**Д.К. СТУКАЛ, В.Е. БЕЛЕНКОВ, И.Б. ФИЛИППОВ***

# МЕТОДЫ НАУК О ДАННЫХ В ПОЛИТИЧЕСКИХ ИССЛЕДОВАНИЯХ: АНАЛИЗ ПРОТЕСТНОЙ АКТИВНОСТИ В СОЦИАЛЬНЫХ СЕТЯХ[1]

*Аннотация.* Появление и рост популярности социальных сетей, а также растущая цифровизация, проникающая в разнообразные сферы экономики и общества, оказали существенное влияние на сферу политики в целом и на процессы политической мобилизации и коммуникации в частности. Методологический арсенал политической науки также оказался затронут указанными трансформационными процессами и начал пополняться новыми подходами и методами, предложенными в рамках недавно возникшей области знания, получившей название наук о данных. В статье предложен обзор ключевых инноваций в методологии исследований политической мобилизации в социальных сетях, которые были заимствованы из области наук о данных. Авторы подробно рассматривают методы обучения с учителем и без учителя и обсуждают их применимость для автома-

* **Стукал Денис Константинович**, кандидат политических наук, PhD, заместитель директора Института прикладных политических исследований, Национальный исследовательский университет «Высшая школа экономики» (Москва, Россия), e-mail: dstukal@hse.ru; **Беленков Вадим Евгеньевич**, магистр, аспирант департамента политики и управления, Национальный исследовательский университет «Высшая школа экономики», редактор Отдела научных изданий Управления научной политики, МГИМО (У) МИД России (Москва, Россия), e-mail: vadim.belenkov@gmail.com; **Филиппов Илья Борисович**, аспирант департамента политики и управления, Национальный исследовательский университет «Высшая школа экономики» (Москва, Россия), e-mail: ibfilippov@gmail.com

[1] Исследование выполнено за счет гранта Российского научного фонда (проект № 20-18-00274), Национальный исследовательский университет «Высшая школа экономики».



DOI: 10.31249/poln/2021.01.02

тизированного сбора данных в почти реальном времени и анализа собранных данных о протестной активности. В контексте методов обучения с учителем особое внимание уделяется методам преодоления переобучения с помощью регуляризации и выбору гиперпараметров с помощью кросс-валидации. В рамках обучения без учителя рассматриваются методы тематического моделирования и методы анализа социальных сетей. Преимущества и недостатки обсуждаемых методов иллюстрируются примерами из современных политических исследований, опубликованных в ведущих рецензируемых журналах. В заключение обсуждаются новейшие методные разработки наук о данных, до сих пор не получившие своего применения в исследованиях политической мобилизации, обладающие высоким аналитическим потенциалом (включая методы с частичным обучением, использование машинного обучения для каузального анализа и использование векторного представления текстов).

*Ключевые слова:* политическая мобилизация; протесты; социальные сети; машинное обучение; науки о данных; обучение с учителем; обучение без учителя; вычислительные социальные науки.

*Для цитирования:* Стукал Д.К., Беленков В.Е., Филиппов И.Б. Методы наук о данных в политических исследованиях: анализ протестной активности в социальных сетях // Политическая наука. – 2021. – № 1. – С. 46–75. – DOI: http://www.doi.org/10.31249/poln/2021.01.02

## Введение: как Интернет изменил методы социальных наук

Изучение факторов и характеристик политической мобилизации, динамики протестных настроений и акций в обществах с различными типами общественно-политического устройства – одно из важных направлений современных политических исследований, переживающее в последнее десятилетие подъем на волне распространения новых информационных технологий и обогащения методологического арсенала социальных наук методами бурно развивающихся наук о данных [Big data ..., 2016; Enikolopov, Makarin, Petrova, 2020]. Какие возможности открывают новые вычислительные методы перед исследователями политической мобилизации? Как эти методы могут быть использованы в процессе сбора и анализа эмпирических данных? Наконец, каковы методологические границы вычислительных методов и анализируемых с их помощью данных?

Развитие методов наук о данных в значительной степени связано с развитием новых информационных технологий и распространением социальных сетей, позволяющих изучать общественные настроения и политическую коммуникацию почти в реаль-

ном времени. Важным фактором оказалось также то, что социальные сети сделали возможным изменение самих технологий политической мобилизации: распространение мобилизирующей информации стало не только более быстрым, но и децентрализованным; появились возможности обновлять информацию о протестах в реальном времени и отслеживать широкий круг источников о протестах. Одновременно с этим социальные сети предложили гражданам альтернативную – онлайновую – форму участия, обусловливающую ощущение вовлеченности в политическую активность без фактического участия в событиях на улице (так называемый slacktivism), что может существенно ослабить реально наблюдаемую политическую мобилизацию. Наконец, социальные сети – вопреки первоначальным оптимистичным прогнозам об их превращении в универсальную технологию продвижения свободы и демократии [Diamond, 2010] – со временем стали использоваться и для политической демобилизации.

Разнообразие форм проникновения новых информационных технологий в мир политики, а также нелинейный характер динамики политической мобилизации под влиянием распространения Интернета и цифровых технологий обусловили два важных методологических тренда. С одной стороны, стала очевидной ограниченность простых методов и моделей, применяемых к анализу неэкспериментальных (observational) данных, полученных с помощью новых информационных технологий. В связи с этим вырос интерес к более широкому кругу методов, развиваемых в рамках наук о данных, что нашло выражение в возникновении такого направления, как вычислительные социальные науки (computational social sciences), включающего в себя как применение методов наук о данных и машинного обучения к исследованию вопросов социальных наук, так и использование для этих целей различных методов вычислительного моделирования, анализа системной динамики и др. [Computational social science, 2009; Cioffi-Revilla, 2010]. С другой стороны, возник запрос на уточнение условий, при которых данные, собранные с помощью новых информационных технологий, могут быть использованы для исследования причинно-следственных связей (т.е. каузального анализа) [Big data ..., 2016], а также стала распространяться практика сочетания методов наук о данных с тщательно продуманными квазиэкспериментальными планами исследования, позволяющими давать численные оценки

каузальным эффектам [Zhuravskaya, Petrova, Enikolopov, 2020; Enikolopov, Makarin, Petrova, 2020].

В данной статье подробно рассматривается первая из отмеченных тенденций, а также предлагается обзор основных методов наук о данных в их приложении к изучению политической мобилизации. Мы рассматриваем два крупных класса методов: обучение с учителем (supervised learning) и без учителя (unsupervised learning). Оба класса методов – несмотря на свои принципиальные различия – получили достаточно широкое распространение в работах, посвященных анализу протестной активности в странах мира. Кроме того, мы рассматриваем новейшую практику использования методов с учителем для сбора эмпирических данных в (почти) реальном времени, что представляет интерес не только для специалистов в области политического поведения, но и исследователей международных отношений и конфликтов.

## Методы обучения с учителем

***Общие методологические вопросы.*** Методы обучения с учителем – это обобщенное название методов анализа данных, предполагающих построение алгоритма для предсказания одной или нескольких зависимых переменных на основе набора объясняющих признаков. Особенность этих методов – наличие размеченных данных, т.е. верных значений зависимых переменных, выступающих в роли учителя, указывающего на ошибки. Наиболее известным примером методов обучения с учителем является линейная регрессия, давно ставшая ключевым методом количественных исследований в области политических наук [Krueger, Lewis-Beck, 2008]. Однако, если традиционное применение регрессионных моделей было сфокусировано на измерении степени взаимосвязи (или даже влияния) между объясняющими и зависимыми переменными (что требовало использования достаточно простых моделей с легко интерпретируемыми параметрами), то многие современные методы машинного обучения ориентированы, прежде всего, на решение задачи *прогнозирования* и в особенности на повышение качества вневыборочного прогноза *(out-of-sample performance)* [Molina, Garip, 2019].

Под качеством вневыборочного прогноза понимается способность модели предсказывать значения зависимых переменных на основе объясняющих признаков за рамками выборки. При этом степень желаемой обобщаемости может различаться и зависит как от амбиций исследователя, так и характера решаемой задачи. Например, при изучении факторов, обусловливающих участие в массовых акциях протеста в Венесуэле, вневыборочными наблюдениями можно считать как венесуэльских протестующих, не попавших в выборку, так и протестующих в других странах в данное время или даже в прошлом или будущем. Естественно, алгоритмы, нацеленные на применение за рамками исходной выборки, не должны быть чрезмерно чувствительны к случайным взаимосвязям, свойственным лишь собранному массиву данных. Наоборот, такие алгоритмы должны улавливать лишь наиболее сильные и устойчивые закономерности, с высокой вероятностью присутствующие в разных массивах наблюдений, и игнорировать слабые и случайные взаимосвязи, которые, скорее всего, свойственны лишь собранным данным.

Для того чтобы избежать переобучения (*overfitting*) алгоритма и повысить его шансы на обобщаемость, методы обучения с учителем используют различные подходы, среди которых наиболее известна регуляризация (*regularization*). Поясним ее суть на примере линейной регрессии, оцениваемой методом наименьших квадратов (МНК): если обычный МНК сводится к поиску таких коэффициентов регрессии, при которых минимальной оказывается суммарная ошибка (квадрат отклонений реальных значений зависимой переменной от модельных), то в линейной регрессии с регуляризацией к этой сумме квадратов добавляется штрафной компонент, который растет по мере роста значений коэффициентов. Выбору штрафного компонента посвящена большая литература (например, см.: [Hastie, Tibshirani, Wainwright, 2016 a]), но наиболее часто на практике используется либо сумма модулей коэффициентов (такая регрессия называется LASSO) [Tibshirani, 1996], либо сумма квадратов коэффициентов (гребневая регрессия) [Hoerl, Kennard, 1970]. Практический смысл штрафного компонента состоит в сознательном занижении (а в случае LASSO-регрессии даже обнулении) коэффициентов при тех переменных, которые относительно слабо связаны с зависимой переменной. Конечно, такое занижение коэффициентов снижает качество алго-

ритма на имеющейся выборке; однако оно же может значительно повысить качество вневыборочного прогноза. Важно, что степень регуляризации (занижения) коэффициентов можно корректировать за счет гиперпараметра (коэффициента при штрафном компоненте), специально подбираемого так, чтобы повысить вневыборочную прогностическую силу модели.

Выбор гиперпараметров (например, коэффициента при штрафном компоненте в регуляризованных моделях) с помощью кросс-валидации (*cross-validation*) – одна из ключевых инноваций, которая была привнесена в методологию социальных наук из наук о данных и которая позволяет регулировать прогностическую силу модели. Технологически кросс-валидация осуществляется путем разделения выборки на k частей и последовательного построения алгоритма на (k-1) частях с измерением качества алгоритма на оставшейся части; полученные таким образом k показателей качества вневыборочного прогноза затем усредняются; в конечном счете выбираются такие гиперпараметры, которые обеспечивают наилучшие в среднем результаты кросс-валидации. Заметим, что описанная здесь процедура кросс-валидации призвана повысить обобщаемость алгоритма в узком смысле (например, на респондентов, не попавших в выборку). В прикладных исследованиях, однако, можно заменить кросс-валидацию на выбор гиперпараметров путем сравнения качества прогноза на других массивах данных (например, протестных данных в других странах или в другое время).

Нацеленность методов обучения с учителем на прогностическую силу (и лишь в меньшей степени – на интерпретируемость результатов) объясняет рост популярности таких ранее не применявшихся в политологических исследованиях методов, как метод опорных векторов (*support vector machine*), случайный лес (*random forest*), ансамбли методов типа бэггинга (*bagging*), бустинга (*boosting*) или экстремального градиентного бустинга (*xgboost*), или, наконец, различные типы нейронных сетей [Hastie, Tibshirani, Friedman, 2016 b]. Все эти методы по отдельности или вместе могут применяться на разных стадиях как для сбора эмпирических данных о протестной мобилизации, так и для их анализа.

***Применение для сбора данных.*** Развитие новых информационных технологий и распространение социальных сетей позволили иначе взглянуть на процедуру сбора эмпирических данных.

И хотя традиционная проблема качества собираемых данных [Freedman, 1991] отнюдь не решается путем простого увеличения объема данных, возможность автоматизации сбора данных позволяет если не заменить традиционные подходы, то хотя бы повысить эффективность некоторых из них. В частности, это касается эмпирических данных о случаях протестных акций, и сегодня зачастую собираемых вручную [Lankina, Tertytchnaya, 2020].

Появились, однако, и более автоматизированные подходы, исторически зародившиеся в исследованиях международных отношений. Именно в этой сфере начиная с конца 1970–1980-х годов предпринимались попытки разработать унифицированные методики кодирования различных международных событий, представленные в то время базами WEIS [McClelland, 1976] и COPDAB [Azar, 1980]. В 2000-е годы исследователями были предложены уже развернутые иерархические классификации (онтологии) событий с сопровождающими их описаниями. К их числу относятся используемые до сих пор IDEA (*integrated data for events analysis*) [Integrated data …, 2003] и CAMEO (*conflict and mediation event observations*) [Schrodt, Gerner, Yilmaz, 2009], а также разрабатываемая в наши дни PLOVER (*political language ontology for verifiable event records*) [Open Event Data Alliance, 2020], которая призвана прийти на смену CAMEO. Несмотря на неизбежную ограниченность онтологий и тенденцию группировать не всегда похожие по своей сути события в одну категорию [Schrodt, Van Brackle, 2013], их наличие позволило в середине 2000-х годов приступить к разработке одной из первых систем автоматического сбора данных о международных событиях (включая протестные акции и ответные действия правительств) на основе публикаций СМИ. Разработка такой системы была выполнена в рамках специального исследования Управления перспективных исследовательских проектов Министерства обороны США (*DARPA*) под названием ICEWS (integrated crisis early warning system) [O'Brien, 2010; Boschee, Natarajan, Weischedel, 2013]. Одновременно с этим была разработана альтернативная система GDELT (*global data on events, location, and tone*) [Leetaru, Schrodt, 2012], также основанная на мониторинге СМИ на 100 языках и автоматическом извлечении информации о 300 категориях событий. Интересно, что сравнение массивов данных, полученных этими двумя автоматизированными системами, указывает на существенные разночтения между ними

[Comparing GDELT and ICEWS Event Data, 2013], что подчеркивает необходимость осторожного обращения с этими данными в прикладных исследованиях.

Стремясь обойти ограничения автоматизированных систем, обусловленные использованием онтологий, М. Кройку и Н. Вайдманн применили обучение с учителем (в частности, ансамбль из опорных векторов и наивного байесовского классификатора) для автоматического отбора публикаций СМИ, посвященных протестным акциям в недемократических государствах [Croicu, Weidmann, 2015]. Их работа является примером сочетания современных методов наук о данных с традиционными методами сбора информации: автоматизированный отбор публикаций в новостных агрегаторах LexisNexis и Factiva производится для оптимизации работы специально обученных людей-кодировщиков и увеличения доли кодируемых ими публикаций, в которых действительно упоминаются протесты. Результатом этой работы является публично доступная база данных «Массовая мобилизация в автократиях» [Weidmann, Rød, 2019][1].

Обратным примером сочетания обучения с учителем с традиционными методами сбора событийных данных является автоматизированная система прогнозирования конфликтов ViEWS [ViEWS …, 2019], предсказывающая на 36 месяцев вперед три типа конфликтов (конфликты с участием правительственных сил, конфликты негосударственных акторов без участия правительственных сил и одностороннее насилие в отношении мирного населения) в Африке по данным базы UCDP (*Uppsala Conflict Data Program*) [Sundberg, Melander, 2013]. Данные UCDP кодируются вручную на основе публикаций из новостного агрегатора Factiva, система же ViEWS обучает на этих данных ансамбль из более чем 20 алгоритмов (логистических регрессий и случайных лесов) для прогнозирования конфликтов (в том числе вооруженных протестов). Полученные прогнозы могут быть далее использованы в ходе целенаправленного сбора эмпирических данных (например, запуска автоматического сбора данных из социальных сетей).

Таким образом, обучение с учителем позволяет не только (и даже не столько) заменить человека в процессе сбора эмпирики,

[1] Mass Mobilization in Autocracies Database. – 2020. – Mode of access: https://mmadatabase.org/ (accessed: 19.10.2020).

но и упростить, оптимизировать или сфокусировать традиционные методы сбора данных.

***Применение в анализе данных о протестах.*** Методы обучения с учителем имеют также мощный аналитический потенциал и широко применяются для анализа данных о протестной активности (в частности, собранных в социальных сетях). Они используются как для прогнозирования протестных событий или действий [Pachinko Prediction …, 2020; Predicting online protest participation …, 2016], так и для решения задачи выявления взаимосвязей между переменными.

Недавнее исследование авторского коллектива под руководством Дж. Тюка [Pachinko Prediction …, 2020] – хороший пример работы, направленной на решение задачи прогнозирования. В частности, решалась задача определения вероятности возникновения протестного события в девяти городах Австралии в каждый из дней с 21 июля 2017 г. по 14 февраля 2018 г. Для этого с помощью публичного API авторы собрали короткие сообщения в сети Твиттер (твиты), описывающие запланированные на будущее действия или события и написанные в одном из девяти крупнейших городов Австралии. Данные о протестных событиях были собраны вручную с крупнейших новостных сайтов Австралии.

В первую очередь, методы обучения с учителем (линейный метод опорных векторов с регуляризацией) были использованы в работе для классификации твитов на те, что описывают и не описывают протесты. В качестве объясняющих признаков в этой задаче выступали слова и словосочетания (уни- и биграммы). В процессе обучения использовалось разбиение массива данных на обучающую и тестовую выборки для выбора значений гиперпараметров, обеспечивающих наилучший вневыборочный прогноз.

Далее решалась задача прогнозирования протестных событий, для чего использовались методы байесовской статистики, позволяющие не только получить бинарные прогнозы (будет протест или нет), но и выразить посредством апостериорного распределения степень уверенности в нем.

Данные из социальных сетей также используются для прогнозирования дальнейшего поведения пользователей социальных сетей. Так, в работе С. Ранганата и его коллег предсказывалось поведение пользователей Твиттера во время протестов, связанных с президентскими выборами в Нигерии в 2015 г. [Predicting online

protest participation …, 2016]. Исследователи поставили перед собой цель предсказать, будут ли сообщения каждого пользователя посвящены протесту, на основе данных об истории коммуникации этого пользователя (о содержании отправленных и полученных ими сообщений, а также о связях с другими пользователями в Твиттере). Решению описанной задачи машинного обучения предшествовал сбор данных с геолокацией в Нигерии через API Твиттера, а также ручная разметка твитов как протестных и непротестных. К собранным данным был применен алгоритм прогнозирования, основанный на модели стохастического процесса, известного как геометрическое броуновское движение. На основе набора метрик качества прогноза (включая долю верных прогнозов, гармоническое среднее точности и полноты, известное как F-мера, и др.) в работе показано, что построенный алгоритм способен лучше прогнозировать содержание последующего твита пользователя (протестное или непротестное), чем целый ряд других моделей поведения пользователей. Разработанный алгоритм позволяет использовать историю сообщений пользователей и их фолловеров, а также информацию о структуре сети для прогнозирования протестного поведения этих пользователей в Твиттере в будущем [Predicting online protest participation …, 2016].

Помимо прогнозирования, методы обучения с учителем, как уже было отмечено выше, применяются для решения более типичной для политической науки задачи измерения связей между переменными и проверки гипотез об этих связях. Для этого используются как традиционные линейные и нелинейные модели (линейная МНК-регрессия, логистическая и пробит-модели), так и более сложные методы обучения с учителем.

Примером такого исследования является работа А.С. Ахременко и др., в которой данные о протестной активности в венесуэльском сегменте Твиттера были использованы для исследования факторов, объясняющих высокую популярность одних твитов о протестах и низкую популярность других [Ахременко, Стукал, Петров, 2020]. Объясняемой величиной в этой работе выступало число ретвитов, полученных каждым твитом, а объясняющие переменные включали в себя как сетевые характеристики автора, так и признаки, характеризующие содержание текста. На основе применения МНК- и LASSO-регрессии к анализу более 5,7 млн твитов авторы показали, что сете-

вые характеристики авторов твитов имеют большую объясняющую силу, чем содержательные признаки.

Помимо МНК-регрессий, особую популярность среди линейных моделей в последние десятилетия завоевала регрессия со смешанными эффектами, позволяющая учитывать разнородный характер объектов наблюдения и даже моделировать неоднородность связей между переменными у разных наблюдений. Примером использования такой модели является работа Р. Мурао и В. Чен, в которой на основе данных Твиттера изучались факторы, обусловливающие отношение журналистов и СМИ к участникам протестов в Бразилии [Mourão, Chen, 2020]. Единицей наблюдения в этом исследовании выступал отдельный твит, а факторы, потенциально влияющие на его содержание, могли действовать на других уровнях. Данное исследование, таким образом, иллюстрирует методы анализа данных с многоуровневой структурой, в которой разные наборы переменных изменяются на разных уровнях. В таких случаях именно модели со смешанными эффектами позволяют корректно моделировать связи между переменными [Steenbergen, Jones, 2002]. С помощью регрессионных моделей со смешанными эффектами Р. Мурао и В. Чен показали, что тональность освещения протестов журналистами в Твиттере положительно связана с их личным отношением к протестам. Кроме того, было показано, что бразильские журналисты освещали в Твиттере протесты правых сил 2015 г. менее позитивно, чем протесты левых сил 2013 г. [Mourão, Chen, 2020].

В современных исследованиях протестной активности широко используются модели, предназначенные для работы с зависимыми переменными, принимающими лишь целые неотрицательные (*далее* – натуральные) значения. Такими переменными в контексте анализа социальных сетей и интернет-трафика могут выступать количество опубликованных сообщений или число поисковых запросов. Для моделирования таких зависимых переменных используются отрицательная биномиальная или пуассоновская регрессии [Wooldridge, 2002]. Примером недавнего исследования, использующего отрицательную биномиальную регрессию для моделирования натуральных зависимых переменных, является работа Дж. Пан и А. Зигель, изучающая влияние репрессий на количество твитов лидеров общественного мнения в Саудовской Аравии и их онлайн-читателей в сети Твиттер [Pan, Siegel, 2020]. В этой работе,

нацеленной на измерение причинно-следственных связей, используется популярный дизайн исследования, получивший название «разности в разностях» (*difference in differences*), а устойчивость результатов проверяется с использованием отрицательной биномиальной модели. Оба подхода показывают, что офлайн-репрессии (аресты) лидеров общественного мнения за активность в Интернете сокращают количество опубликованных ими твитов после их освобождения, но не сокращают (по крайней мере, в течение месяца после ареста) количество твитов, схожих с репрессированными, но не репрессированных лидеров общественного мнения [Pan, Siegel, 2020]. В другой недавней работе отрицательная биномиальная регрессия применялась для изучения того, как эмоции, вызываемые изображениями в твитах сторонников протестного движения Black Lives Matter, влияют на количество ретвитов [Casas, Williams, 2019]. Исследователи пришли к выводу, что сообщения, содержащие изображения, в среднем получали большее (по сравнению с сообщениями без изображений) количество ретвитов как в целом, так и от ранее не участвовавших в обсуждении протеста пользователей социальной сети. Наибольшее количество ретвитов получали сообщения с изображениями, вызывающими энтузиазм или страх. Изображения же, провоцирующие грусть, напротив, были отрицательно связаны с количеством ретвитов [Casas, Williams, 2019].

В современных исследованиях протестной активности для анализа данных из социальных сетей широко используются разнообразные методы машинного обучения с учителем. Они используются как с традиционной целью выявления характера и направленности связей между переменными, так и с целью прогнозирования протестных действий и протестных событий в будущем. При этом проведенный обзор позволяет говорить о разнообразии применяемых методов и задач, для решения которых они используются.

## Методы обучения без учителя

***Общие методологические вопросы.*** Альтернативный класс методов наук о данных, широко применяемый в исследованиях политической мобилизации, – это обучение без учителя (например, факторный [Иберла, 1980; Basilevsky, 1994] или кластерный

анализ [Cluster analysis, 2011]). Как следует из названия, методы этого класса не используют в своей работе значения зависимых переменных, и потому имеют ограниченную применимость в задачах предсказания. Вместо этого обучение без учителя используется для *поиска структуры в данных* и чрезвычайно полезно в задачах разведывательного анализа, позволяя исследователю оперативно выявить потенциально интересные группы наблюдений или закономерности в собранных данных.

В современных исследованиях политической мобилизации методы обучения без учителя в основном используются для обработки естественного языка и тематического моделирования (*topic modeling*), т.е. автоматического выявления тем в большом массиве текстовой информации (чаще всего – в публикациях СМИ или социальных сетей).

Из методов тематического моделирования в прикладных политических исследованиях наиболее часто применяется латентное размещение Дирихле (*LDA, latent Dirichlet allocation*) [Blei, Ng, Jordan, 2003] или его усовершенствованные версии типа модели коррелирующих тем (*correlated topic model*) [Blei, Lafferty, 2007] или предложенная группой американских политологов, социологов и статистиков структурная тематическая модель (*structural topic model*) [The structural topic model ..., 2013; Structural topic models ..., 2014]. Все эти модели объединены рядом важных особенностей, отличающих их от классических методов кластерного анализа (иерархической кластеризации или метода *k*-средних). Во-первых, это вероятностные модели смеси распределений, оцениваемые методами байесовской статистики. С одной стороны, это позволяет исследователям включать в процесс оценивания моделей свои априорные представления о взаимосвязях между содержанием текстов и другими наблюдаемыми характеристиками (например, партийной принадлежности авторов текстов). С другой стороны, в результате оценивания этих моделей получаются распределения вероятностей, позволяющие должным образом характеризовать меру неопределенности относительно групповой принадлежности текстов. Во-вторых, эти модели относятся к классу моделей смешанного членства (*mixed membership*), т.е. один и тот же текст может одновременно (и с разными вероятностями) принадлежать к разным тематическим группам. Иными словами, эти модели допускают, что в каждом тексте речь идет не об одной, а

сразу о многих темах; удельный же вес каждой темы отражается в присваиваемом каждому тексту распределении вероятностей на множестве тем.

Одним из существенных недостатков многих методов тематического моделирования, однако, является неидентифицируемость модели из-за неупорядоченности тем (*label-switching*): выявляемые темы можно перенумеровать без потери прогностической силы модели. Из-за этого итоговое распределение вероятностей имеет сложную форму и множество мод [Roberts, Brandon, Dustin, 2016], что отрицательно сказывается на реплицируемости результатов анализа, а также существенно осложняет сам процесс статистического оценивания. Наконец, еще одна сложность практической реализации методов тематического моделирования состоит в необходимости выбирать общее число тем (что, впрочем, требуется и в большинстве методов кластерного анализа). Существует литература, посвященная разработке методов систематического решения проблемы выбора общего числа тем [Griffiths, Steyvers, 2006; Reading tea leaves ..., 2009; Evaluation methods for topic models, 2009]; кроме того, предложены методы, в которых число тем оказывается одной из моделируемых величин и потому также подлежит оценке в рамках модели [Hierarchical Dirichlet processes, 2006].

Отдельным подклассом методов, нацеленным на определение структуры в имеющихся данных, является анализ социальных сетей. Анализ социальных сетей опирается на базовое представление о том, что структура связей между объектами в выборке может оказывать существенное влияние на эти объекты и выборку в целом. Применение методов анализа социальных сетей может быть полезно для анализа процессов, в которых необходимо учитывать локальный контекст для каждого объекта и / или его связь с глобальным контекстом всей выборки [Kadushin, 2012].

Анализ социальных сетей применим в тех случаях, где данные представимы в виде социального графа. Социальный граф является совокупностью вершин и ребер между ними [Kadushin, 2012]. Вершинами (точками, узлами) становятся объекты – в рамках социальных наук обычно объектами становятся индивиды [Kadushin, 2004], страны [Hafner-Burton, Kahler, Montgomery, 2009] или сообщества [Padjett, Ansell, 1993]; ребрами (дугами) выступают связи между этими объектами. Рёбра могут иметь или не иметь

направление связи, в зависимости от этого граф является направленным или ненаправленным.

Представление данных в виде вершин и ребер между ними позволяет осуществлять расчет расстояния между объектами на основании числа ребер, лежащих между соответствующими вершинами. Кроме того, возникает возможность выявить свойства объекта, которые «объективно» задаются его положением в общей структуре связей между элементами сети [Emirbayer, Goodwin, 1994]. Наконец, появляются основания для предположений о неизвестных характеристиках объекта исходя из близких к нему объектов с известными характеристиками [Golbeck, 2013]. Объекты в сети становится возможным группировать и выделять с помощью различных методов выявления сообществ [Grandjean, 2016].

Одним из ключевых свойств, задаваемых положением объекта в сети непосредственно, является центральность. Центральность – это метрика «важности» вершины в структуре сети и – потенциально – «влияния» вершины на течение процессов, которые происходят в контексте рассматриваемой сети. Существует множество подходов к определению центральности, ключевыми из которых являются «центральность связности», которая опирается исключительно на число вершин, с которыми у данной вершины есть общее ребро; «центральность близости», которая учитывает кратчайшее расстояние между вершиной и самой удаленной от нее вершиной, и «центральность посредничества», которая указывает на то, сколько кратчайших путей между остальными вершинами проходит через данную [Vera, Schupp, 2006].

Помимо того что анализ социальных сетей позволяет исследовать некоторые свойства каждого объекта в сети, важным является исследование свойств сети в целом. Топология и плотность сети (доля наличествующих ребер по отношению к числу всех возможных связей), среднее расстояние между объектами, склонность к группированию пользователей в кластеры, степень гомогенности или гомофилии (проявление тенденции к группировке похожих объектов), а также распределение числа связей и центральностей – все это является важным источником информации о самой выборке объектов и процессах, породивших наблюдаемое распределение связей [Tweeting from left to right ..., 2015].

***Применение в сборе и анализе данных о протестах.*** Поскольку методы обучения без учителя, в первую очередь, нацеле-

ны на выявление структуры в данных, их применимость для сбора информации несколько ограниченна и встречается обычно лишь на ранней, разведывательной, стадии сбора информации. Примером такого подхода является работа Н. Кальдерон и др., в которой на основе большого массива твитов исследовались общественные настроения в Бразилии в период чемпионата мира по футболу 2014 г. [Mixed-initiative social media analytics ..., 2015]. Проведенное в этом исследовании тематическое моделирование опиралось на инструмент визуального анализа IN-SPIRE, который кластеризует документы на основе совместного появления в них одинаковых слов и отображает их двумерном пространстве [IN-SPIRE …, 2004]. По наиболее распространенным в каждом кластере словам определялась тема, объединяющая документы в кластере. Если она не была связана с деятельностью политиков и государственных институтов или отношением к ним, то документы удалялись из рассмотрения. Процесс последовательного выявления тем и их удаления проходил многократно до тех пор, пока не остались лишь те темы, которые необходимы для дальнейшего исследования [Mixed-initiative social media analytics, 2015]. Такой же метод отбора документов можно использовать и посредством латентного размещения Дирихле (*далее* – LDA), и посредством моделей коррелированных тем.

Последние из указанных моделей особенно широко применяются на стадии анализа данных о протестах. Так, например, LDA и модель коррелирующих тем применялись для тематического моделирования твитов в работах К. Мангера и др.[1] [Elites Tweet ..., 2019] и С. Линдгрена [Lindgren, 2019]. В первой из них по результатам LDA была рассчитана энтропия количества тем, о которых писали в твитах во время протестов 2015 г. сторонники и противники Н. Мадуро среди депутатов парламента Венесуэлы. Это позволило оценить изменения информационной повестки дня у сторонников и противников руководства Венесуэлы и описать их коммуникационные стратегии во время протестов [Elites Tweet ..., 2019]. Во второй из этих работ результаты тематического моделирования позволили определить, насколько движение MeToo сохраняло сконцентрированность на своей изначальной повестке с

[1] В этой работе для проверки робастности результатов также использовали модель коррелированных тем (Correlated Topic Model).

течением времени. С. Линдгрен оценил ее как количество тем, вероятность отнесения документов к которым превышает заданный порог [Lindgren, 2019]. В обоих упомянутых исследованиях ученые стремились с помощью латентного размещения Дирихле выявить как можно большее количество тем, чтобы затем рассчитать характеристики их распределения.

LDA может быть использовано не только для определения тем в сообщениях в социальных сетях, но и для группировки аккаунтов по степени их схожести. Такой подход был использован в работе Дж. Ларсон и др. при исследовании влияния сетевого окружения на участие в протестной активности [Social networks ..., 2019]. В этой работе, в частности, с помощью LDA с 12 темами была определена тема каждого из более чем 68 тыс. аккаунтов, на которые были подписаны участники протестных мероприятий и люди схожих с ними политических взглядов, но не посетившие протестные мероприятия. Затем Дж. Ларсон и ее коллеги выяснили, что протестующие и воздержавшиеся от протеста не отличаются по темам аккаунтов, на которые они подписаны. Это может указывать, как считают исследователи, на схожесть уровня интереса к политике у этих двух групп [Social networks ..., 2019].

В отмеченных выше работах тематическое моделирование использовалось для получения обобщенных описательных статистик, характеризующих массив документов, и не требовало детального изучения содержания текстов и выделенных тем. Иное применение тематического моделирования представлено в работе К. Кларка и К. Коцака о роли Твиттера и Фейсбука в первоначальной протестной мобилизации в ходе египетской революции 2011 г. [Clarke, Kocak, 2020]. Авторы не только оценили параметры тематической модели, примененной к массиву сообщений в сети Твиттер, но и содержательно проинтерпретировали полученные темы, опираясь на наиболее часто встречающиеся в них слова. Их исследование выявило, что Твиттер прежде всего использовался для распространения новостей о ходе протестных мероприятий в режиме реального времени, что доказывается преобладанием темы news and updates («новости и обновления») в первый день египетских протестов 2011 г. Роль Фейсбука была выявлена с помощью качественных методов и состояла в рекрутировании протестующих, планировании и координации протестных действий [Clarke, Kocak, 2020].

Для анализа протестной активности с помощью данных из социальных сетей широко используется сетевой анализ (*social network analysis*), который направлен на выявление структуры связей между участниками протестных движений (в том числе и в сравнении с тем, кто не участвует в них). Данные из социальных сетей действительно предоставляют возможности для применения именно методов сетевого анализа, так как в них содержится информация об отношениях между людьми, которую можно представить в виде сетевого графа с набором признаков, описывающих как вершины, так и ребра.

В исследованиях протестной активности применяются разные методы сетевого анализа. Так, в исследовании структуры социальных связей в Твиттере среди участников движения Occupy Wall Street на этапе его зарождения М. Тремайне применял для выявления наиболее влиятельных людей и хэштегов различные метрики центральности, включая степени близости (*closeness*), посредничества (*betweenness*), влияния (*eigenvalue*), полустепени захода и исхода (*indegree and outdegree*) [Tremayne, 2014].

Другим примером современного исследования с применением методов сетевого анализа является уже упоминавшаяся выше работа Дж. Ларсон и др. [Social networks and protest participation ..., 2019]. В ней структура социальных связей в Твиттере среди 764 участников марша (далее – группа протестующих) против терроризма в Париже в январе 2015 г. сравнивалась со структурой социальных связей такого же числа случайно отобранных пользователей Твиттера, использовавших в своих сообщениях те же хэштеги, но не принявших участие в марше (далее – контрольная группа). Для каждого пользователя из каждой группы были найдены пользователи, на которых они были подписаны (далее – сосед первого порядка), и пользователи, на которых были подписаны соседи первого порядка (далее – сосед второго порядка). Было построено две сети. В первую (далее – протестную) сеть вошли пользователи из группы протестующих, а также их соседи первого и второго порядков; во вторую (далее – контрольную) – пользователи из контрольной группы, а также их соседи первого и второго порядков. Как показывают результаты, у пользователей из группы протестующих доля других пользователей из своей группы среди соседей первого и второго порядков в среднем выше, чем у пользователей из контрольной группы. Кроме того, у пользователей из группы

протестующих выше доля триад, содержащих как минимум еще одного пользователя из своей группы. Наконец, в протестной сети количество направленных в обе стороны ребер между пользователями из группы протестующих выше, чем количество таких ребер между пользователями из контрольной группы в контрольной сети. Такие результаты могли быть получены, как считают авторы, если прямые и сильные связи с другими высокомотивированными к участию в протесте людьми повышают ценность протеста для самого индивида и его готовность к участию в нем [Social networks and protest participation ..., 2019][1].

Как показывает этот краткий обзор, методы обучения без учителя активно используются в современных исследованиях протестной активности. Они позволяют решать разнообразные задачи от предварительного отбора необходимых для дальнейшего исследования данных до выявления степени динамики размывания повестки протестных движений и определения наиболее влиятельных движений. Методы обучения без учителя обладают большим потенциалом для применения в ходе исследования протестной активности с помощью данных из социальных сетей.

## Заключение

Представленный в данной работе обзор применения методов машинного обучения к исследованию процессов политической мобилизации акцентирует внимание на двух основных классах методов: обучении с учителем и без учителя (к числу последних отнесены также некоторые методы анализа социальных сетей). За рамками обзора остались методы с частичным обучением (*semi-supervised learning*), в разной форме сочетающие характеристики обучения с учителем и без него [Zhu, Goldberg, 2009]. Это молодое направление, которое пока не получило широкого применения в политиче-

[1] Отметим, однако, важную особенность данного исследования, затрудняющую перенесение полученных в нем выводов на многие другие протестные акции: марш против терроризма в Париже в январе 2015 г. не являлся протестом против руководства страны и / или политического режима; во главе марша стоял сам Президент Франции Ф. Олланд. См.: Погибших в парижских терактах почтили минутой молчания // Коммерсантъ. – 2015. – 11 января. – Режим доступа: https://www.kommersant.ru/doc/2644377 (дата посещения: 31.08.2020).

ской науке вообще и исследовании политической мобилизации в частности, но которое потенциально интересно в типичных для политологических исследований ситуации, когда количество размеченных данных крайне ограниченно, а неразмеченных – велико.

Другое направление исследований, практически не затронутое в данном обзоре и до сих пор остающееся в значительной степени белым пятном в прикладных политологических исследованиях, – это использование машинного обучения для каузального анализа (исследования причинно-следственных связей). В этом направлении машинное обучение может быть использовано на тех этапах исследования, на которых требуется решить задачу прогнозирования (например, на первом шаге оценивания регрессионных моделей методом инструментальных переменных).

Наконец, в данной работе не рассматривается широкий и динамично развивающийся класс методов обработки естественного языка, опирающихся на векторное представление слов (*word embeddings*) как способ квантификации их значения [Distributed representations ..., 2013; Pennington, Socher, Manning, 2014]. В политологической литературе данные методы в последние годы получили свое применение в рамках исследования идеологических расколов на основе массивов текстовых данных о парламентских дебатах [Rheault, Cochrane, 2020]. Их применение в исследованиях политической мобилизации (например, для выявления различий в контекстах использования имен политиков, названий институтов и организаций) еще ждет своих исследователей.

В целом растущий объем доступных для анализа данных позволяет ожидать, что распространение методов машинного обучения при исследовании политической мобилизации и других вопросов политической науки будет расти, что превращает их в важную и перспективную составляющую методологического арсенала современного политолога.

## Список литературы

*Ахременко А.С., Стукал Д.К., Петров А.П.* Сеть или текст? Факторы распространения протеста в социальных медиа: теория и анализ данных // Полис. Политические исследования. – 2020. – № 2. – С. 73–91. – DOI: https://doi.org/10.17976/jpps/2020.02.06

*Иберла К.* Факторный анализ / пер. с нем. В.М. Ивановой. – М. : Статистика, 1980. – 398 с.

*Azar E.* The conflict and peace data bank (COPDAB) project // Journal of Conflict Resolution. – 1980. – Vol. 24, N 1. – P. 143–152. – DOI: https://doi.org/10.1177/002200278002400106

*Basilevsky A.* Statistical factor analysis and related methods. – N.Y. : Wiley, 1994. – 759 p.

Big data, social media, and protest: foundations for a research agenda // Computational social science: discovery and prediction / J. Tucker [et al.]; M. Alvarez (ed.). – N.Y. : Cambridge University Press, 2016. – P. 199–224.

*Blei D., Lafferty J.* A correlated topic model of science // Annals of Applied Statistics. – 2007. – Vol. 1, N 1. – P. 17–35. – DOI: https://doi.org/10.1214/07-aoas114

*Blei D., Ng A., Jordan M.* Latent Dirichlet Allocation // Journal of Machine Learning Research. – 2003. – Vol. 3, N 3. – P. 993–1022.

*Boschee E., Natarajan P., Weischedel R.* Automatic extraction of events from open source text for predictive forecasting // Handbook of computational approaches to counterterrorism / V. Subrahmanian (ed). – N.Y. : Springer, 2013. – P. 51–67.

*Casas A., Williams N.* Images that matter: online protests and the mobilizing role of pictures // Political Research Quarterly. – 2019. – Vol. 72, N 2. – P. 360–375. – DOI: https://doi.org/10.1177/1065912918786805

*Cioffi-Revilla C.* Computational social science // Wiley interdisciplinary reviews: Computational statistics. – 2010. – Vol. 2, N 3. – P. 259–271. – DOI: https://doi.org/10.1002/wics.95

*Clarke K., Kocak K.* Launching revolution : social media and the Egyptian uprising's first movers // British Journal of Political Science. – 2020. – Vol. 50, N 3. – P. 1025–1045. – DOI: https://doi.org/10.1017/s0007123418000194

Cluster analysis / B. Everitt, S. Landau, M. Leese, D. Stahl. – Chichester : Wiley, 2011. – 330 p.

Comparing GDELT and ICEWS Event Data / M. Ward, A. Beger, J. Cutler, [et al]. – 2013. – 10 p. – Mode of access: https://www.researchgate.net/profile/Andreas_Beger2/publication/303211430_Comparing_GDELT_and_ICEWS_event_data/links/57f7d9bb08ae886b89836115/Comparing-GDELT-and-ICEWS-event-data.pdf (accessed: 19.10.2020).

Computational social science / D. Lazer, A. Pentland, L. Adamic, [et al.] // Science. – 2009. – Vol. 323, N 5915. – P. 721–723. – DOI: https://doi.org/10.1126/science.1167742

*Croicu M., Weidmann N.* Improving the selection of news reports for event coding using ensemble classification // Research and Politics. – 2015. – Vol. 2, N 4. – P. 1–8. – DOI: https://doi.org/10.1177/2053168015615596

*Diamond L.* Liberation technology // Journal of Democracy. – 2010. – Vol. 21, N 3. – P. 69–83. – DOI: https://doi.org/10.1353/jod.0.0190

Distributed representations of words and phrases and their compositionality / T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean // Proceedings of the 26 th International Conference on Neural Information Processing Systems. – Neural Information Processing Systems Foundation, 2013. – P. 3111–3119. – Mode of access:

https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf (accessed: 19.10.2020).

Elites Tweet to get feet off the streets: measuring regime social media strategies during protest / K. Munger, R. Bonneau, J. Nagler, J. Tucker // Political science research and methods. – 2019. – Vol. 7, N 4. – P. 815–834. – DOI: https://doi.org/10.1017/psrm.2018.3

*Emirbayer M., Goodwin J.* Network analysis, culture, and the problem of agency // American journal of sociology – 1994. – Vol. 99, N 6. – P. 1411–1454. – DOI: https://doi.org/10.1086/230450

*Enikolopov R., Makarin A., Petrova M.* Social media and protest participation: evidence from Russia // Econometrica. – 2020. – Vol. 88, N 4. – P. 1479–1514. – DOI: https://doi.org/10.3982/ecta14281

Evaluation methods for topic models / H. Wallach, I. Murray, R. Salakhutdinov, D. Mimno // ICML'09: Proceedings of the 26 th Annual International Conference on Machine Learning. – 2009. – P. 1105–1112 – Mode of access: https://mimno.infosci.cornell.edu/papers/wallach09evaluation.pdf (accessed: 19.10.2020).

*Freedman D.* Statistical models and shoe leather // Sociological Methodology – 1991. – Vol. 21 – P. 291–313. – DOI: https://doi.org/10.2307/270939

*Golbeck J.* Analyzing the social web. – Amsterdam : Morgan Kaufmann, 2013. – 290 p.

*Grandjean M.* A social network analysis of Twitter: Mapping the digital humanities community // Cogent Arts & Humanities. – 2016. – Vol. 3, N 1. – P. 1–14. – DOI: https://doi.org/10.1080/23311983.2016.1171458

*Griffiths T., Steyvers M.* Probabilistic topic models // Latent Semantic Analysis: A Road to Meaning / E. Laurence, D. Landauer, S. McNamara, D. Kintsch (eds). – Mahwah, NJ : Laurence Erlbaum, 2006. – P. 427–448.

*Hafner-Burton E., Kahler M., Montgomery A.* Network analysis for international relations // International organization. – 2009. – Vol. 63, N 3. – P. 559–592. – DOI: https://doi.org/10.1017/s0020818309090195

*Hastie T., Tibshirani R., Friedman J.* The elements of statistical learning: data mining, inference, and prediction. – N.Y. : Springer, 2016 b. – 745 p. – DOI: https://doi.org/10.1007/978-0-387-84858-7

*Hastie T., Tibshirani R., Wainwright M.* Statistical learning with sparsity: the lasso and generalizations. – Boca Rato : CRC Press, 2016 a. – 354 p.

Hierarchical Dirichlet processes / Y. Teh, M. Jordan, M. Beal, D. Blei // Journal of the American Statistical Association. – 2006. – Vol. 101, N 476. – P. 1566–1581. – DOI: https://doi.org/10.1198/016214506000000302

*Hoerl E., Kennard R.* Ridge regression: biased estimation for nonorthogonal problems // Technometrics. – 1970. – Vol. 12, N 1. – P. 55–67. – DOI: https://doi.org/10.1080/00401706.1970.10488634

IN-SPIRE InfoVis 2004 Contest Entry / P. Wong, E. Hetzler, S. Posse, M. Whiting, S. Havre, N. Cramer, A. Shah, M. Singhal, A. Turner, J. Thomas // IEEE Symposium on Information Visualization. InfoViz. – 2004. – 2 p. – Mode of access: https://www.cs.umd.edu/hcil/InfovisRepository/contest-2004/3/PNNLsummary2004.pdf (accessed: 19.10.2020).

Integrated data for events analysis (IDEA): an event typology for automated events data development / D. Bond, J. Bond, C. Oh, J. Jenkins, C.L. Taylor // Journal of peace research. – 2003. – Vol. 40, N 6. – P. 733–745. – DOI: https://doi.org/10.1177/00223433030406009

*Kadushin C.* Too much investment in social capital? // Social networks. – 2004. – Vol. 1, N 26. – P. 75–90. – DOI: https://doi.org/10.1016/j.socnet.2004.01.009

*Kadushin C.* Understanding social networks: Theories, concepts, and findings. – Oxford : Oxford university press, 2012. – 264 p.

*Krueger J., Lewis-Beck M.* Is OLS dead? // The political methodologist. – 2008. – Vol. 15, N 2. – P. 2–4.

*Lankina T., Tertytchnaya K.* Protest in electoral autocracies: a new dataset // Post-Soviet affairs. – 2020. – Vol. 36, N 1. – P. 20–36. – DOI: https://doi.org/10.1080/1060586x.2019.1656039

*Leetaru K., Schrodt P.* Gdelt: Global data on events, location, and tone, 1979–2012 // ISA annual convention. – 2012. – Vol. 2. – P. 1–49.

*Lindgren S.* Movement mobilization in the age of hashtag activism: examining the challenge of noise, hate, and disengagement in the #MeToo campaign // Policy and Internet. – 2019. – Vol. 11, N 4. – P. 418–438. – DOI: https://doi.org/10.1002/poi3.212

*McClelland C.* World event/interaction survey codebook (ICPSR 5211). – Ann Arbor : University consortium for political and social research, 1976. – 22 p.

Mixed-initiative social media analytics at the World Bank. Observations of citizen sentiment in Twitter Data to explore “Trust” of political actors and state institutions and its relationship to social protest / N. Calderon, B. Fisher, J. Hemsley, B. Ceskavich, G. Jansen, R. Marciano, V. Lemieux // 2015 IEEE International Conference on Big Data (Big Data). – Santa Clara, CA : IEEE, 2015. – P. 1678–1687. – DOI: https://doi.org/10.1109/BigData.2015.7363939

*Molina M., Garip F.* Machine learning for sociology // Annual review of sociology. – 2019. – Vol. 45. – P. 27–45. – DOI: https://doi.org/10.1146/annurev-soc-073117-041106

*Mourão R., Chen W.* Covering protests on Twitter: The influences on journalists’ social media portrayals of Left- and Right-Leaning demonstrations in Brazil // The International Journal of Press/Politics. – 2020. – Vol. 25, N 2. – P. 260–280. – DOI: https://doi.org/10.1177/1940161219882653

*O’Brien S.* Crisis early warning and decision support: contemporary approaches and thoughts on future research // International studies review. – 2010. – Vol. 12, N 1. – P. 87–104. – DOI: https://doi.org/10.1111/j.1468-2486.2009.00914.x

*Open Event Data Alliance.* PLOVER: Political language ontology for verifiable event records. Event, actor and data interchange specification. – 2020. – Mode of access: https://github.com/openeventdata/PLOVER/blob/master/PLOVER_MANUAL.pdf (accessed: 19.10.2020).

Pachinko Prediction: A Bayesian method for event prediction from social media data / J. Tuke, A. Nguyen, M. Nasim, D. Mellor, A. Wickramasinghe, N. Bean, L. Mitchell // Information processing and management. – 2020. – Vol. 57, N 2. – P. 1–13. – DOI: https://doi.org/10.1016/j.ipm.2019.102147

*Padjett J., Ansell C.* Robust Action and the Rise of the Medici // American Journal of Sociology. – 1993. – Vol. 98, N 6. – P. 1259–1319. – DOI: https://doi.org/10.1086/230190

*Pan J., Siegel A.* How Saudi crackdowns fail to silence online dissent // American Political science review. – 2020. – Vol. 114, N 1. – P. 109–125. – DOI: https://doi.org/10.1017/s0003055419000650

*Pennington J., Socher R., Manning C.* Glove: global vectors for word representation // Conference on Empirical methods in natural language processing (EMNLP). – Association for computational linguistics, 2014. – P. 1532–1543.

Predicting online protest participation of social media users / S. Ranganath, F. Morstatter, X. Hu, J. Tang, S. Wang, H. Liu // 30 th AAAI Conference on artificial intelligence, AAAI 2016. – Phoenix : AAAI press, 2016. – P. 208–214.

Reading tea leaves: how humans interpret topic models / J. Chang, J. Boyd-Graber, S. Gerrish, W. Chong, D. Blei // Advances in neural information processing systems 22 (neural information processing systems 2009) / Y. Bengio [et al.] (eds). – 2009. – Mode of access: https://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf (accessed: 19.10.2020).

*Rheault L., Cochrane C.* Word embeddings for the analysis of ideological placement in parliamentary corpora // Political analysis. – 2020. – Vol. 22, N 1. – P. 112–133. – DOI: https://doi.org/10.1017/pan.2019.26

*Roberts M., Brandon S., Dustin T.* Navigating the Local Modes of Big Data: The Case of Topic Models // Data Analytics in Social Science, Government, and Industry / M. Alvarez (ed.). – N.Y. : Cambridge University Press, 2016. – P. 51–97.

*Schrodt P., Gerner D., Yilmaz O.* Conflict and mediation event observations (CAMEO) : an event data framework for a post Cold War world // International conflict mediation: new approaches and findings / J. Bercovitch, S. Gartner (eds). – N.Y. : Routledge, 2009. – P. 287–304.

*Schrodt P., Van Brackle D.* Automated Coding of Political Event Data // Handbook of Computational Approaches to Counterterrorism / V. Subrahmanian (ed.). – N.Y. : Springer, 2013. – P. 23–49.

Social networks and protest participation: evidence from 130 million Twitter users / *J. Larson, J. Nagler, J. Ronen, J. Tucker* // American Journal of Political Science. – 2019. – Vol. 63, N 3. – P. 690–705. – DOI: https://doi.org/10.1111/ajps.12436

*Steenbergen M., Bradford J.* Modeling multilevel data structures // American Journal of Political Science. – 2002. – Vol. 46, N 1. – P. 218–237. – DOI: https://doi.org/10.2307/3088424

Structural topic models for open-ended survey responses / M. Roberts, S. Brandon, T. Dustin, C. Lucas, L.-L. Jetson, S. Gadarian, B. Albertson, D. Rand // American journal of political science. – 2014. – Vol. 58, N 4. – P. 1064–1082. – DOI: https://doi.org/10.1111/ajps.12103

*Sundberg R., Melander E.* Introducing the UCDP georeferenced event dataset // Journal of peace research. – 2013. – Vol. 50, N 4. – P. 523–532. – DOI: https://doi.org/10.1177/0022343313484347

The structural topic model and applied social science / *M. Roberts, S. Brandon, T. Dustin, E. Airoldi* // Advances in neural information processing systems workshop

on topic models: computation, application, and evaluation. – NIPS, 2013. – Mode of access: https://scholar.princeton.edu/files/bstewart/files/stmnips2013.pdf (accessed: 19.10.2020).

*Tibshirani R.* Regression Shrinkage and selection via the Lasso // Journal of the Royal statistical society. Series B (methodological). – 1996. – Vol. 58, N 1. – P. 267–288. – DOI: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

*Tremayne M.* Anatomy of protest in the digital era: a network analysis of Twitter and Occupy Wall Street // Social movement studies. – 2014. – Vol. 13, N 1. – P. 110–126. – DOI: https://doi.org/10.1080/14742837.2013.830969

Tweeting from left to right: Is online political communication more than an echo chamber? / *P. Barberá, J.T. Jost, J. Nagler, J.A. Tucker, R. Bonneau* // Psychological science. – 2015. – Vol. 26, N 10. – P. 1531–1542. – DOI: https://doi.org/10.1177/0956797615594620

*Vera E., Schupp T.* Network analysis in comparative social sciences // Comparative Education. – 2006. – Vol. 42, N 3. – P. 405–429. – DOI: https://doi.org/10.1080/03050060600876723

ViEWS: A political violence early-warning system / H. Hegre, M. Allansson, M. Basedau, [et al.] // Journal of Peace Research. – 2019. – Vol. 56, N 2. – P. 155–174. – DOI: https://doi.org/10.1177/0022343319823860

*Weidmann N., Rød E.* The Internet and political protest in autocracies. Chapter 4 // Weidmann N., Rød E. The Internet and Political Protest in Autocracies. – N.Y. : Oxford university press, 2019. – P. 39–62.

*Wooldridge J.* Econometric analysis of cross section and panel data. – Cambridge : The MIT Press, 2002. – 741 p.

*Zhu X., Goldberg A.* Introduction to semi-supervised learning // Synthesis lectures on artificial intelligence and machine learning. – 2009. – Vol. 3, N 1. – P. 1–130. – DOI: https://doi.org/10.2200/S00196ED1V01Y200906AIM006

*Zhuravskaya E., Petrova M., Enikolopov R.* Political effects of the internet and social media // Annual review of economics. – 2020. – Vol. 12. – P. 415–438. – DOI: https://doi.org/10.1146/annurev-economics-081919-050239

## D.K. Stukal, V.E. Belenkov, I.B. Philippov*
## Data science methods in political science research: analyzing protest activity in social media[1]

*Abstract.* The advent of social media and increased digitization of social processes have had a dramatic impact on politics and, particularly, on political mobilization

* **Stukal Denis**, HSE University (Moscow, Russia), e-mail: dstukal@hse.ru; **Belenkov Vadim**, HSE University; Moscow State Institute of International Relations, MFA Russia (Moscow, Russia), e-mail: vadim.belenkov@gmail.com; **Philippov Ilya**, HSE University (Moscow, Russia), e-mail: ibfilippov@gmail.com

[1] This research was supported by RSF (project No. 20-18-00274), HSE University.

and communication. The political science methodology and toolkit have also adapted to these changes and absorbed a variety of new approaches and methods from the burgeoning field of data science. This paper provides an overview of some of the key methodological innovations to the political science toolkit drawn from data science and discusses the advantages and limitations of these new methods for studying protest activity and political mobilization in social media. We focus on supervised and unsupervised learning as two major groups of methods that can be applied to either facilitate data collection in almost real time or the analysis of big data on protest activity. We discuss overfitting, regularization, and hyperparameter selection via cross-validation in the context of supervised methods, and present topic modeling and social network analysis techniques within unsupervised methods. The strengths and weaknesses of these methods are illustrated with references to recent articles published in peer-reviewed journals. We conclude the paper with a discussion of the emerging methods that have not been used in political mobilization research yet and are open for further exploration by political scientists.

*Keywords*: political mobilization; protest; social media; machine learning; data science; supervised learning; unsupervised learning; computational social sciences

*For citation:* Stukal D.K., Belenkov V.E., Philippov I.B. Data science methods in political science research: analyzing protest activity in social media. *Political science (RU)*. 2021, N 1, P. 46–75. DOI: http://www.doi.org/10.31249/poln/2021.01.02

## References

Akhremenko A.S., Stukal D.K., Petrov A.P. Network vs message in protest diffusion on social media: theoretical and data analytics perspectives. *Polis. Political Studies*. 2020, N 2. P. 73–91. DOI: https://doi.org/10.17976/jpps/2020.02.06 (In Russ.)

Azar E. The conflict and peace data bank (COPDAB) project. *Journal of Conflict Resolution*. 1980, Vol. 24, N 1, P. 143–152. DOI: https://doi.org/10.1177/002200278002400106

Barberá P., Jost J.T., Nagler J., et al. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*. 2015, Vol. 26, N 10, P. 1531–1542. DOI: https://doi.org/10.1177/0956797615594620

Basilevsky A. *Statistical factor analysis and related methods*. New York : Wiley, 1994, 759 p.

Blei D, Lafferty J. A correlated topic model of science. *Annals of Applied Statistics*. 2007, Vol. 1, N 1, P. 17–35. DOI: https://doi.org/10.1214/07-aoas114

Blei D., Ng A., Jordan M. Latent dirichlet allocation. *Journal of machine learning research*. 2003, Vol. 3, N 3, P. 993–1022.

Bond D., Bond J., Oh C. et al. Integrated data for events analysis (IDEA): An event typology for automated events data development. *Journal of peace research*. 2003, Vol. 40, N. 6, P. 733–745. DOI: https://doi.org/10.1177/00223433030406009

Boschee E, Natarajan P., Weischedel R. Automatic extraction of events from open source text for predictive forecasting. In: Subrahmanian V. (ed.). *Handbook of computational approaches to counterterrorism*. New York : Springer, 2013, P. 51–67.

Calderon N., Fisher B., Hemsley J., et al. Mixed-initiative social media analytics at the World Bank. Observations of citizen sentiment in Twitter Data to explore "Trust" of political actors and state institutions and its relationship to social protest, *2015 IEEE International Conference on Big Data (Big Data)*. Santa Clara, CA : IEEE, 2015, P. 1678–1687. DOI: https://doi.org/10.1109/BigData.2015.7363939

Casas A., Williams N. Images that Matter: Online Protests and the Mobilizing Role of Pictures. *Political Research Quarterly*. 2019, Vol. 72, N 2, P. 360–375. DOI: https://doi.org/10.1177/1065912918786805

Chang J., Boyd-Graber J., Gerrish S., et al. Reading Tea Leaves: How Humans Interpret Topic Models. In: Bengio Y. et al. (eds). *Advances in neural information processing systems 22 (neural information processing systems 2009)*. 2009. Mode of access: https://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf (accessed: 19.10.2020).

Cioffi-Revilla C. Computational Social Science. *Wiley interdisciplinary reviews: Computational Statistics*. 2010, Vol. 2, N 3, P. 259–271. DOI: https://doi.org/10.1002/wics.95

Clarke K., Kocak K. Launching revolution: social media and the Egyptian uprising's first movers. *British Journal of Political Science*. 2020, Vol. 50, N 3, P. 1025–1045. DOI: https://doi.org/10.1017/s0007123418000194

Croicu M, Weidmann N. Improving the selection of news reports for event coding using ensemble classification. *Research and politics*. 2015, Vol. 2, N 4, P. 1–8. DOI: https://doi.org/10.1177/2053168015615596

Diamond L. Liberation technology. *Journal of democracy*. 2010, Vol. 21, N 3, P. 69–83. DOI: https://doi.org/10.1353/jod.0.0190

Emirbayer M., Goodwin J. Network analysis, culture, and the problem of agency. *American journal of sociology*. 1994, Vol. 99, N 6, P. 1411–1454. DOI: https://doi.org/10.1086/230450

Enikolopov R., Makarin A., Petrova M. Social media and protest participation: evidence from Russia. *Econometrica*. 2020, Vol. 88, N 4, P. 1479–1514. DOI: https://doi.org/10.3982/ecta14281

Everitt B., Landau S., Leese M. et al. *Cluster analysis*. Chichester : Wiley, 2011, 330 p.

Freedman D. Statistical models and shoe leather. *Sociological Methodology*. 1991, Vol. 21, P. 291–313. DOI: https://doi.org/10.2307/270939

Golbeck J. *Analyzing the social web*. Amsterdam : Morgan Kaufmann, 2013, 290 p.

Grandjean, M. A social network analysis of Twitter: Mapping the digital humanities community. *Cogent Arts & Humanities*. 2016, Vol. 3, N 1, P. 1–14. DOI: https://doi.org/10.1080/23311983.2016.1171458

Griffiths T., Steyvers M. Probabilistic topic models. In: Laurence E., Landauer D., McNamara S. Kintsch D. (eds). *Latent Semantic Analysis: A Road to Meaning*. Mahwah, NJ : Laurence Erlbaum, 2006, P. 427–448.

Hafner-Burton E., Kahler M., Montgomery A. Network analysis for international relations. *International organization*. 2009, Vol. 63, N 3, P. 559–592. DOI: https://doi.org/10.1017/s0020818309090195

Hastie T., Tibshirani R., Wainwright M. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Rato : CRC Press, 2016 a, 354 p.

Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York : Springer, 2016 b, 745 p. DOI: https://doi.org/10.1007/978-0-387-84858-7

Hegre H., Allansson M., Basedau M., et al. ViEWS: A political violence early-warning system. *Journal of Peace Research*. 2019, Vol. 56, N. 2, P. 155–174. DOI: https://doi.org/10.1177/0022343319823860

Hoerl E., Kennard R. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*. 1970, Vol. 12, N. 1, P. 55–67. DOI: https://doi.org/10.1080/00401706.1970.10488634

Iberla K. *Factor analysis*. Moscow : Statistika, 1980, 398 p. (In Russ.)

Kadushin C. Too much investment in social capital? *Social networks*. 2004, Vol. 1, N 26, P. 75–90. DOI: https://doi.org/10.1016/j.socnet.2004.01.009

Kadushin C. *Understanding social networks: theories, concepts, and findings*. Oxford : Oxford university press, 2012, 264 p.

Leetaru K., Schrodt P. Gdelt: Global data on events, location, and tone, 1979–2012. *ISA annual convention*. 2012, Vol. 2, P. 1–49.

Krueger J., Lewis-Beck M. Is OLS dead? *The Political Methodologist*. 2008, Vol. 15, N 2, P. 2–4.

Lankina T., Tertytchnaya K. Protest in electoral autocracies: A new dataset. *Post-Soviet affairs*. 2020, Vol. 36, N 1, P. 20–36. DOI: https://doi.org/10.1080/1060586x.2019.1656039

Larson J., Nagler J., Ronen J., et al. Social networks and protest participation: evidence from 130 million Twitter users. *American journal of political science*. 2019, Vol. 63, N 3, P. 690–705. https://doi.org/10.1111/ajps.12436

Lazer D., Pentland A., Adamic L., et. al. Computational social science. *Science*. 2009, Vol. 323, N 5915, P. 721–723. DOI: https://doi.org/10.1126/science.1167742

Lindgren S. Movement mobilization in the age of hashtag activism: examining the challenge of noise, hate, and disengagement in the #MeToo campaign. *Policy and Internet*. 2019, Vol. 11, N 4, P. 418–438. DOI: https://doi.org/10.1002/poi3.212

McClelland C. *World event/interaction survey codebook (ICPSR 5211)*. Ann Arbor : University Consortium for Political and Social Research, 1976, 22 p.

Mikolov T., Sutskever I., Chen K., et al. Distributed representations of words and phrases and their compositionality. In: *Proceedings of the 26 th International Conference on Neural Information Processing Systems. Neural Information Processing Systems Foundation*. 2013, P. 3111–3119. Mode of access: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf (accessed: 19.10.2020).

Molina M., Garip F. Machine learning for sociology. *Annual Review of Sociology*. 2019, Vol. 45, P. 27–45. DOI: https://doi.org/10.1146/annurev-soc-073117-041106

Mourão R., Chen W. Covering protests on Twitter: the influences on journalists' social media portrayals of left- and right-leaning demonstrations in Brazil. *The international journal of press/politics*. 2020, Vol. 25, N 2, P. 260–280. DOI: https://doi.org/10.1177/1940161219882653

Munger K., Bonneau R., Nagler J., et al. Elites Tweet to get feet off the streets: measuring regime social media strategies during protest. *Political Science Research and Methods*. 2019, Vol. 7, N 4, P. 815–834. DOI: https://doi.org/10.1017/psrm.2018.3

O'Brien S. Crisis early warning and decision support: contemporary approaches and thoughts on future research. *International Studies Review*. 2010, Vol. 12, N 1, P. 87–104. DOI: https://doi.org/10.1111/j.1468-2486.2009.00914.x

Open Event Data Alliance. PLOVER: Political Language Ontology for Verifiable Event Records. Event, Actor and Data Interchange Specification. 2020. Mode of access: https:.github.com/openeventdata/PLOVER/blob/master/PLOVER_MANUAL.pdf (accessed: 19.10.2020).

Padjett J., Ansell C. Robust Action and the Rise of the Medici. *American Journal of Sociology*. 1993, Vol. 98, N. 6, P. 1259–1319, DOI: https://doi.org/10.1086/230190

Pan J., Siegel A. How Saudi crackdowns fail to silence online dissent. *American Political Science Review*. 2020, Vol. 114, N 1, P. 109–125. DOI: https://doi.org/10.1017/s0003055419000650

Pennington J., Socher R., Manning C. Glove: global vectors for word representation. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014, P. 1532–1543.

Ranganath S., Morstatter F., Hu X., Tang J., Wang S., Liu H. Predicting online protest participation of social media users. *30 th AAAI Conference on Artificial Intelligence, AAAI 2016*. Phoenix : AAAI press, 2016, P. 208–214.

Rheault L., Cochrane C. Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political analysis*. 2020, Vol. 28, N 1, P. 112–133. DOI: https://doi.org/10.1017/pan.2019.26

Roberts M., Brandon S., Dustin T. Navigating the local modes of big data: the case of topic models. In: Alvarez M. (ed.) *Data Analytics in Social Science, Government, and Industry*. New York : Cambridge University Press, 2016, P. 51–97.

Roberts M., Brandon S., Dustin T., et al. The structural topic model and applied social science. In: *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*. NIPS, 2013. Mode of access: https://scholar.princeton.edu/files/bstewart/files/stmnips2013.pdf (accessed: 19.10.2020).

Roberts M., Brandon S., Dustin T., et al. Structural topic models for open-ended survey responses. *American journal of political science*. 2014, Vol. 58, N 4, P. 1064–1082. DOI: https://doi.org/10.1111/ajps.12103

Schrodt P., Gerner D., Yilmaz O. Conflict and mediation event observations (CAMEO): an event data framework for a post Cold War world. In: Bercovitch J., Gartner S. (eds). *International conflict mediation: new approaches and findings*. Routledge : New York, 2009, P. 287–304.

Schrodt P., Van Brackle D. Automated coding of political event data. In: Subrahmanian V. (ed.) *Handbook of computational approaches to counterterrorism*. New York : Springer, 2013, P. 23–49.

Steenbergen M., Bradford J. Modeling multilevel data structures. *American journal of political science*. 2002, Vol. 46, N 1, P. 218–237. DOI: https://doi.org/10.2307/3088424

Sundberg R., Melander E. Introducing the UCDP Georeferenced Event Dataset. *Journal of peace research*. 2013, Vol. 50, N 4, P. 523–532. DOI: https://doi.org/10.1177/0022343313484347

Tuke J., Nguyen A., Nasim M., et al. Pachinko prediction: a bayesian method for event prediction from social media data. *Information processing and management*. 2020, Vol. 57, N 2, P. 1–13. DOI: https://doi.org/10.1016/j.ipm.2019.102147

Teh Y., Jordan M., Beal M., et al. Hierarchical Dirichlet processes. *Journal of the american statistical association*. 2006, Vol. 101, N 476, P. 1566–1581. DOI: https://doi.org/10.1198/016214506000000302

Tibshirani R. Regression Shrinkage and Selection via the lasso. *Journal of the Royal statistical society. Series B (methodological)*. 1996, Vol. 58, N 1, P. 267–288. DOI: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tremayne M. Anatomy of protest in the digital era: a network analysis of Twitter and Occupy Wall Street. *Social Movement Studies*. 2014, Vol. 13, N 1, P. 110–126. DOI: https://doi.org/10.1080/14742837.2013.830969

Tucker J., Nagler J., Metzger M., et al. Big data, social media, and protest: foundations for a research agenda. In: Alvarez M. (ed). *Computational social science: discovery and prediction*. New York : Cambridge University Press, 2016, P. 199–224.

Vera E., Schupp T. Network analysis in comparative social sciences. *Comparative Education*. 2006, Vol. 42, N 3, P. 405–429. DOI: https://doi.org/10.1080/03050060600876723

Wallach H., Murray I., Salakhutdinov R., et al. Evaluation methods for topic models. *ICML'09: Proceedings of the 26 th Annual International Conference on Machine Learning*. 2009, P. 1105–1112 Mode of acess: https:.mimno.infosci.cornell.edu/papers/wallach09evaluation.pdf (accessed: 19.10.2020).

Ward M., Beger A., Cutler J., et al. Comparing GDELT and ICEWS Event Data. 2013, 10 p. Mode of access: https:.www.researchgate.net/profile/Andreas_Beger2/publication/303211430_Comparing_GDELT_and_ICEWS_event_data/links/57f7d9bb08ae886b89836115/Comparing-GDELT-and-ICEWS-event-data.pdf (accessed: 19.10.2020).

Weidmann N., Rød E. The Internet and Political Protest in Autocracies. Chapter 4. In: Weidmann N., Rød E. *The Internet and Political Protest in Autocracies*. New York: Oxford university press. 2019, P. 39–62.

Wong P., Hetzler E., Posse S., et al. IN-SPIRE InfoVis 2004 Contest Entry. *IEEE Symposium on Information Visualization*. InfoViz. 2004, 2 p. Mode of access: https://www.cs.umd.edu/hcil/InfovisRepository/contest-2004/3/PNNLsummary2004.pdf (accessed: 19.10.2020).

Wooldridge J. *Econometric analysis of cross section and panel data*. Cambridge : The MIT Press. 2002, 741 p.

Zhu X., Goldberg A. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*. 2009, Vol. 3, N 1, P. 1–130. DOI: https://doi.org/10.2200/S00196ED1V01Y200906AIM006

Zhuravskaya E., Petrova M., Enikolopov R. Political effects of the internet and social media. *Annual review of economics*. 2020, Vol. 12, P. 415–438. DOI: https://doi.org/10.1146/annurev-economics-081919-050239