DOI: http://dx.doi.org/10.21686/2413-2829-2024-1-67-75

CC BY

АВТОРСКАЯ МЕТОДИКА ПРОГНОЗИРОВАНИЯ РЕЗУЛЬТАТОВ ВЫБОРОВ НА ОСНОВЕ ИСПОЛЬЗОВАНИЯ БОЛЬШИХ ДАННЫХ

А. С. Козин
Финансовый университет при Правительстве Российской Федерации,
Москва, Россия

Интернет-СМИ предлагают своим читателям возможность высказать мнение по опубликованным темам (комментариям). С этой целью с каждой статьей, размещенной в Интернете, связаны средства коммуникации. Эти комментарии серьезно воспринимаются средствами массовой информации, так как они содержат ценную информацию, позволяющую узнать чувства пользователей Интернета. Социальные сети генерируют огромное количество неструктурированных данных благодаря их использованию тысячами пользователей. Таким образом, наше исследование направлено на обоснование и использование подхода машинного обучения (искусственного интеллекта) для анализа чувств пользователей через комментарии, собранные на YouTube по поводу пресс-конференций, проводимых кандидатами в депутаты. Полученные результаты показывают осуществимость этого подхода, а также автором предложены другие пути его совершенствования.
Ключевые слова: интернет-СМИ, искусственный интеллект, машинное обучение, социальные сети.

THE AUTHOR'S METHODOLOGY OF FORECASTING ELECTION RESULTS ON THE BASIS OF BIG DATA

Anton S. Kozin
Financial University under the Government of the Russian Federation,
Moscow, Russia

Internet mass media provides their readers with an opportunity to express their opinions on topics (comments) being published. For this purpose each article inserted in the Internet is connected with certain means of communication. These comments are taken seriously by mass media as they contain valuable information that helps find out feelings of Internet users. Social networks generate a vast amount of unstructured data due to their use by thousands of users. Therefore, our research aims at substantiation and use of machine learning (artificial intellect) approach in order to analyze users' feelings through comments on YouTube dealing with press-conferences given by nominated candidates. Obtained results show feasibility of the approach. The author proposes other ways of its modernization.
Keywords: Internet mass media, artificial intellect, machine learning, social networks.

Введение

Изобретенный в 1955 г. Джоном Маккарти искусственный интеллект (ИИ) считается областью науки и техники, которая занимается пониманием, использованием компьютера, интеллектуального поведения и созданием искусственных систем, воспроизводящих это поведение. Это симуляция человеческого интеллекта. ИИ часто связан с человеческим прогрессом и современностью. Он проникает в несколько областей, включая журналистику, будь то написание, расследование или проверка информации. ИИ вмешивается в анализ комментариев, размещенных на веб-сайте пользователями Интернета по определенной теме. Таким образом, ИИ «также проявляет себя в от-

ношениях между информационным онлайн-сайтом и его получателями, а также в интерпретации его способов консультаций и практики чтения» [2. – С. 10]. Интернет-СМИ предоставляют своим читателям возможность высказать свое мнение по опубликованным темам. Для этого с каждой статьей, размещенной в Интернете, связаны средства коммуникации (дискуссионный форум, социальные сети). Эти комментарии серьезно воспринимаются средствами массовой информации, поскольку содержат ценную информацию, анализ которой может способствовать определению чувств пользователей Интернета по размещенной теме.

ИИ состоит из анализа текстового контента с онлайн-бирж или понимания преобладающих настроений рынка путем анализа альтернативных наборов данных, таких как статьи и пресс-релизы, квартальные отчеты о прибылях и убытках. Средства коммуникации, такие как социальные сети, генерируют огромные объемы данных в результате их использования тысячами пользователей. Эти данные, подлежащие обработке, являются большими и потенциально неструктурированными. Для анализа этого большого количества неструктурированных данных (комментариев, размещенных в социальных сетях) очень сложно использовать традиционные методы, такие как информационные и коммуникационные науки (ICS). Таким образом, проблема исследования связана исключительно с анализом текстовых материалов (комментариев), которыми обмениваются читатели интернет-СМИ. Цель исследования – авторская методика прогнозирования результатов выборов с использованием больших данных, основанная на анализе настроений пользователей Интернета.

Исследование основано на эпистемологической позиции, называемой междисциплинарной, которая заключается в импорте инструмента анализа данных ИИ. В данном случае будет определена взаимосвязь между ICS и AI (Artificial Intelligence – искусственный интеллект), а затем дано определение ML (машинного обучения), показано применение ML в анализе настроений в контексте онлайн-СМИ (случай выступления кандидатов на YouTube).

Информационные и коммуникационные науки и искусственный интеллект

Взаимосвязь между ICS и AI сложна и характеризуется тремя аспектами. Во-первых, эти две дисциплины имеют общую информацию в качестве общей основы. Таким образом, социальные сети, такие как YouTube, считаются СМИ.

Средства массовой информации «не передают информацию, а создают смысловые пространства в рамках культуры знаков» [4. – С. 50]. Понятие «медиа» относится к двум типам информации: информации в математической природе термина (серия символов, беспорядочная в глазах читателя, и часть технической процедуры), которая способствует передаче определенного количества свойств сообщения и не может быть интерпретирована, и информации, которая представляет собой сообщение или социальную информацию, созданную пользователем средства массовой информации. В Национальном информационном центре термин «информация» основывается, в частности, на данных [9. – С. 68]. В полях, которые рассматривают данные и информацию как сущности, информация транскрибируется в данные и значения. Согласно этому определению данные составляют «материю» информации, где информация играет решающую роль в представлении реальности. Таким образом, мнение, или точка зрения, относится к отношению, мысли или суждению, вдохновленному чувством. Эмоция – это утонченное чувство: тонкая чувствительность, особенно выраженная в произведении искусства, романтическое или ностальгическое чувство, граничащее с сентиментальностью [8; 10].

Во-вторых, ИИ связан с определенными вопросами ICS, такими как информирование и общение, анализ данных, опублико-

ванных в социальных сетях, поиск информации, создание контента и т. д. В этом смысле ИИ можно рассматривать как новую методологию в СНГ.

В-третьих, ICS может изучать социальное использование ИИ и его роль в обществе и, таким образом, способствовать развитию научных знаний, вкладу ИИ в общество и этике его использования, а также методологиям проектирования человеко-машинного интерфейса. Это сбалансированное научное партнерство, которое должно быть построено между двумя дисциплинами [2. – С. 10].

В ИИ для обработки любой информации мы должны использовать массивные наборы многомерных данных, накопленных с беспрецедентной скоростью об отдельных лицах и даже об организациях через Интернет и социальные сети, которые называются большими данными [11. – С. 54]. Таким образом, важность ИИ требует, чтобы мы понимали его не только как набор алгоритмов, но и как технокультуру, где задействованные науки (математика, информационные науки и когнитивные науки) вписаны в учреждения, которые настраивают научную практику, но так же, как и набор, в определенных культурных контекстах и в воображении, включающем идеологии, фикции, чувствительные представления. ИИ не является изолированной областью, поскольку он соединяет несколько дисциплин, таких как информатика, когнитивные науки, нейронаука, гуманитарные и социальные науки, лингвистика.

Зачем использовать машинное обучение?

За последние десять лет машинное обучение вторглось в ИТ-системы компаний. В настоящее время в крупных компаниях, таких как Google, в момент запуска поиска активируется несколько обучающих систем, поэтому одни убеждаются, что вы не робот (вы не мошенник или хакер), другие анализируют ваш профиль и предлагают вам рекламу или контакты. Машинное обучение вмешивается во множество областей. Если такие компании, как онлайн-СМИ, не используют машинное обучение, они не могут не идти в ногу с развитием сферы журналистики [5. – С. 54].

Исследования по машинному обучению выдвигают следующие аргументы, оправдывающие его использование:

1) *неструктурирование обрабатываемых данных*. Обработка данных, извлеченных из социальных сетей, стала очень сложной из-за неструктурированности этих больших данных, их субъективности и неоднородности. Методы машинного обучения использовались для преодоления этой трудности в реальных приложениях [7. – С. 104]. Большие и неструктурированные данные, также называемые большими данными, имеют четко определенные характеристики: более важные, чем у обычных баз данных, варьируются в зависимости от типов и форматов, поступая очень быстрыми темпами. Традиционные методы оценки качества, актуальности и точности данных неприменимы, учитывая, что ценность данных велика;

2) *машинное обучение против традиционного программирования*. Применение традиционного программирования в случае анализа тональности (текстовых комментариев) требует от программиста составления исчерпывающего перечня всех возможных и мыслимых случаев в анализируемых документах. С этой целью этот алгоритм требует каждый раз обновления и добавления текстов на вновь обнаруженные слова в комментариях. Следовательно, он будет содержать тысячи сложных инструкций (правил). Машинное обучение, с другой стороны, представляет собой принципиально иной подход по сравнению с традиционным подходом: машина учится на примерах, а не явно запрограммирована на конкретный результат. Анализатор тональности на основе машинного обучения предоставляет предварительную информацию об отношении читателей онлайн-СМИ и автоматически узнает, какие слова и фразы являются хорошими предикторами.

Применение машинного обучения к анализу комментариев в интернет-СМИ: кейс выступления кандидата от «Единой России»

В рамках исследования был проведен анализ эмоций через комментарии, собранные на YouTube, по поводу пресс-конференций, проведенных кандидатом от «Единой России» [5. – С. 20]. Цель данного анализа – получение информации о мнениях интернет-пользователей по отношению к выступлению кандидата от «Единой России» с использованием ML-подхода, заключающегося в построении модели анализа настроений. Эта модель должна будет учиться на этих собранных данных (корпус обучающих данных) и уметь предсказывать чувства пользователей Интернета, т. е. классифицировать невидимые данные.

Основываясь на методе разработки проекта машинного обучения, были выбраны следующие шаги:

1. *Определение проблемы*.

Построение модели прогнозирования настроений в стихах с речью кандидата от «Единой России» послужит, с одной стороны, для демонстрации того, что подход ML эффективен для гигантских объемов данных и может рассматриваться как новая методология в ICS.

С другой стороны, эта модель будет использоваться для подпитки другой системы ML, такой как система президентских выборов. Эта система будет использовать входную прогнозную информацию и другую информацию, чтобы определить, полезна ли она для прогнозирования результатов выборов.

2. *Извлечение данных*.

На этом этапе выбрали видеоролики, связанные с пресс-конференциями кандидата от «Единой России», и те, которые содержат большое количество взаимодействий с пользователями. Извлечение комментариев осуществляется с сайта YouTube с помощью Python (наиболее часто используемый язык программирования в области искусственного интеллекта) и API YouTube: googleapiclient.discovery. API представляет собой интерфейс прикладного программирования, который позволяет подключать одно программное обеспечение или услугу к другому или сервису для обмена данными и функциями. Команда Python: from google apiclient. discovery import build – позволяет импортировать эту библиотеку для запроса и извлечения комментариев с сайта YouTube. Набор аналитических данных состоит из 600 комментариев. Комментарии сохраняются один за другим в одном файле Excel.

После устранения повторяющихся и пустых комментариев мы получили 500 различных их видов. Для каждого комментария вручную проверяли, были ли полученные комментарии связаны с речью кандидата от «Единой России». Набор данных помечается (ранжируется). Следовательно, мы точно знаем, какие комментарии выражают положительное (3), отрицательное (2), нейтральное (1) и нерелевантное отношение (0). В табл. 1 приведены примеры комментариев с их аннотациями.

Таблица 1

Примеры комментариев с аннотациями к ним

Обратная связь	Аннотации
Противостояние политике	2 (отрицательный)
Посланник народа	2 (отрицательный)
Приветствую Вас	3 (положительный)
Готов помочь	3 (положительный)

Распределения, полученные из различных классов при аннотировании всех комментариев, представлены в табл. 2.

Таблица 2

Положения аннотированных комментариев

Метка	Значение	Количество комментариев
0	Не имеет отношения к делу	60
1	Нейтральный	28
2	Отрицательный	177
3	Положительный	235

Из табл. 2 видно, что количество комментариев к двум меткам 0 и 1 очень мало по сравнению с другими метками, поэтому мы решили оставить только комментарии с метками 2 и 3 для проведения нашего эксперимента. Мы стремимся построить бинарный классификатор для прогнозирования полярности комментария (положительный, отрицательный) о пресс-конференциях, проводимых кандидатом от «Единой России».

3. *Предварительная обработка комментариев*.

Предварительная обработка осуществляется автоматически с помощью программной библиотеки NLTK, которая доступна бесплатно в сети. NLTK (Natural Language ToolKit) – справочная библиотека в области NLP (Natural Language Processing), представляющая собой программную библиотеку на языке программирования Python (используется в области ИИ), позволяющую обрабатывать естественный язык.

Нами были проведены различные предварительные обработки:

– устранены знаки препинания;

– убраны специальные символы и оставлены только буквы;

– убраны диакритические знаки;

– удалены URL-адреса;

– удалены цифры;

– убраны лишние пробелы;

– удалены инструментальные слова или не имеющие смысла слова (стоп-слова);

– удалены эмодзи;

– удалены комментарии, состоящие только из одного слова;

– выполнена токенизация (разрезание текста на фундаментальные единицы – токены).

После этого этапа мы получили аннотированный очищенный корпус из 412 комментариев (235 положительных и 177 отрицательных).

На рис. 1 показано распределение меток в нашем корпусе.

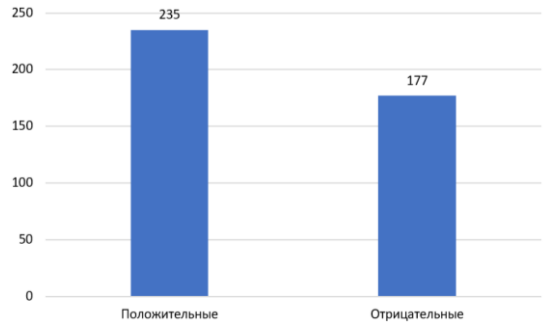

Рис. 1. Распространение меток наборов данных

Этот корпус включает в себя словарь из 2 479 различных терминов. В следующем примере представлены 10 наиболее часто встречающихся терминов в нашем корпусе (каждый термин сопровождается его частотой): президент – 32; народ – 32; страна – 22; да здравствует – 22; министр – 21; хайр – 21; человек – 2; для страны – 2; история – 2; хорошо – 2. Мы видим, что термины «президент», «народ», «страна» являются наиболее цитируемыми. Это может дать нам первую информацию о содержании проанализированных комментариев.

Частотный анализ также позволил заметить, что большинство терминов повторяются только один раз. На рис. 2 показано распределение вхождений в нашем корпусе.

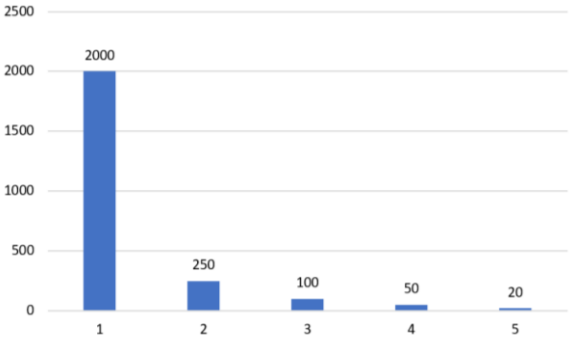

Рис. 2. Распределение вхождений в комментариях

В табл. 3 приводятся данные о размерах комментариев с точки зрения терминов и символов.

Таблица 3

Размер терминов и символов в комментариях

Типы комментариев	Размер терминов и символов
Самый длинный комментарий по количеству терминов	117
Самый короткий комментарий по количеству терминов	2
Самый длинный комментарий по количеству символов	697
Самый короткий комментарий по количеству символов	9

На рис. 3 показано распределение длин по количеству комментариев. Из этого рисунка мы видим, что большинство комментариев короткие, их размер варьируется от 2 до 12 терминов. Для других размеров (от 13 до 117) количество комментариев всегда меньше 10. В противном случае несколько комментариев, не превышающих 10, имеют размер более 13 терминов.

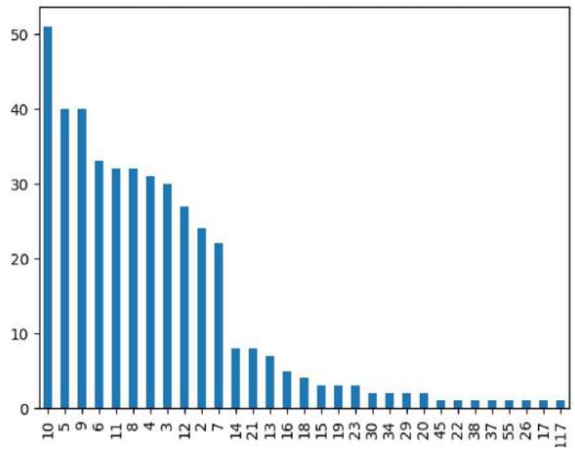

Рис. 3. Распределение длин по количеству комментариев

4. *Классификация и прогнозирование*.

В контексте машинного обучения необходимо протестировать контролируемый подход (при условии, что у нас есть аннотированный корпус). Для этого существует несколько классификаторов (алгоритмов ранжирования) (*K* Ближайшие соседи, машины поддержки векторов (или SVM), Наивный Байес, деревья решений, нейронные сети и др.). Со своей стороны, мы протестировали несколько классификаторов, чтобы определить лучший. Для этого мы использовали две программные библиотеки Python: keras и sklearn.

Keras – это библиотека с открытым исходным кодом для манипулирования нейронными сетями, цель которой – проведение быстрых экспериментов с наборами данных. Sklearn – еще одна библиотека машинного обучения, написанная на Python. Она предназначена для предоставления сборника эффективных реализаций алгоритмов машинного обучения.

На рис. 4 показаны шаги, предпринятые для получения модели классификации комментариев.

При классификации необходимо выполнить следующие процедуры:

а) генерирование из корпуса, предварительно обработанного и аннотированного, трех подкорпусов: один – для обучения (Train), один – для теста (Test) и еще один – для проверки (Validation);

б) определение модели (архитектуры интеллектуальной системы или классификатора);

в) кодирование комментариев;

г) подготовка вектора чувств;

д) оценка модели.

Этот подход повторяется несколько раз до тех пор, пока не будет получена наиболее релевантная проверенная и валидированная модель. На рис. 5 показан классификационный подход. Например, для нейросетевой модели мы используем один скрытый слой с 50 нейронами. Каждый слой представляет собой один нейрон, который предсказывает значение 2 для отрицательных отзывов и значение 3 для положительных отзывов. Сеть управляется параметрами, адаптированными к задачам двоичной классификации. Мы будем следить за точностью при обучении и оценке модели. На рис. 5 показана модель, описанная в предыдущем примере.

Для прогноза вводится комментарий к модели, построенной на предыдущем этапе, и классификатор возвращает результат ее классификации – положительный или отрицательный.

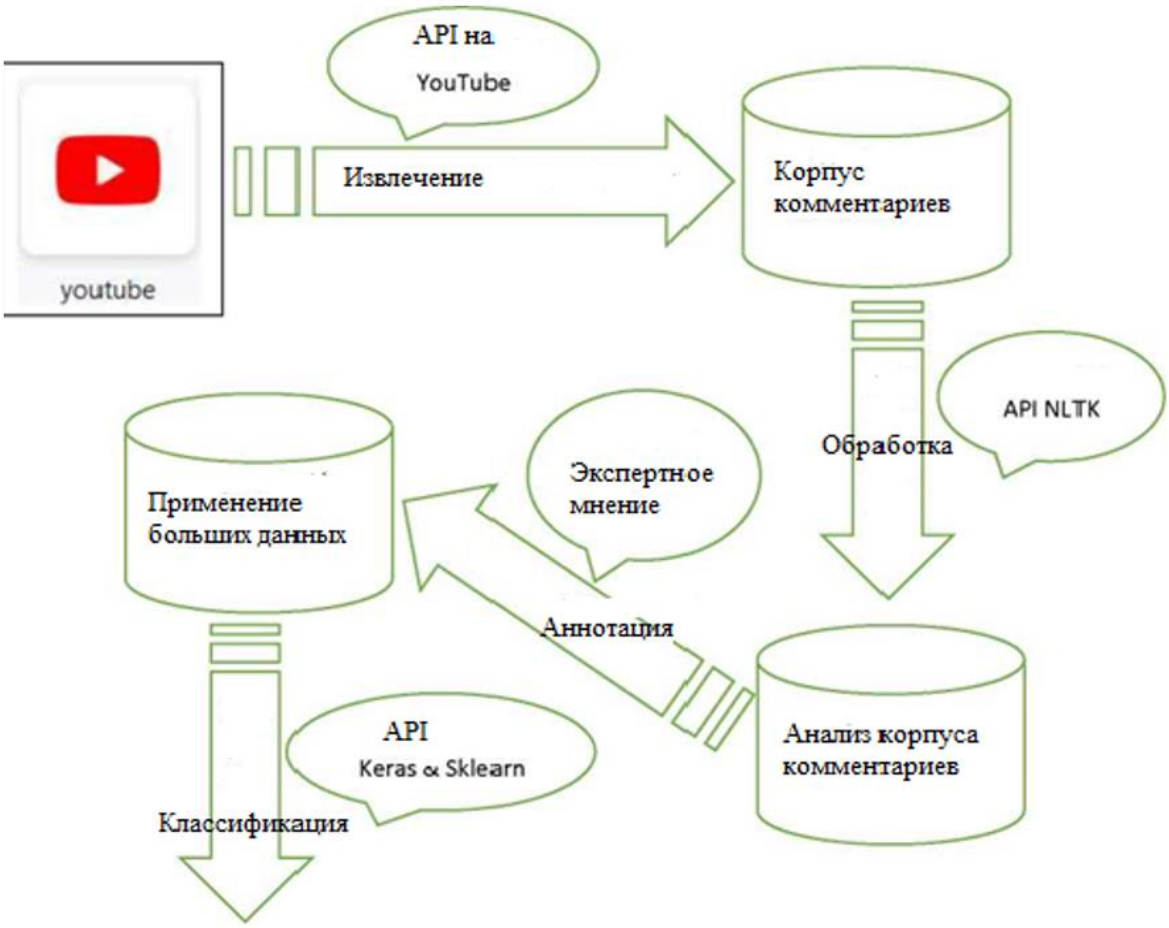

Рис. 4. Этапы классификации комментариев

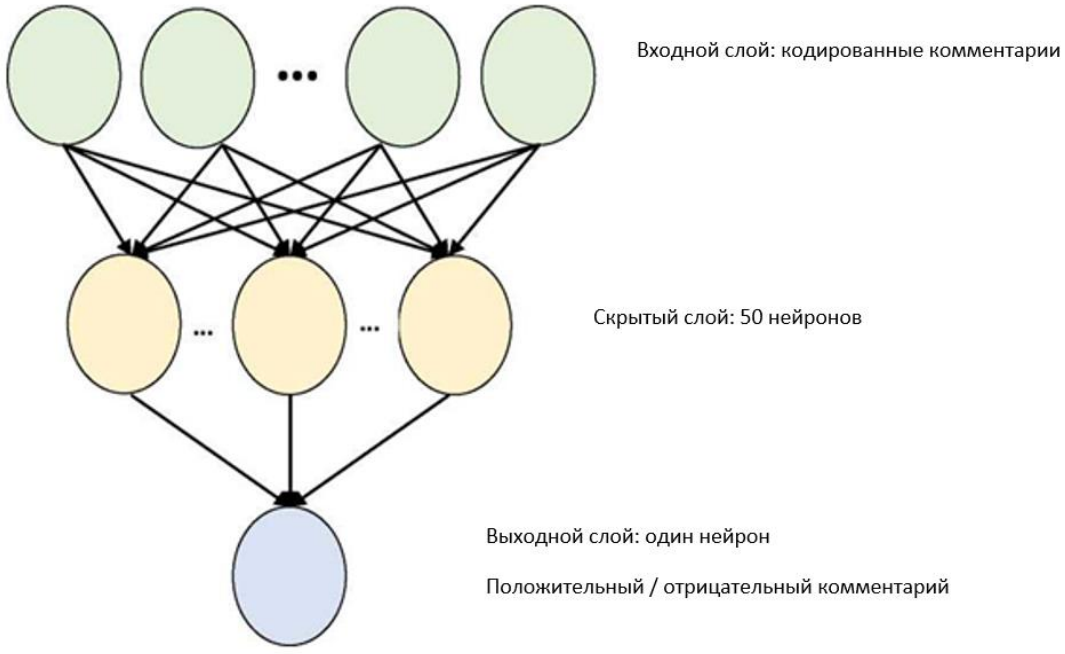

Рис. 5. Пример нейросетевой модели классификации комментариев

Заключение

В ходе исследования показана взаимосвязь между ICS и AI, а также обосновано использование машинного обучения при анализе комментариев в онлайн-СМИ вместо традиционных методов ICS. В перспективе можно предложить оценку других моделей типа Transformers, которые относятся к моделям глубокого обучения и могут преобразовывать набор данных, а также выполнять классификацию данных в случае анализа настроений, например, BERT.

Список литературы

1. *Окунев И. Ю., Горелова Ю. С., Груздева Е. Е.* Региональные особенности электорального поведения в Польше: опыт сравнительного пространственного анализа // Сравнительная политика. – 2021. – № 12 (1). – С. 149–160.

2. *Окунев И. Ю., Жирнова Л. С.* Этнический раскол в электоральном поведении Латвии // Научно-аналитический журнал «Обозреватель – Observer». – 2019. – № 11. – С. 78–90.

3. *Осколков П. В.* Партийная система Эстонии на современном этапе: электоральная турбулентность и смена этнорегиональных паттернов // Балтийский регион. – 2020. – Т. 12. – № 1. – С. 4–15.

4. *Осколков П. В.* Этнические партии и этническое голосование: проблемы концептуализации и факторы развития // Мировая экономика и международные отношения. – 2020. – Т. 64. – № 8. – С. 112–118.

5. *Панов П. В.* Этнические и региональные партии в странах Центральной и Восточной Европы: специфика и траектории развития // Вестник Пермского университета. Политология. – 2020. – Т. 14. – № 3. – С. 20–34.

6. *Сироткина Е. В., Карандашова С. А.* Лояльность элит и выборы глав регионов: роль предвыборных конфликтов в исходе голосования // Полис. Политические исследования. – 2017. – № 6. – С. 76–91.

7. *Солопенко А. В.* Сравнение электорального поведения русскоязычных граждан Латвии и Эстонии на парламентских выборах в 2011–2015 годах // Прибалтийские исследования в России : материалы Международной научной конференции / под ред. А. П. Клемешева, Н. М. Межевича, Г. М. Федорова. – Калининград : Балтийский федеральный университет имени Иммануила Канта, 2016. – С. 163–167.

8. *Тарасов И. Н.* Инструменты этнической политики в Латвии // Южно-российский журнал социальных наук. – 2019. – Т. 20. – № 2. – С. 34–44.

9. *Тарасов И. Н.* Выборы в Польше и Литве в 2020 году // Вызовы и перспективы развития Калининградской области: геополитика и геоэкономика / под ред. Г. М. Федорова. – Калининград : Балтийский федеральный университет имени Иммануила Канта, 2021. – С. 68–70.

10. *Тарасов И. Н., Фидря Е. С.* География электоральной волатильности в Варминьско-Мазурском воеводстве Польши // Балтийский регион. – 2016. – № 8 (4). – С. 78–89.

11. *Уразбаев Е. Е., Ямалова Э. Н.* Этнополитическое движение как проводник институционализации национализма в современной Латвии // Балтийский регион. – 2020. – Т. 12. – № 2. – С. 54–69.

References

1. Okunev I. Yu., Gorelova Yu. S., Gruzdeva E. E. Regionalnye osobennosti elektoralnogo povedeniya v Polshe: opyt sravnitelnogo prostranstvennogo analiza [Regional Features of Election Behavior in Poland: Experience of Comparative Space Analysis]. *Sravnitelnaya politika* [Comparative Policy], 2021, No. 12 (1), pp. 149–160. (In Russ.).

2. Okunev I. Yu., Zhirnova L. S. Etnicheskiy raskol v elektoralnom povedenii Latvii [Ethnic Split in Election Behavior in Latvia]. *Science and Analytical Journal 'Observer'*, 2019, No. 11, pp. 78–90. (In Russ.).

3. Oskolkov P. V. Partiynaya sistema Estonii na sovremennom etape: elektoralnaya turbulentnost i smena etnoregionalnykh patternov [Party System in Estonia on the Current Stage: Election Turbulence and Change in Ethno-Regional Patterns]. *Baltiyskiy region* [Baltic Region], 2020, Vol. 12, No. 1, pp. 4–15. (In Russ.).

4. Oskolkov P. V. Etnicheskie partii i etnicheskoe golosovanie: problemy kontseptualizatsii i faktory razvitiya [Ethnic Parties and Ethnic Voting: Problems of Conceptualization and Factors of Development]. *Mirovaya ekonomika i mezhdunarodnye otnosheniya* [Global Economics and International Relations], 2020, Vol. 64, No. 8, pp. 112–118. (In Russ.).

5. Panov P. V. Etnicheskie i regionalnye partii v stranakh Tsentralnoy i Vostochnoy Evropy: spetsifika i traektorii razvitiya [Ethnic and Regional Parties in Countries of Central and Eastern Europe: Specific Features and Trajectories of Development]. *Vestnik Permskogo universiteta. Politologiya* [Bulletin of the Perm University. Politology], 2020, Vol. 14, No. 3, pp. 20–34. (In Russ.).

6. Sirotkina E. V., Karandashova S. A. Loyalnost elit i vybory glav regionov: rol predvybornykh konfliktov v iskhode golosovaniya [Elite Loyalty and Electing Regional Heads: the Role of Pre-Election Conflicts in Vote Results]. *Polis. Politicheskie issledovaniya* [Polis. Political Research], 2017, No. 6, pp. 76–91. (In Russ.).

7. Solopenko A. V. Sravnenie elektoralnogo povedeniya russkoyazychnykh grazhdan Latvii i Estonii na parlamentskikh vyborakh v 2011–2015 godakh [Comparing Election Behavior of Russian-Speaking Citizens of Latvia and Estonia during Parliamentary Elections in 2011–2015]. *Baltic Research in Russia: materials of the In ternational Conference*, edited by P. Klemeshev, N. M. Mezhevich, G. M. Fedorov. Kaliningradm, Baltiyskiy federalnyy universitet imeni Immanuila Kanta, 2016, pp. 163–167. (In Russ.).

8. Tarasov I. N. Instrumenty etnicheskoy politiki v Latvii [Tools of Ethnic Policy in Latvia]. *South-Russian Journal of Social Science*, 2019, Vol. 20, No. 2, pp. 34–44. (In Russ.).

9. Tarasov I. N. Vybory v Polshe i Litve v 2020 godu [Elections in Poland and Latvia in 2020]. *Challenges and Prospects of the Kaliningrad Area Development: Geo-Politics and Geo-Economics*], edited by G. M. Fedorov. Kaliningrad, the Baltic Federal I. Kant University, 2021, pp. 68–70. (In Russ.).

10. Tarasov I. N., Fidrya E. S. Geografiya elektoralnoy volatilnosti v Varminsko-Mazurskom voevodstve Polshi [Geography of Election Volatility in Varminsko-Mazurskiy Province in Poland]. *Baltiyskiy region* [Baltic Region], 2016, No. 8 (4), pp. 78–89. (In Russ.).

11. Urazbaev E. E., Yamalova E. N. Etnopoliticheskoe dvizhenie kak provodnik institutsionalizatsii natsionalizma v sovremennoy Latvii [Ethno-Political Movement as Guide of Nationalism Institutionalization in Today's Latvia]. *Baltiyskiy region* [Baltic Region], 2020, Vol. 12, No. 2, pp. 54–69. (In Russ.).

Сведения об авторе

Антон Сергеевич Козин
аспирант Департамента политологии
факультета социальных наук
и массовых коммуникаций
Финансового университета.
Адрес: ФГОБУ ВО «Финансовый университет
при Правительстве Российской Федерации»,
125167, Москва,
Ленинградский проспект, д. 49/2.
E-mail: Antonkozinx@gmail.com

Information about the author

Anton S. Kozin
Post-Graduate Student of the Department
of Political Science of the Faculty of Social
Sciences and Mass Communications
of the Financial University.
Address: Financial University under
the Government of the Russian Federation,
49/2 Leningradskiy Avenue, Moscow,
125167, Russian Federation.
E-mail: Antonkozinx@gmail.com