают различные конструкции с предлогами или союзами, но это вряд ли дало бы значительно больше результатов, поскольку чем сложнее грамматическая конструкция, тем она реже встречается, и вместе с этим такие грамматики значительно увеличили бы время работы самого алгоритма, и так достаточно медленного по своей архитектуре.

Существенным недостатком этого лингвистического анализатора является то, что в нем не снята морфологическая неоднозначность. Морфологическая неоднозначность обозначает ситуацию, когда одна словоформа соответствует разным по смыслу словам, и разрешить эту неоднозначность можно только при помощи синтаксического контекста. Например, в выражении «… рожу Навального …» не ясно, собирается ли комментатор родить Навального, либо он грубо высказывается о его лице. Используемый парсер в этом случае просто выбирает первую граммему (часть речи), которую предлагает библиотека pymorphy2.

Полученные от парсера слова нормализуются, а затем оцениваются тональным анализом. В нашем распоряжении был словарь тональностей, разработанный НИУ «Высшая школа экономики»[6]. Словарь содержит 6860 уникальных слов, каждое представлено в словаре не менее трех раз с определенной оценкой тональности от –2 (негативная окраска слова) до 2 (позитивная окраска слова), ноль означает нейтральную окраску. Эта тональность присуждена словам на основе контекста отрывков, в которых они присутствовали. Разметка словаря проводилась на краудсорсинговой платформе, то есть эту работу осуществляли добровольцы, а не эксперты. Для предотвращения существенных ошибок или недобросовестной работы их результаты контролировались и проверялись. Недостаток словаря в том, что в нем отсутствует частица «не» перед словом в тексте. Отсутствие этой информации, безусловно, повлияло на качество анализа. Так как в тональном

[5] GLRParser [Электронный ресурс]. URL: https://github.com/vas3k/python-glr-parser (дата обращения: 19.04.2017).

[6] Лаборатория Интернет-исследований ВШЭ. Тональный словарь [Электронный ресурс]. URL: http://linis-crowd.org/ (дата обращения: 22.03.2017).

словаре несколько оценок одного слова и контекст употребления их не известен, то для нашего анализа использовалась средняя оценка слова.

Полностью алгоритм извлечения мнений имеет следующую структуру: каждому слову, выделенному лингвистическом анализатором и найденному в словаре, присваивается среднее арифметическое оценок тональности слова. Частица «не» перед словом меняет знак тональности на противоположный. В результате понятие «Навальный» оценивается 99 словами, 61 из которых уникальное, максимально негативная оценка слова, используемая в отношении «Навального»,—1.67, максимально позитивная — 1.33, средняя оценка –0.15, распределение оценок тональностей представлено на рис. 2. Понятие «Путин» оценивается 94 словами, из которых 70 уникальные, максимально негативная оценка –1.75, максимально позитивная — 1.17, средняя оценка –0.25, распределение оценок тональностей представлено на рис. 3.

Как видно из представленных диаграмм (рис. 2 и 3), несмотря на то, что доли положительных оценок не сильно различаются (17.17% «Навальный» и 15.96% — «Путин») и отличия в долях негативных оценок также не критичны (38.38% «Навальный» и 44.68% — «Путин»), доля негативных оценок тональности слов превысила долю нейтральных оценок в отношении понятия «Путин», что может свидетельствовать о более эмоциональном восприятии обсуждающими фигуры действующего президента, чем оппозиционера А. Навального. В случае распределения оценок тональности по отношению к А. Навальному неожиданным оказался перевес негативных оценок над позитивными. Если исходить из предположения, что люди, смотревшие фильм и оставлявшие комментарии, в большинстве своем критически оценивают существующую власть, которая, как показало тематическое моделирование, персонифицирована в фигуре действующего президента, то, скорее всего, они должны бы более позитивно воспринимать любую оппозицию. Однако этого не происходит.

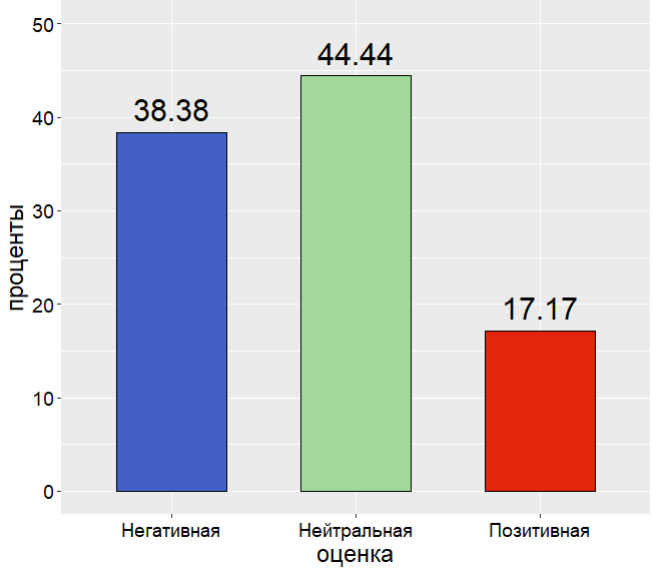

Рисунок 2. Распределение оценок тональности слов в отношении понятия «Навальный»

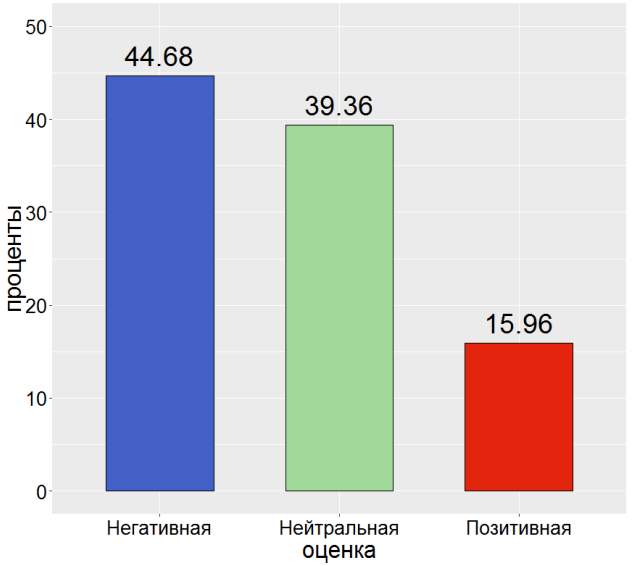

Рисунок 3. Распределение оценок тональности слов в отношении понятия «Путин»

Более наглядной визуализацией полученных результатов является облако слов, цвет которых отражает степень эмоциональности оценки. Для создания этих графиков был использован пакет R «wordcloud»[7]. Синие оттенки слов обозначают негативную оценку, красные — положительную, зеленый цвет — нейтральную. Чем интенсивнее цвет, тем выше степень тональности. Облако слов, связанных с понятием «Навальный», изображено на рис. 4, связанных с понятием «Путин» — на рис. 5.

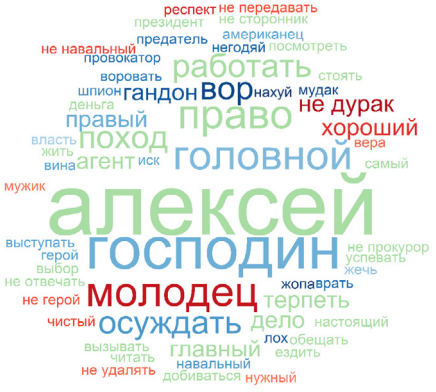

Рисунок 4. Облако слов, связанных с понятием «Навальный»

[7] Fellows I. Wordcloud: Word Clouds. R package version 2.5. 2014. [Электронный ресурс]. URL: https://CRAN.R-project.org/package=wordcloud (дата обращения: 22.03.2017).

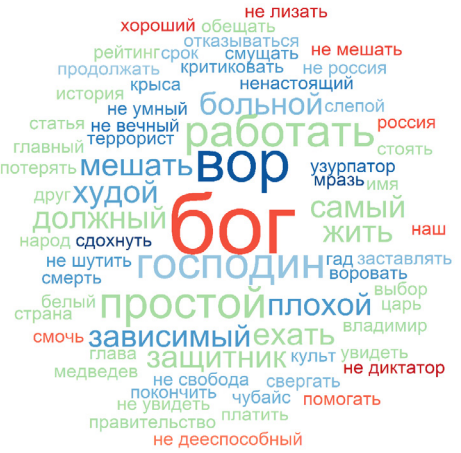

Рисунок 5. Облако слов, связанных с понятием «Путин»

Из представленных рисунков видно, что слова, положительно оцененные алгоритмом, не всегда являются таковыми. Например, фраза «не дееспособный» в облаке Путина была оценена как положительная, потому что люди писали ее с ошибкой (правильно — недееспособный), или самое часто повторяемое слово «бог» использовалось как элемент сарказма: «Ну да, Путин наверно бог или гений. Среди 140 миллионов россиян невозможно найти ему замену», «…ну, конечно, Путин бог и царь и без него России путь только в сторону Украины» и т. д. В облаке Навального тоже можно увидеть неоднозначно оцениваемые фразы, вроде положительной «не герой», и отрицательно оцененного слова «герой».

Этот пример по извлечению мнений демонстрирует очевидный факт: чем точнее хочется получить подсчет и оценку, тем более продвинутые инструменты для этого необходимы. Альтернативой уже готовым средствам может стать более конкретная гипотеза о структуре мнений. Например, в использованном парсере эту функцию может выполнять более сложная грамматика. Улучшит или подтвердит валидность результатов анализа также ручная проверка контекста, в котором присутствуют отобранные слова. Если слов или фраз получилось довольно много, то можно ограничиться самыми частотными. Такая проверка поможет улучшить точность оценок, но не исправит полноту.

В исследованном нами примере нет смысла сравнивать полученные рейтинги одобрения/неодобрения политиков, так как выборка текстов явно в большей степени содержит мнения людей, критически относящихся к действующей власти. Тем не менее результаты извлечения мнений можно сравнить со шкалами измерений отношения к политикам, которые используют полстерские организации. Например, в Левада-Центре и ВЦИОМ используются вопросы, измеряющие

эмоциональное отношение к персоналиям[8]. Шкалы этих вопросов более детальные, чем у разработанного нами анализатора, но важнее то, что в результате работы парсера удается получить слова, характеризующие объект интереса, вне зависимости от того, информацию о каком именно аспекте мнений мы хотим получить. Например, для того чтобы с помощью опроса узнать, ассоциируется ли действующая власть с коррупцией, необходимо задать хотя бы один соответствующий вопрос. Метод анализа мнений через парсинг текстов обходит эту проблему и имеет еще ряд достоинств. Во-первых, он позволяет получать детализированную картину представлений изучаемой группы, сравнимую с результатами качественных социологических методов (интервью, фокус-группы). Во-вторых, дается количественная и тональная оценка мнений, как в массовых опросах. В-третьих, этот метод более универсален, чем обычные статистические методы — его можно применять как на достаточно большой выборке текстов, так и на начальной стадии анализа транскриптов интервью.

Выводы

Проведенное сравнение результатов массового опроса с результатами, полученными в ходе анализа больших текстовых данных из сети Интернет, позволяет сделать следующие выводы, касающиеся итогов тематического моделирования и анализатора по извлечению мнений. Во-первых, существующие без посредничества интервьюера и заранее заданного опросного листа мнения и обсуждения в сети Интернет, будучи подвергнуты соответствующему анализу, позволяют исследователям получать представление о контексте формирования общественного мнения относительно события или персоналии. Во-вторых, результаты подобного анализа можно использовать для улучшения качества структуры опросника и самих вопросов.

В то же время пока результаты анализа текстов из сети Интернет не могут рассматриваться как полноценная альтернатива массовым опросам. Один из камней преткновения для распространения выводов, полученных в исследованиях социальных медиа, на нецифровую реальность — отсутствие теоретической базы для генерализации данных на более широкие группы населения [Дудина, 2016: 28]. Традиционная модель массового опроса предполагает привязку мнений к социально-демографическим группам, а результаты анализа текстов из сети Интернет дают возможность представить лишь спектр обсуждаемых тематик и их относительную популярность, но не позволяет сопоставить мнения с их носителями, поскольку при анализе данных из социальных медиа возникает проблема с получением надежной демографической информации. Более доступная информация о сетевой активности или культурных предпочтениях пользователей пока не вписывается в модель общественного мнения. Другая проблема состоит в нерешенности вопроса о том, как формировать выборку, представляющую «генеральную» совокупность. Возможно, эта проблема может быть решена через

[8] См. Пресс-выпуск № 3342. Александр Лукашенко: штрихи к портрету [Электронный ресурс] // ВЦИОМ. URL: https://wciom.ru/index.php?id=236&uid=116142 (дата обращения: 16.07.2017); Владимир Путин: отношение и оценки [Электронный ресурс] // Левада-Центр. URL: https://www.levada.ru/2017/04/24/15835/ (дата обращения: 16.07.2017).

поиск и мониторинг наиболее адекватных и надежных источников представления общественного мнения в сети Интернет.

Тематическое моделирование и алгоритмы по извлечению мнений — инструменты, выполняющие отчасти разные функции. Тематическая модель отражает структуру текстов и обсуждения в целом. Конкретные объекты внимания в таких моделях можно измерить только через долю темы, которая с ними связана. Алгоритмы по извлечению мнений разрабатываются, напротив, именно для оценки отношения к конкретным объектам, о которых говорится в тексте. Такие алгоритмы более универсальны и не так требовательны к длине и количеству анализируемых текстов, как тематические модели. Возможно, со временем анализаторы станут главным инструментом для оценки общественного мнения в Интернете, так как дают возможность оценивать конкретные слова, представляющие разные аспекты мнения. В то же время тематическое моделирование дает лучшее понимание контекста общественного мнения о событии или персоне.

Примеры сравнений результатов алгоритмизированного анализа данных из онлайн-источников и опросов демонстрируют большую гибкость опросных методов относительно набора характеристик, переменных, по которым социологии могут измерить отношение к интересующему предмету. Например, чтобы выяснить с какими достижениями или провалами ассоциируется политик, при использовании опроса можно просто задать соответствующий вопрос и классифицировать полученные ответы, в то время как методы анализа неструктурированных текстов не гарантируют обнаружения такой специфичной тематики, если подобные мнения присутствуют в постах и комментариях и их можно обнаружить «ручным» анализом. Поэтому основным перспективным направлением разработки методов анализа онлайн-текстов является совершенствование инструментария, а именно создание специфично социологических анализаторов. Такие примеры уже есть: это широко используемый в основном для английского языка словарь, учитывающий психологические характеристики текста (LIWC[9]); среди русскоязычных разработок можно отметить список измеряемых параметров пользователей и их текстов, который предлагает коммерческий проект «Social Data Hub»[10]. На данный момент глобальной перспективной задачей становится создание необходимого для социологов набора методов и вспомогательных инструментов анализа текста, способных воспроизвести структурно традиционный опрос.

Список литературы (References)

Дудина В. И. Цифровые данные — потенциал развития социологического знания // Социологические исследования. 2016. № . 9. С. 21—30. [Dudina V. I. (2016) Digital data potentialities for development of sociological knowledge. *Sociological Studies*. No. 9. P. 21—30.] (In Russ.)

Юдина Д. И., Дудина В. И. Семантическая сеть на биграммах как метод валидизации результатов тематического моделирования в социологическом исследовании

[9] LIWC. URL: http://liwc.wpengine.com/ (дата обращения: 21.09.2017).

[10] Построение скоринговых моделей [Электронный ресурс] // Social Data Hub. URL: https://socialdatahub.com/ru/postroenie_skoringovyih_modeley (дата обращения: 21.09.2017).

// Журнал социологии и социальной антропологии. 2016. Т. 19. № . 4. С. 71—83. [Iudina D. I., Dudina V. I. (2016) Semanticheskaya set' na bigrammakh kak metod validizatsii rezul'tatov tematicheskogo modelirovaniya v sotsiologicheskom issledovanii [Semantic Network on Bigrams as a Method for Validation of Topic Modelling Results in the Sociological Research]. *Zhurnal sotsiologii i sotsial'noi antropologii [Journal of Sociology and Social Anthropology]*. Vol. 19. No. 4. P. 71—83.] (In Russ.)

Berstel J., Boasson L. (1990) Context-Free Languages. In: Jan van Leeuwen, ed., Handbook of Theoretical Computer Science. Volume A — Algorithms and Complexity. Amsterdam: Elsevier. P. 59—102.

Cambria E., Gasta P., Bisio F., Zunino R. (2015) An ELM-based model for affective analogical reasoning. *Neurocomputing*. Vol. 149. P. 443—455.

Cody E. M., Reagan, A.J., Mitchell L., Dodds, P.S., Danforth C. M. (2015) Climate change sentiment on Twitter: An unsolicited public opinion poll. PLoS ONE. Vol. 10. No. 8. DOI: https://doi.org/10.1371/journal.pone.0136092

Cody E. M., Reagan A. J., Dodds P. S., Danforth C. M. (2016) Public Opinion Polling with Twitter. *arXiv preprint arXiv*:1608.02024. URL: http://www.uvm.edu/pdodds/research/papers/cody2016b/ (accessed: 10.10.2017)

Coussement K., Van den Poel D. (2009) Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Systems with Applications*. Vol. 36. № . 3. P. 6127—6134.

DiGrazia J., McKelvey K., Bollen J., Rojas F. (2013) More tweets, more votes: Social media as a quantitative indicator of political behavior. *PloS one*. Vol. 8. № . 11. P. e79449. DOI: https://doi.org/10.1371/journal.pone.0079449.

Forelle M. C., Howard P. N., Monroy-Hernández A., Savage S. (2015) Political bots and the manipulation of public opinion in Venezuela. URL: https://ssrn.com/abstract=2635800 (accessed: 23.09.2017). DOI: https://doi.org/10.2139/ssrn.2635800.

Gimpel K., Schneider N., O'Connor B., Das D., Mills D., Eisenstein J., Heilman M., Yogatama D., Flanigan J., Smith N. A. (2011) Part-of-speech tagging for twitter: Annotation, features, and experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers. Vol. 2. Association for Computational Linguistics. P. 42—47.

Grün B., Hornik K. (2011) topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software.* Volume 40, Issue 13. DOI: https://doi.org/10.18637/jss.v040.i13.

Hu N., Liu L., Sambamurthy V. (2011) Fraud detection in online consumer reviews. *Decision Support Systems*. Vol. 50. No. 3. P. 614—626.

Jiang L., Yu M., Zhou M., Liu X., Zhao T. (2011) Target-dependent twitter sentiment classification. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.* Vol. 1.— Association for Computational Linguistics. P. 151—160.

Justo R., Corcoran T., Luki, S.M., Walker M., Torres M. I. (2014) Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems*. Vol. 69. P. 124—133.

Lin C., He Y., Everson R., Ruger S. (2012) Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data engineering*. Vol. 24. № . 6. P. 1134—1145.

O'Connor B., Balasubramanya R., Routledg B., Smith N. (2010) From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*. Vol. 11. No. 122—129. P. 1—2.

Pang B., Lee L. (2008) Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*. Vol. 2. No. 1—2. P. 1—135.

Poria S., Cambria E., Winterstein G., Huang G. B. (2014) Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*. Vol. 69. P. 45—63.

Prabowo R., Thelwall M. (2009) Sentiment analysis: A combined approach. *Journal of Informetrics*. Vol. 3. No. 2. P. 143—157.

Ravi K., Ravi V. (2015) A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*. Vol. 89. P. 14—46.

Reyes A., Rosso P. (2012) Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision Support Systems*. Vol. 53. No. 4. P. 754—760.

Rill S., Reinel D., Scheidt J., Zicari R. V. (2014) Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. *Knowledge-Based Systems*. Vol. 69. P. 24—33.

Tang H., Tan S., Cheng X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications*. Vol. 36. No. 7. P. 10760—10773.

Tumasjan A. et al. (2010) Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsm*. Vol. 10. No. 1. P. 178—185.