# Методология дискурс-анализа: от ручного подсчёта данных к машинному обучению

**О.Г. Филатова[1], Д.В. Волковский[1], П.Н. Беген[2]**

*[1] Санкт-Петербургский государственный университет*
*[2] Университет ИТМО*

**Аннотация.**

**Ключевые слова:** интернет-дискурс, методология дискурс-анализа, машинное обучение, искусственный интеллект, сентимент-анализ

# The Methodology of Discourse Analysis: from Manuel Data Counting to Machine Learning

**O.G. Filatova[1], D.V. Volkovskii[1], P.N. Begen[2]**

*[1] Saint Petersburg State University*
*[2] ITMO University*

**Abstract.**

**Keywords:** Internet discourse, discourse analysis methodology, machine learning, artificial intellect, sentiment analysis

Изучение интернет-дискурса является методологически и эмпирически сложной задачей, поскольку в настоящее время встречаются ограничения научной области знаний и недостаток обоснованных теоретических и аналитических работ, связанных с анализом интернет-дискурса и репрезентацией его результатов. В целом, это ведёт к широким и разнородным по содержанию и формату интерпретациям эмпирического доказательства не только среди исследователей и экспертов, но и самих участников делиберативного онлайн-процесса. Концепция делиберативности, подробно разработанная в теории коммуникативной этики Юргена Хабермаса, обычно рассматривается как повседневная практика политических онлайн-обсуждений, которые возникают в связи с какими-то событиями и процессами в социально-политической сфере.

Однако особые трудности возникают с методологией исследования онлайн-дискурса, и на это есть несколько причин.

Во-первых, большая часть исследований виртуальной публичной сферы еще не материализовалась в аналитических рамках, которые позволили бы эмпирически изучать дискурсивные практики граждан. Другими словами, проблема заключается в том, как перевести нормативные ценности публичной сферы и дискурсивные этические теории в эмпирически дружественные аналитические инструменты для изучения дискурсивных процессов.

Во-вторых, эффективность конкретной методологии исследования для сбора эмпирических данных зависит от ее способности учитывать роль технологических и конструктивных свойств, которые позволяют функционировать онлайн-дискурсам. Отсутствие разграничения между технологическими и социальными свойствами веб-пространств может привести к менее надежным доказательствам и противоречивым интерпретациям дискурсов.

Более того, понимание обсуждения как разговора только о проблемах, а не действия по ним, все равно будет означать, что эффективность политического участия остается без внимания, поскольку участие также должно быть политически мотивированным гражданским действием. Такое узкое толкование публичной сферы является одной из причин, по которой исследования часто не дают убедительных доказательств прагматической полезности онлайн публичной сферы.

Большинство исследований массовых коммуникаций по-прежнему фокусируется на «аудиториях», «отправителях» и «получателях», которые не могут быть удовлетворительно применены к новым цифровым сообществам и их дискурсам. В эпоху электронных коммуникаций средства массовой информации утратили монополию на информирование общественности. Отсутствие инноваций в исследовании публичных обсуждений в интернете является одной из основных причин существующей двусмысленности и радикально противоположных взглядов на коммуникативные практики в интернете.

Выбирая методологию анализа интернет-дискурса, мы решили остановиться на методике дискурс-анализа, разработанной и описанной Ю. Г. Мисниковым в его PhD-диссертации, защищённой в университете Лидса (Великобритания). Учёный разработал «deliberative standard to assess discourse quality» [1], в котором были описаны семь тематически различных дискурсивных параметров, соответствующих конкретным вопросам исследования, для руководства процессом кодирования сообщений интернет-дискуссий. Отметим, что Ю. Г. Мисников сделал это первым, так как на момент выхода его диссертации в научной литературе не существовало прямых аналогов. Каждый параметр содержит

определённых набор эмпирических характеристик, предназначенных для отражения дискурсивных качеств.

*Первый параметр* коррелирует с *партисипативным равенством* и пост-активностью, содержащими семь характеристик: номер участника, его имя, статус, номер поста, номер и дата поста участника. При исследовании уровня активности граждан часто возникает проблема, связанная с неравномерным распределением долей участия в обсуждениях. Ещё заметим, что превалирование высоко интерактивных, сильно персональных и часто невежливых черт в интернет-дискуссиях приводит к их слабому, низкому, «неудовлетворительному» качеству.

*Второй параметр* раскрывает «*цивильность*», которая используется для характеристики качественного характера публичной онлайн-дискуссии и связано с демонстрацией толерантного отношения. Данные о цивильности не так просто интерпретировать, поскольку нет универсального подхода её определения. Бывают ситуации, когда сообщения содержат одновременно вежливые и невежливые речевые аспекты, что вызывает трудность в кодировке поста. Помимо использования грубых выражений, которые отчётливо демонстрируют умышлённую невежливость, некоторые сообщения могут только подразумевать неприятный подтекст. В некоторых случаях реакция субъектов онлайн-обсуждений на такие посты может выступать надёжным индикатором, отражающим всю сложность субъективных отношений, которые формируются между участниками в процессе дискуссий. Если говорить о вежливых сообщениях, то они могут иметь специальную цель и быть адресованы определённым участникам в более персонализированной манере или же с выделением акцентов на некоторых аспектах темы, что способствует большему вовлечению людей в обсуждение. В целом цивильность включает следующие тональности выражения отношения:

1) цивильные (экспрессивно вежливые, приветственные, необязательно поддерживающие, могут быть критичными);

2) нормальные (амбивалентные, нейтральные, могут как поддерживать, так и критиковать);

3) ацивильные (экспрессивно недружественные, пренебрежительные, враждебные, грубые, не обязательно критические);

4) другие (их сложно квалифицировать, что-то смешанное между цивильными и ацивильными).

*Следующий параметр – действенность утверждений*: истинность высказываний, нормативная правота, субъективная правдивость, согласие и несогласие. В качестве ещё одного параметра указаны намерения речевых актов, которые могут быть директивными (чётко направленными, не подлежащими спору), коммиссивными (допускаются некие поправки и корректировки) и экспрессивными (носят преимущественно эмоциональный характер).

Значимыми составляющими обсуждений являются такие *параметры,* как дискурсивные *интерактивность* и *диалогичность*, охватывающие персонально адресованные посты с использованием имени адресата; неперсонально адресованные; прямые ответы участников, включая цитаты; открытые ответы, обратную связь на сообщения и цитирование постов.

Диалогичность концептуально акцентирует внимание на окружающих. Если коммуникатор понимает их и знает, как найти к н им подход, то он лучше поймёт себя и свои коммуникативные действия. Однако, когда концепт «Я» спрятан в концепт «Другие», когда самореализация идёт через других, то тут и возникает сложность. Наши изречения нельзя определить как первоначальные или конечные, все они имеют предварительную историю и одновременно содержат предчувствие, связанное с реакциями других на то, что сказано или написано. Диалог – это распознавание нужд и интересов других через взаимность, которая включает не только согласие, но и оппозицию, противоречие. Когда есть обратная связь, аргументация, возражение субъектов дискурса, диалогические отношения между ними становятся более значимы, ибо избегают пассивное состояние. Рядом с понятием диалогичности находится понятие интерактивности, которую принято считать ключом к изучению публичных онлайн-дискурсов. Одним из её преимуществ является то, что не требуется, чтобы участники, вовлеченные в публичный диалог, встречались лицом к лицу. Интерактивность не обязательно стимулирует публичное взаимодействие индивидов. Скорее всего это возможность быть диалогичным и кооперативным с равными себе людьми, что привлекает остальных граждан к участию в онлайн-дискурсе. Несогласия, полемика рассматриваются как часть интерактивности в том числе. Существует спор насчёт того, кого из участников интернет-дискуссии можно считать интерактивным: того, кто отвечает только на предыдущее сообщение, или того, кто отвечает на многие сообщения. Мы считаем, что те и те будут являться интерактивными участниками, однако степень их интерактивности будет значительно различаться.

*Аргументация,* как существенный *параметр* при изучении онлайн-дискуссий, изменчива, многогранна, никогда не находится в статичном состоянии, направлена прежде всего на обеспечение понимания между участниками обсуждений и поддержание диалога между ними при каждом взаимодействии. Аргументация всегда важна, так как она помогает увидеть «крайние позиции», т.е. позиции согласия и несогласия, выступающие, в свою очередь, демократическими формами публичного рассуждения посредством межличностной интеракции.

Аргументирование является актом взаимного понимания между коммуникаторами и обоюдного признания других индивидов и их позиций. Соответственно, аргументация на взаимной основе, по сути, выступает как коммуникативная и дискурсивная. Качество аргументации

зависит не столько от говорящего, сколько от слушающего, поскольку нет смысла там, где нет диалога. Значение приобретает та коммуникация, в которой происходит ответ общественности на волнующие её проблемы, в ином случае коммуникативный акт становится бесполезным и бесчувственным [2]. Изолированные дискурсы не будут иметь практически никакого смысла для анализа, особенно в поляризированных социально-политических взаимоотношений, поскольку их участники не в достаточной степени представляются риторически убедительными и диалогически адаптивными [3].

Согласно Ю. Г. Мисникову, аргументация как дискурсивный стандарт включает два набора параметров: первый из них содержит факты, выводы, примеры, сравнения, обобщения, логические умозаключения, другие виды доказательств, второй – ссылки на онлайн-ресурсы, печатные медиа, радио и телевидение. Именно с этих позиций мы проанализировали аргументацию в онлайн-дискуссиях на тему пенсионной реформы.

И, наконец, *последний параметр – тематическое многообразие*. Вопросы обсуждения могут быть связаны с государством и его управлением, обществом и политикой, экономикой, социальными проблемами, регионами, международными отношения в пределах бывшего СССР и, далее, с культурой, здравоохранением, стилем жизни, медиа и интернетом.

Методика дискурсивного анализа, основанная на концепции Хабермаса и развивающая её (Хабермас никогда не подсчитывал результаты), была выбрана по нескольким причинам. Во-первых, если мы изучаем онлайн-дискурс с позиций политического PR, политической коммуникации, нам важна коммуникативная и социально-политическая стороны дискуссий, которые мы можем выявить с помощью используемого подхода.

Во-вторых, методика является понятной и несложной, ей удобно пользоваться в программе Excel. К данному методологическому подходу могут прибегать эксперты, PR-специалисты, обычные граждане, что расширяет сферу дискурсивного анализа, делает её разнообразной с точки зрения участников и придаёт демократичности с позиций общественного контроля. При изучении каждого комментария интернет-дискуссии ему присваивается свой уникальный числовой идентификатор в виде комбинации трех параметров: (i) уникальный идентификатор сообщения, закодированный (начиная с 1) в последовательном порядке публикации независимо от автора (последняя цифра представляет общее количество опубликованных высказываний всех участников); (ii) уникальный идентификатор участника, закодированный (начиная с 1) в порядке начала обсуждения (последняя цифра представляет общее количество всех участников); (iii) уникальный идентификатор высказывания, закодированный (начиная с 1) в последовательном порядке

публикации каждым участником (последняя цифра представляет общее количество опубликованных высказываний каждым участником). Таким образом, каждое опубликованное сообщение может быть однозначно идентифицировано с помощью трехзначной комбинации с точки зрения того, когда оно было отправлено, кем и в каком порядке. Другие атрибуты (например, статус участника форума или дата регистрации) могут быть добавлены по мере необходимости. Например, сообщение, закодированное как «12-4-2», означает, что это было 12-е сообщение в ветке, опубликованной участником, который вступил в дискуссию под номером 4, и это была его или ее вторая публикация до сих пор.

В-третьих, использование методики позволяет сформулировать суть мнений участников дискурса в отношении различных общественных проблем. В итоге формируются массивы данных, связывающих оригинальный авторский текст постов с неким обобщенным мнением, что в дальнейшем позволит предсказывать тип мнений в зависимости от содержания текста, выявлять множественность и неоднозначность мнений, высказываемых людьми по актуальной теме.

В-четвёртых, ручной подсчёт некоторых данных и их кодировка может сводиться к машинному обучению, что заметно ускоряет и облегчает работу исследователя, особенно когда количество постов превышает тысячи. Однако здесь не всё так легко. Чтобы компьютерная программа могла правильно обрабатывать информацию, необходимо её «тренировать» путём предоставления определённого количества постов, комментариев онлайн-обсуждений. По нашим данным их должно быть не менее одной тысячи.

Изучение искусственного интеллекта включает такое междисциплинарное направление, как машинное обучение, сочетающее в себе математическую статистику, методы оптимизации, методы извлечения информации, интеллектуальный анализ данных. Исследовательские работы в области машинного обучения обязательно подразумевают проведение экспериментов на модельных или реальных данных с целью проверки подлинности и качества работы методов, подтверждения гипотез, расчета статистических метрик, формирования списка критериев, обладающих статистической значимостью.

Центральными методами машинного обучения являются линейная и логистическая регрессия (linear and logistic regression), метод опорных векторов (SVM, support vector machines), деревья решений (decision trees), случайный лес (random forest), градиентный бустинг (boosting), нейронные сети (neural networks), глубокое обучение (deep learning), самоорганизующиеся карты (self-organizing maps) и др. [4,5]

В сочетании с алгоритмами машинного обучения используется обработка естественного языка (NLP), поскольку она позволяет идентифицировать диалогические акты, речь, обнаруживать эмоции,

анализировать тональность текста и т.д. Благодаря NLP естественный язык преобразуется в формат, используемый методами машинного обучения, чтобы реализовывать собственные алгоритмы. В качестве основных методов и подходов обработки естественного языка выделяются токенизация, составление списка стоп-слов, стемминг, лемматизация, извлечение именованных сущностей, модель «мешок слов», вычисление функции TF-IDF, алгоритмы Word2Vec [6,7,8,9].

Применению методов машинного обучения, в частности, искусственных нейронных сетей, в сентимент-анализе (анализе тональности текста) и компьютерной лингвистике послужили развитие информационно-коммуникационных технологий, значительное увеличение количества данных и рост вычислительных мощностей [11, 12]. На их основе нам удалось провести пилотное исследование с применением машинного обучения для сентимент-анализа дискуссий граждан на такую актуальную общественно-политическую тему, как российская пенсионная реформа [13]. Цель данного исследования заключалась в разработке автоматизированного инструментария для проведения анализа интернет-дискурса, в частности, проведение эксперимента по применению искусственных нейронных сетей в определении позиций граждан (сентимент-анализ) в онлайн-обсуждениях на тему повышения пенсионного возраста.

На основании результатов, полученных в ходе апробации автоматизированной программы для анализа интернет-дискуссий, было отмечено несколько перспектив для дальнейших исследований, первая из которых заключается в использовании машинного обучения как исследовательского инструментария для кодировки и анализа описанных в статье параметров делиберативного стандарта.

Вторая перспектива связана с созданием методов по распознаванию таких параметров, как аргументация и цивильность. Например, идентификация аргументации и некоторых её типов (ссылок на источники, цитат) будет базироваться на парсинге и использовании регулярных выражений для поиска ссылок. Выделение типов цивильности может осуществляться в автоматическом режиме, что похоже на анализ тональности, но, на самом деле, здесь совершенно другой подход в техническом плане.

Третья возможность упирается в предоставлении исследователям статистического анализа с элементами визуализации, к примеру, типов цивильности по городам и их классификации. Результаты могут быть выведены в таблице, а также представлены на графике или в виде диаграммы.

Таким образом, методология дискурс-анализа Ю.Г. Мисникова может постепенно переводиться в машинный формат и осуществляться с помощью возможностей искусственного интеллекта

Работа выполнена при поддержке РНФ, проект №18-18-00360 «Электронное участие как фактор динамики политического процесса и процесса принятия государственных решений».

**Литература**

1. Misnikov Y. Public Activism Online in Russia: Citizens' Participation in Webbased Interactive Political Debate in the Context of Civil Society. Development and Transition to Democracy: PhD thesis … Ph. D. / Y. Misnikov. Leeds, 2011.
2. Бахтин М. М. Проблема текста в лингвистике, филологии и других гуманитарных науках. Искусство, 1986. С. 297—325.
3. Habermas J. The Theory of Communicative Action. Reason and the Rationalization of Society, vol. 1. Beacon, Boston, 1984.
4. Zhang C., Ma Y. Ensemble Machine Learning. Methods and Applications. Springer, Boston, MA, 2012. DOI: https://doi.org/10.1007/978-1-4419-9326-7
5. Dietterich T.G. Ensemble Methods in Machine Learning. In: Multiple Classifier Systems // MCS 2000. Lecture Notes in Computer Science. 2000. Vol. 1857. DOI: https://doi.org/10.1007/3-540-45014-9_1
6. Goldberg Y. Neural Network Methods in Natural Language Processing. Morgan & Claypool Publishers, 2017. 310 p.
7. Goldberg Y. A Primer on Neural Network Models for Natural Language Processing // Journal of Artificial Intelligence Research. 2016. № 57. pp. 345–420.
8. Батура Т. В. Методы автоматической классификации текстов // Программные продукты и системы / Software & Systems. 2017. Т. 30. № 1. С. 85–99. DOI: 10.15827/0236-235X.117.085-099
9. Зиберт А. О., Хрусталев В. И. Разработка системы определения наличия заимствований в работах студентов высших учебных заведений. Методы предварительной обработки текста // Universum: Технические науки: электрон. научн. журн. 2014. №4 (5). URL: http://7universum.com/ru/tech/archive/item/1258 (дата обращения: 21.04.2020)
11. Wang P., Xu J., Xu B., Liu C., Wang H.Z.F., Hao H. Semantic Clustering and Convolutional Neural Network for Short Text Categorization // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China, July 26-31, 2015. P. 352–357. URL: http://www.aclweb.org/anthology/P15-2058.
12. Socher R., Perelygin A., Wu J.Y., Chuang J., Manning C.D., Ng A.Y., Potts C. Recursive deep models for semantic compositionality over a sentiment treebank // EMNLP. 2013. P. 1631–1642.

13. Begen P.N., Misnikov Y.G., Filatova O.G. Application of automated tools in researching internet discourses: Experience of using the recurrent neural networks for studying discussions on pension reform // 21st Conference on Scientific Services and Internet, SSI- 2019. T.2543. 2019. P. 336-344/

**References**

1. Misnikov Y. Public Activism Online in Russia: Citizens' Participation in Webbased Interactive Political Debate in the Context of Civil Society. Development and Transition to Democracy: PhD thesis … Ph. D. / Y. Misnikov. Leeds, 2011.
2. Bahtin M. M. Problema teksta v linguistike, filologii i drugih gumanitarnuh naykah. Iskusstvo, 1986. S. 297—325.
3. Habermas J. The Theory of Communicative Action. Reason and the Rationalization of Society, vol. 1. Beacon, Boston, 1984.
4. Zhang C., Ma Y. Ensemble Machine Learning. Methods and Applications. Springer, Boston, MA, 2012. DOI: https://doi.org/10.1007/978-1-4419-9326-7
5. Dietterich T.G. Ensemble Methods in Machine Learning. In: Multiple Classifier Systems // MCS 2000. Lecture Notes in Computer Science. 2000. Vol. 1857. DOI: https://doi.org/10.1007/3-540-45014-9_1
6. Goldberg Y. Neural Network Methods in Natural Language Processing. Morgan & Claypool Publishers, 2017. 310 p.
7. Goldberg Y. A Primer on Neural Network Models for Natural Language Processing // Journal of Artificial Intelligence Research. 2016. № 57. pp. 345–420.
8. Batyra T. V. Metody avtomaticheckoi klassifikacii tekstov // Programnye producti i sistemi / Software & Systems. 2017. T. 30. № 1. C. 85–99. DOI: 10.15827/0236-235X.117.085-099
9. Zibert A. O., Hrystalev V. I. Razrabotka sistemi opredelenia nalachia zaimstvovanii v rabotah studentov vychih uchebnih zavedenii. Metodi predvaritelnoi obrabotki teksta // Universum: Tehnicheskie nayki: elektron. naychn. zhur., 2014. №4 (5). URL: http://7universum.com/ru/tech/archive/item/1258 (data obrachenia: 21.04.2020)
10.
11. Wang P., Xu J., Xu B., Liu C., Wang H.Z.F., Hao H. Semantic Clustering and Convolutional Neural Network for Short Text Categorization // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China, July 26-31, 2015. P. 352–357. URL: http://www.aclweb.org/anthology/P15-2058.

12. Socher R., Perelygin A., Wu J.Y., Chuang J., Manning C.D., Ng A.Y., Potts C. Recursive deep models for semantic compositionality over a sentiment treebank // EMNLP. 2013. P. 1631–1642.
13. Begen P.N., Misnikov Y.G., Filatova O.G. Application of automated tools in researching internet discourses: Experience of using the recurrent neural networks for studying discussions on pension reform // 21st Conference on Scientific Services and Internet, SSI- 2019. T.2543. 2019. P. 336-344/