УДК 339.13 JEL M31 DOI 10.26425/1816-4277-2024-4-61-72

Методы машинного обучения (токенизация) в маркетинговых исследованиях

Ганебных Елена Викторовна

Канд. экон. наук, доц. каф. менеджмента и маркетинга
ORCID: 0000-0003-0669-8318, e-mail: ganebnykh@mail.ru

Савельева Надежда Константиновна

Д-р экон. наук, дир. Института экономики и менеджмента
ORCID: 0000-0002-9497-6172, e-mail: nk_savelyeva@vyatsu.ru

Созинова Анастасия Андреевна

Д-р экон. наук, проф. каф. менеджмента и маркетинга
ORCID: 0000-0001-5876-2823, e-mail: aa_sozinova@vyatsu.ru

Фокина Ольга Васильевна

Канд. экон. наук, зав. каф. менеджмента и маркетинга
ORCID: 0000-0002-6697-3353, e-mail: fokina@vyatsu.ru

Алцыбеева Ирина Георгиевна

Канд. экон. наук, доц. каф. менеджмента и маркетинга
ORCID: 0000-0002-2204-1757, e-mail: irinaal81@yandex.ru

Вятский государственный университет, г. Киров, Россия

Аннотация

Полевые исследования представляют особый интерес в маркетинге, так как часто формируют уникальную статистику. Закрытые вопросы в ходе сбора информации упрощают обработку данных, но одновременно значительно ограничивают глубину изучения вопроса. Открытые вопросы позволяют дать более глубокое понимание мнений респондентов, однако обработка ответов в форме естественного языка (качественных данных) сложна и трудоемка, поскольку происходит обычно вручную. Современные методы машинного обучения, в частности токенизация, могут быть использованы для автоматизации обработки таких данных. Целью настоящего исследования является апробация применения указанного метода к обработке данных полевого исследования «Мониторинг состояния и развития конкуренции на товарных рынках Новосибирской области». Поставлены и решены следующие задачи: собрана и подготовлена для обработки первичная информация, выделены и сформированы группы токенов, на основании которых далее ответы респондентов были объединены в относительно однородные кластеры, включающие схожие ответы на открытые вопросы. Последующая проверка качества проведенного исследования проводилась на основании метрик точности (Precision), полноты (Recall) и F-меры, которые показали приемлемый уровень качества обработки данных. Сбор информации реализован через социологические опросы (раздачу опросных листов) и CAWI-опросы и включал открытые вопросы. В результате исследования выявлено, что даже крайне незначительные упоминания не были упущены. Полученные данные позволили сделать вывод о необходимости формирования аннотированных баз данных и библиотек токенов для целей маркетинговых исследований.

Ключевые слова

Машинное обучение, токен, токенизация, полевое исследование, открытые вопросы, качественные данные, обработка данных, естественный язык

Для цитирования: Ганебных Е.В., Савельева Н.К., Созинова А.А., Фокина О.В., Алцыбеева И.Г. Методы машинного обучения (токенизация) в маркетинговых исследованиях//Вестник университета. 2024. № 4. С. 61–72.

Machine learning methods (tokenization) in marketing research

Elena V. Ganebnykh
Cand. Sci. (Econ.), Assoc. Prof. at the Management and Marketing Department
ORCID: 0000-0003-0669-8318, e-mail: ganebnykh@mail.ru

Nadezhda K. Savelieva
Dr. Sci. (Econ.), Director of the Institute of Economics and Marketing
ORCID: 0000-0002-9497-6172, e-mail: nk_savelyeva@vyatsu.ru

Anastasia A. Sozinova
Dr. Sci. (Econ.), Prof. at the Management and Marketing Department
ORCID: 0000-0001-5876-2823, e-mail: aa_sozinova@vyatsu.ru

Olga V. Fokina
Cand. Sci. (Econ.), Head of the Management and Marketing Department
ORCID: 0000-0002-6697-3353, e-mail: fokina@vyatsu.ru

Irina G. Altsybeeva
Cand. Sci. (Econ.), Assoc. Prof. at the Management and Marketing Department
ORCID: 0000-0002-2204-1757, e-mail: irinaal81@yandex.ru

Vyatka State University, Kirov, Russia

Abstract

Field research is of particular interest in marketing because it often generates unique statistics. Closed-ended questions during data collection simplify data processing, but at the same time significantly limit the research subject depth. Open-ended questions provide a deeper understanding of respondents' opinions, but processing responses in the form of natural language (qualitative data) is difficult and time-consuming, as it is usually done manually. Modern machine learning techniques, particularly tokenization, can be used to automate such data processing. The purpose of the study is to test this method application to data processing of the field research "Monitoring of the competition state and development in the commodity markets of the Novosibirsk Region". The following tasks have been set and solved: primary information has been collected and prepared for processing, and token groups identified and formed. Based on the groups, the respondents' answers have been further combined into relatively homogeneous clusters including similar answers to open-ended questions. Subsequent quality control of the conducted research has been carried out on the basis of Precision, Recall and F-measure metrics, which showed an acceptable level of data processing quality. Information collection has been realized through sociological surveys (questionnaire distribution) and CAWI surveys and included open-ended questions. The study reveals that even extremely insignificant references were not missed. The obtained data allowed us to conclude that it is necessary to form annotated databases and token libraries for the marketing research purposes.

Keywords

Machine learning, token, tokenization, field research, open-ended question, qualitative data, data processing, natural language

For citation: Ganebnykh E.V., Savelieva N.K., Sozinova A.A., Fokina O.V., Altsybeeva I.G. (2024) Machine learning methods (tokenization) in marketing research. *Vestnik universiteta*, no. 4, pp. 61–72.

ВВЕДЕНИЕ

Цифровая революция в области искусственного интеллекта базируется на огромном количестве данных, что в свою очередь порождает новые методы их интеллектуального анализа [1]. Одним из наиболее динамичных инструментов обработки больших данных является машинное обучение, которое представляет «область искусственного интеллекта, выявляющую и изучающую закономерности, отношения и темы в данных для поддержки обучения, обработки и принятия решений без участия человека»[1] [2].

Развитие методов машинного обучения продолжается, а пристальный интерес к их возможностям обусловливает рост интенсивности соответствующих исследований. В связи с лавинообразным ростом исследований с использованием новых интеллектуальных технологий их систематизация в настоящее время затруднена, однако очевидный тренд в повсеместном применении методов машинного обучения в различных областях деятельности очевиден.

Первоначальное применение методов машинного обучения в маркетинге происходило под эгидой консьюмеризма и потребительской инженерии и было направлено на манипулирование потребителями [10; 11]. Основной задачей применения данных методов было продвижение товаров и услуг, поэтому наиболее востребованными из них стали:

1) классификация (диагностика и удержание клиентов);
2) регрессия (прогнозирование поведения клиентов и рынка в целом);
3) уменьшение размерности (обнаружение структуры и признаков целевой аудитории);
4) кластеризация (сегментация клиентов, целевой маркетинг, рекомендательные системы).

Эволюция маркетинговой социальной инженерии от «атаки на человека» с основной задачей продать продукт к инжинирингу самого продукта, а также социальной реформации через удовлетворенность потреблением, подразумевающим понимание проблем и болей клиента, вынуждает маркетологов смещать фокальный объект. По сути, непосредственный процесс исследования не меняется – используются все те же методы опросов, глубинных интервью, фокус-групп и др., однако важными становятся контекст, среда, мысли клиента.

Повсеместная цифровизация упрощает проведение опросов, позволяет собирать данные респондентов без участия интервьюеров, а также в дальнейшем довольно быстро и эффективно их обрабатывать при помощи специализированного программного обеспечения. Тем не менее проводимые социологами исследования показывают, что онлайн-опросы эффективны только с использованием крайне небольшого количества вопросов или с использованием исключительно закрытых вопросов [12].

Особый интерес в маркетинговой деятельности вызывают полевые исследования, так как часто представляют собой уникальную статистику, способную дать более глубокое понимание всех элементов взаимодействия рынка и потребителей. Очевидным является тот факт, что большие объемы собранных данных повышают точность или репрезентативность информации, но одновременно затрудняют ее обработку. Если с анализом количественных данных справляются программы статистической обработки, такие как SPSS или MiniTab, то качественные данные чаще всего обрабатываются вручную, что замедляет и удорожает аналитический процесс.

Методы обработки качественной информации применяются в машинном обучении для анализа текстов, в том числе отзывов клиентов или их постов в сети «Интернет» (далее – Интернет). Однако чаще всего данные методы сводятся к некой классификации для определения класса отзыва (положительный или отрицательный) или темы обращения (жалоба, предложение, благодарность и т.п.) [13]. Они отчасти могут помочь в статистической обработке качественных данных полевых исследований, если целью стоит классификация мнений респондентов. Важной особенностью классификации в машинном обучении является заранее определенное количество классов, то есть фактически введение ограничений на ответы в опросных листах. Это приводит к необходимости «закрывать» вопросы и ограничивать глубину понимания мнения респондентов.

При сборе глубинной информации действительно важными являются открытые вопросы, позволяющие появиться новым данным, которые не рассматривались исследователями. Таким образом, происходит приращение знаний за счет вовлечения респондента в исследование проблематики. Если речь идет о небольшом количестве глубинных интервью или фокус-группах, то обработка информации происходит в ручном режиме, а данные проходят через фильтры субъективной оценки исследователя.

[1] Machine Learning in Manufacturing. Режим доступа: https://techreviewer.co/blog/machine-learning-in-manufacturing (дата обращения: 23.01.24).

Если же речь идет о массовых опросах (часто о панельных наблюдениях), то обработка такой информации становится трудоемкой, а также возникает высокая вероятность упустить данные, которые не упоминаются часто.

МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

Наряду с традиционным применением методов машинного обучения в сфере информационных технологий (далее – ИТ) отдельные публикации свидетельствуют о попытках их применения в промышленности, химии, логистике, археологии, медицине и т.д. Примеры применяемых в различных отраслях методов машинного обучения представлены в табл. 1.

Таблица 1

Применение методов машинного обучения вне ИТ-отрасли

Отрасль	Кейс	Методы
Промышленность	Прогнозирование отказов оборудования	Логистическая регрессия Градиентный бустинг Нейронные сети
	Оптимизация ресурсов	Обучение с подкреплением Глубокое обучение
	Распознавание предметов для перемещения	Машинное зрение (семантическая сегментация)
Нефтегазовая и горнодобывающая отрасли	Оптимизация процессов	Нейронные сети
	Разведка новых месторождений	Цифровые двойники
Финансы	Оценка рисков	Кластеризация
	Борьба с мошенничеством	Случайный лес Логистическая регрессия Нейронные сети
Медицина	Диагностика заболеваний	Логистическая регрессия Нейронные сети Случайный лес Байесовский классификатор Метод опорных векторов
Ритейл	Предсказание поведения клиента	Логистическая регрессия Нейронные сети Глубокое обучение
Логистика	Оптимизация маршрутов	Логистическая регрессия Нейронные сети
	Распознавание поврежденных предметов Распознавание навигационных знаков Отслеживание уровней запасов	Машинное зрение

Составлено авторами по материалам источника[2,3,4,5] [3–9]

Проанализированные кейсы не составляют сколько-нибудь значимой доли всех существующих случаев применения методов машинного обучения и не охватывают всех отраслей его применения, однако демонстрируют горизонты применения и свидетельствуют о возможности или необходимости пересмотра существующих «традиционных» методов анализа.

[2] Широков Д.Б. Применение машинного обучения для решения задач в IT-отрасли: практические примеры и перспективы развития. Режим доступа: https://www.securitylab.ru/analytics/538199.php (дата обращения: 23.01.24).

[3] Применение машинного обучения в промышленной автоматизации. Режим доступа: https://skine.ru/articles/440432/?ysclid=lpbmtng9ij881561249 (дата обращения: 23.01.24).

[4] Зуйкова А. Как цифровые двойники помогают добывать нефть и газ. Режим доступа: https://trends.rbc.ru/trends/industry/613895d29a79477154fec314 (дата обращения: 23.01.24).

[5] Shinkarenko A. Machine learning in logistics: technology breakdown & 10 issue cases. Режим доступа: https://www.itransition.com/machine-learning/logistics?ysclid=lpci8gut6829091412 (дата обращения: 23.01.24).

В октябре 2023 г. авторами (в числе коллектива исследователей) проводился мониторинг состояния и развития конкуренции на товарных рынках Новосибирской области, в ходе которого наряду с кабинетным исследованием проводилось полевое исследование мнения потребителей (жителей региона) посредством социологических опросов (раздача опросных листов) и CAWI-опроса. Опросные листы были предоставлены заказчиком исследования – Министерством экономического развития Новосибирской области – и содержали такие открытые вопросы, как:

– на какие товары и/или услуги, по Вашему мнению, цены в Новосибирской области выше по сравнению с другими регионами;

– качество каких товаров и/или услуг, по Вашему мнению, в Новосибирской области выше по сравнению с другими регионами.

В основу формирования выборки респондентов – потребителей товаров, работ и услуг – для сбора данных об удовлетворенности качеством товаров, работ и услуг на товарных рынках Новосибирской области был положен подход многоступенчатой многомерной стратификации. Ее базовыми признаками стали половозрастные характеристики населения Новосибирской области на 1 января 2023 г. Согласно данным территориального органа Федеральной службы государственной статистики по Новосибирской области, генеральная совокупность респондентов составила 2,794 млн чел., из которых 1,249 млн (46 %) – мужчины и 1,515 млн (54 %) – женщины. Далее был проанализирован количественный состав возрастных групп (табл. 2).

Таблица 2

Половозрастной состав генеральной совокупности респондентов

Всего	Квота по полу		Квота по возрасту				
	муж.	жен.	18–24 лет	25–34 лет	35–44 лет	45–54 лет	старше 55 ktn
2 794 266	1 278 702	1 515 564	105 991	168 119	229 421	171 060	306 591
100 %	46 %	54 %	12 %	17 %	23 %	17 %	31 %

Составлено авторами по материалам источника[6]

Согласно требованиям Министерства экономического развития Новосибирской области, общий объем выборки для проведения опроса потребителей товаров, работ и услуг должен был составить 1,2 тыс. чел. Для формирования квот выборки данная совокупность была поделена пропорционально общему половозрастному составу генеральной совокупности в разрезе 30 муниципальных районов и пяти городских округов региона.

Для проведения мониторинга был осуществлен сбор данных следующими методами:

– методом CAWI-опроса с использованием сервиса для проведения онлайн-опросов «Яндекс.Формы» и размещением ссылок на анкеты на официальных сайтах областных исполнительных органов государственной власти Новосибирской области, Инвестиционном портале Новосибирской области, на официальных сайтах органов местного самоуправления в информационно-телекоммуникационной сети «Интернет»;

– методом социологического опроса, реализованного при помощи полевого исследования с раздачей опросных листов (анкет), в том числе по месту работы или жительства респондентов, в бюджетных учреждениях и организациях частных форм собственности, а также личного и почтового опросов.

В связи с активным участием населения Новосибирской области в опросе квотное требование было перевыполнено. Вместо 1,2 тыс. респондентов опрос прошли 1,829 тыс. чел., и по согласованию с заказчиком анализ результатов исследования проведен в отношении всех опрошенных.

РЕЗУЛЬТАТЫ ОСНОВНОЙ ВЫБОРКИ

Количество данных, полученных от 1,829 тыс. респондентов, с одной стороны, значительно для обработки мнений потребителей, выраженных в свободной форме, а с другой – не попадает под формат «больших данных» [14]. В целях выработки метода обработки больших объемов количественных

[6] Территориальный орган Федеральной службы государственной статистики по Новосибирской области. Распределение численности населения Новосибирской области по полу и возрастным группам на 1 января 2023 г. Режим доступа: https://54.rosstat.gov.ru/storage/mediabank/Распределение%20численности%20населения%20Новосибирской%20области%20по%20полу%20и%20возрастным%20группам%20на%201%20января%20 2023%20г..pdf (дата обращения 24.01.2024).

данных авторы провели апробацию применения метода машинного обучения (токенизации) к данной совокупности, принимая ее за условную обучающую выборку, с дальнейшей кластеризацией полученных данных. Проверка качества алгоритма, традиционно производящаяся в машинном обучении на тестовой выборке, была проведена в данном случае вручную за неимением другой совокупности данных.

Еще одной сложностью стала оцифровка данных полевого исследования, так как практически половина респондентов была опрошена через заполнение опросных листов, то есть на бумажном носителе.

Метод токенизации используется в машинном обучении для обработки естественного языка, зафиксированного в форме текста и в подготовке входных данных для дальнейшей обработки нейросетями [15]. Опуская программное описание сущности токенов, используемое в ИТ-сфере, общий смысл токенизации сводится к выявлению в тексте фонем, выделяющихся вдоль естественной границы слов, как простейших структурных элементов текста. Эти границы имеют мало общего с морфемной структурой слова и определяются на основе механики речи и когнитивных процессов. На практике данные токены представляют «обрезки» слов, чаще всего часть или весь корень слова, иногда включая части других структурных элементов, чаще всего суффиксов.

Процесс обучения алгоритма проводился в несколько итераций.

1. На первоначальном этапе в качестве «Токена 0» было принято каждое слово. В результате был составлен перечень из 38,625 тыс. таких токенов.

2. Далее из перечня были исключены знаки препинания, союзы, предлоги, частицы, междометия, местоимения, сокращения «и т.п.» и «и т.д.» как не несущие смысловой нагрузки. Также было произведено разделение слов, написанных через дефис, на два отдельных токена. В результаты количество токенов сократилось до 14,763 тыс. ед.

3. На следующем этапе весь перечень был подвергнут поиску одинаковой последовательности символов, причем в ходе эксперимента сперва был запущен поиск одинаковой последовательности из трех символов, что привело к необходимости отсечения отдельных частей слов, например, последовательности «очн», характерной как для «молочной», так и «хлебобулочной» продукции.

4. После отсечения неинформативных частей слов был запущен повторный поиск последовательности одинаковых пяти, четырех, а затем трех символов. Количество символов было определено остаточными частями слов. Таким образом, было получено 109 групп токенов.

5. Далее группы токенов были объединены в 11 смысловых кластеров (табл. 3):

- топливо;
- такси;
- транспорт (общественный транспорт, автомобили, автозапчасти, ремонт автомобилей);
- медикаменты и медицинские услуги;
- недвижимость и строительные материалы;
- жилищно-коммунальное хозяйство;
- продукты питания;
- непродовольственные товары;
- интернет и связь;
- образование;
- электроника и техника.

Таблица 3

Объединенные кластеры

Кластер	Токены
Топливо	бенз, топл, диз, нефт, соля, горю, смаз
Такси	такс
Транспорт (общественный транспорт, автомобили, автозапчасти, ремонт автомобилей)	тран, обус, езд, воз, пасс, марш, омоб, легк, озап, ремо, маши
Медикаменты и медицинские услуги	лека, меди, фарм, апте, больн, клин, леч, здра, стом, отези, анали
Недвижимость и строительные материалы	жил, квар, дом, помещ, недви, стро

Окончание табл. 3

Кластер	Токены
Жилищно-коммунальное хозяйство	комму, отоп, тепл, канал, вод, трич, энерг, жкх, жку, хозя, тбо, гвс, хвс, газ, свет, дров, уго
Продукты питания	прод, пита, еда, мяс, моло, масл, сыр, сме, рож, рыб, кур, каль, море, хле, печк, яйц, фру, апель, колб, осис, говя, сви, алко, сах, греч, соль, мук, овощ, шоко
Непродовольственные товары	одеж, обув, текст, вещ, ошок, мыл
Интернет и связь	интер, моб, связ, сото
Образование	высш, образ, дошк
Электроника и техника	трон, быт, техн

Составлено авторами по материалам исследования

При беглом просмотре ответов респондентов логичным выглядело разделение кластера «Транспорт» на «Общественный транспорт» и «Автомобили, автозапчасти и ремонт автомобилей» и кластера «Медикаменты» на «Лекарства» и «Медицинские услуги», так как это разные отрасли. Разделение данных кластеров возможно лишь при ручной обработке, поскольку процент содержания общих слов велик и зачастую разница кроется в окончаниях или комбинации слов.

Кроме того, произошло смешение слов, относящихся по смыслу к разным кластерам, но имеющим общий токен:

– дом (кластер «Недвижимость и стройматериалы») и общедомовой (кластер «Жилищно-коммунальное хозяйство»);

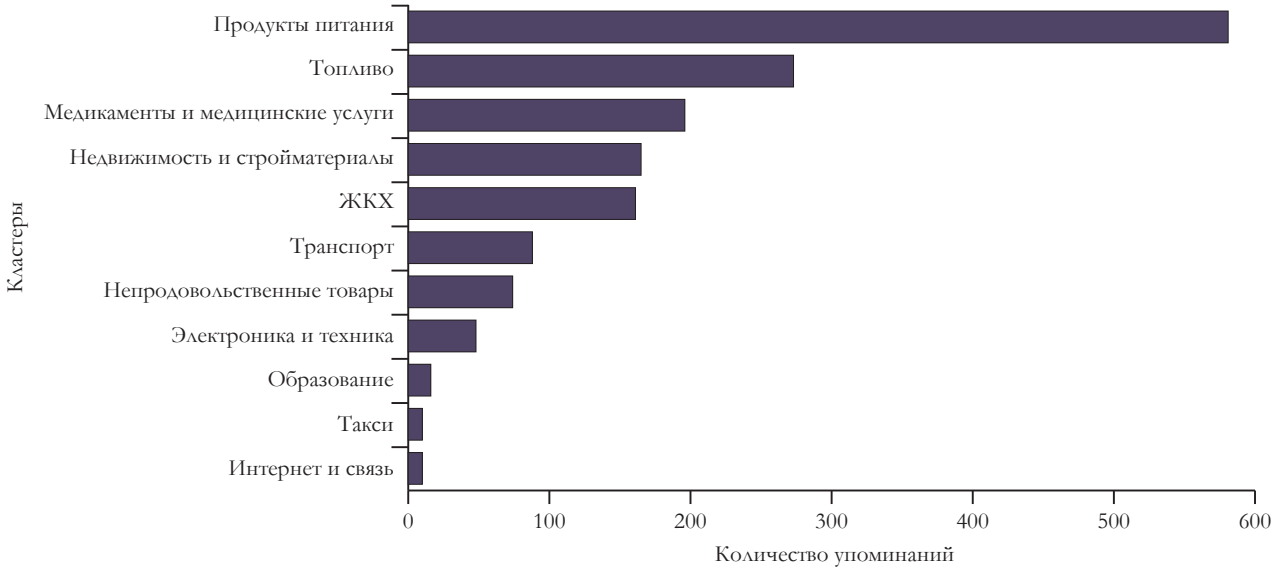
– анализ (кластер «Медикаменты и медицинские услуги») и канализация (кластер «Жилищно-коммунальное хозяйство»);

– электричество, электроэнергия (кластер «Жилищно-коммунальное хозяйство») и электроника (кластер «Электроника и техника»).

В данных словах были внесены ручные правки и произведено принудительное отнесение к соответствующим кластерам.

Особую сложность представляли слова с опечатками и сокращения слов, так как они отбрасывались в отдельную группу или попадали в неправильный кластер.

Полученные данные позволили получить объективную картину (рис. 1).



Примечание: ЖКХ – жилищно-коммунальное хозяйство

Составлено авторами по материалам исследования

Рис. 1. Распределение мнений респондентов о товарах и услугах, цены на которые в Новосибирской области выше, чем в других регионах

Как видно из рис. 1, даже крайне незначительные упоминания не были нивелированы: по 10 упоминаний (0,5 %) в кластерах «Такси» и «Интернет и связь». При ручной обработке эти упоминания вполне могли выпасть из фокуса зрения аналитика.

Качество примененного метода оценивалось по двум показателям: точность (Precision) и полнота (Recall). Первый показатель измеряет, какой процент токенов, отнесенных к определенному кластеру, действительно к нему относится, а второй измеряет, для какого процента токенов конкретного кластера действительно был определен правильный кластер. Так, Precision составил 80,6 %, а Recall – 94,6 %. Для расчета среднего гармонического значения точности и полноты был использован показатель F-меры, который составил 87 %, что является довольно высоким показателем качества, несмотря на то что проверка фактически велась вручную.

РЕЗУЛЬТАТЫ КОНТРОЛЬНОЙ ВЫБОРКИ

Далее обученный алгоритм был применен к следующему вопросу: «Качество каких товаров и/или услуг, по Вашему мнению, в Новосибирской области выше по сравнению с другими регионами?». Поскольку большинство ответов в целом совпадали с представленными в предыдущем вопросе, теоретически данную итерацию можно считать проверкой на тестовой выборке.

При запуске алгоритма были обнаружены новые токены, не использованные ранее в анализе предыдущей базы. Алгоритм необходимо было дообучить, в результате чего библиотека токенов пополнилась значениями, которые в процессе интерпретации были отнесены к кластерам «Легкая промышленность», «Досуг» и «Туризм».

Так, было выделено 10 кластеров:

- продукты питания;
- жилищно-коммунальное хозяйство;
- образование;
- медицина;
- недвижимость и строительные материалы;
- транспорт;
- дороги;
- легкая промышленность;
- досуг;
- туризм.

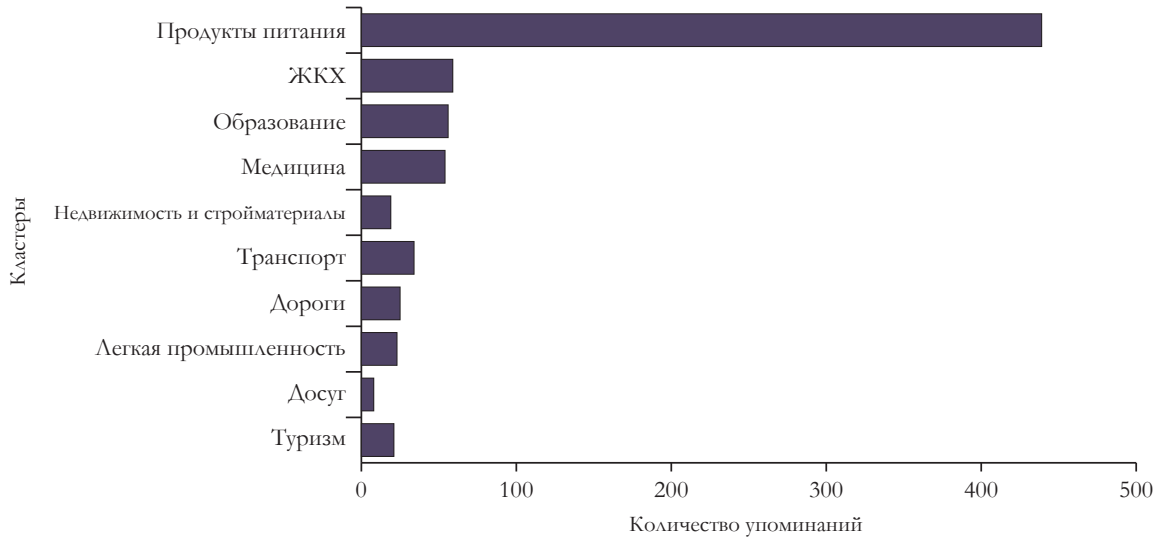
Некоторые затруднения были обнаружены при разделении кластеров «Досуг» и «Туризм». В итоге к первому кластеру была отнесена интерпретация «культурного досуга», включающего театры, кино, выставки, музеи и сам токен «досуг», а все, что связано с путешествиями и походами, было отнесено к кластеру «Туризм».

Количественные показатели зафиксированного мнения респондентов в отношении товаров и услуг, качество которых в Новосибирской области выше по сравнению с другими регионами, представлены на рис. 2.

По-прежнему проблему составили слова, написанные с ошибками, и сокращения слов, однако показатели тестовой выборки были выше тех, что были получены на обучающей: Precision = 82,3 %, Recall = 95,1 %, F = 88,2 %.

Проведенный эксперимент показал хороший уровень пригодности использования алгоритма для анализа неструктурированных данных, касающихся мнения потребителей товаров и услуг о качестве конкуренции на товарных рынках.

В результате еще одной проверки алгоритм был применен к уже обработанной вручную базе данных опроса населения в отношении доступности финансовых услуг и удовлетворенности деятельностью в указанной сфере, осуществляемой на территории Новосибирской области. Опрос проводился также двумя методами аналогично опросу потребителей товаров, работ и услуг – при помощи раздачи опросных листов и при помощи CAWI-опроса через «Яндекс.Формы». Квотное задание в 300 респондентов также было превышено, и общая совокупность опрошенных составила 719 чел., из которых 32,4 % – полевое исследование и 67,6 % – CAWI-опрос.



Примечание: ЖКХ – жилищно-коммунальное хозяйство

Составлено авторами по материалам исследования

Рис. 2. Распределение мнений респондентов о товарах и услугах, качество которых в Новосибирской области выше, чем в других регионах

Единственным открытым вопросом данного исследования был вопрос: «Укажите, какие трудности у Вас возникают при получении или использовании финансовых продуктов (услуг)». Чуть более 1/5 респондентов (20,6 %) не ответили. Еще 59,4 % ответили «Нет» в различных вариациях, что может трактоваться как «Нет трудностей» или как нежелание отвечать на вопрос. Таким образом, качественные данные были предоставлены лишь пятой частью (20,0 %) опрошенных респондентов.

В связи с тем, что тематика опроса была иной, чем в обучающей выборке, уже собранная библиотека токенов в данном случае не подошла. Для подбора токенов был использован алгоритм, аналогичный использованному при анализе ответов потребителей товаров, работ и услуг. В результате было сформировано 7 кластеров:

– расстояние до места оказания финансовых услуг (далеко, дорога, идти, ехать, место);

– ожидание предоставления финансовой услуги (очередь, ждать, время, работать, выходной, режим, зависает);

– недостаточное количество пунктов обслуживания, включая банкоматы (офис, отделение, банк, банкомат, терминал, наличные, доступ, магазин, мало, вечер, ночь);

– навязчивость услуг (навязчиво, ненужный, страховка, рассылка);

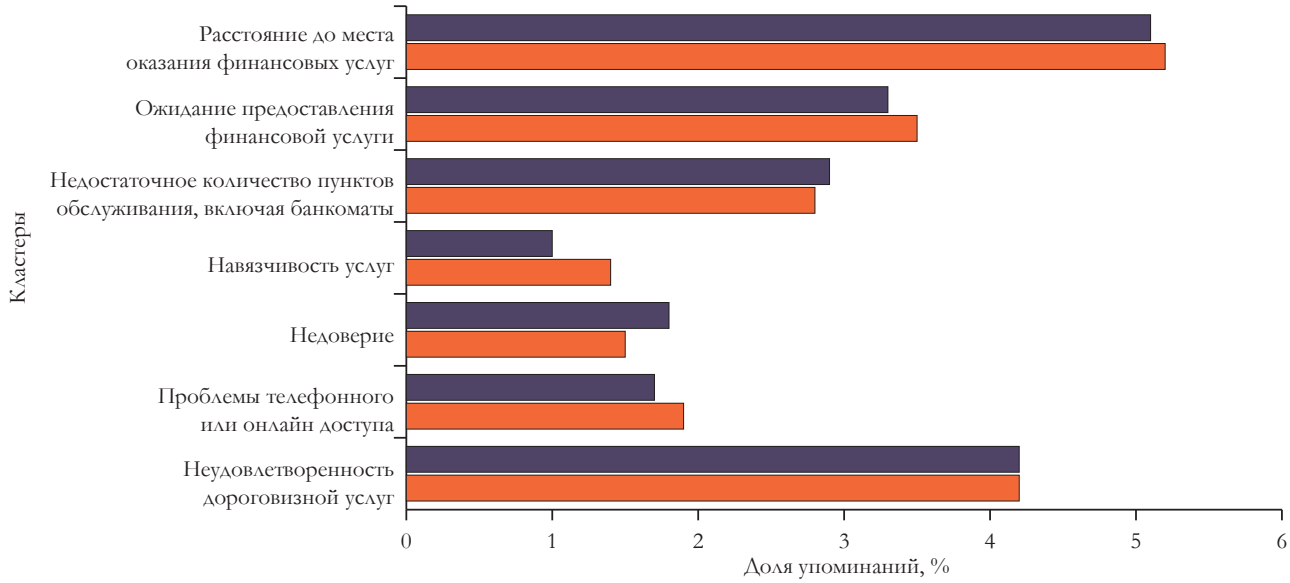
– недоверие (недоверие, скрытый, мошенник, условие, доверие, понятно, обман, правда);

– проблемы телефонного или онлайн доступа (интернет, онлайн, автоответчик, бот, телефон, обращение, дозвон, обслуживать, сбой, приложение, связь);

– неудовлетворенность дороговизной услуг (%, процент, ставка, взнос, ипотека, кредит, высокий, огромный, дорогой, комиссия).

Кроме вышеперечисленных упоминаний, были обнаружены универсальные ответы, которые проблематично однозначно отнести к тому или иному кластеру: «проблема», «сложность», «трудность», «оплата», «информация». Тем не менее полученные данные позволили получить общую картину трудностей, возникающих у потребителей при получении или использовании финансовых продуктов (услуг), и сравнить ее с обработанными вручную данными (рис. 3).

Сравнительный анализ данных, полученных при помощи токенизации и ручной обработки, показал незначительные отличия в пределах 0,1–0,4 %. Наибольшее отклонение выявлено в кластерах «Навязчивость услуг» и «Недоверие». Для понимания причин такого отклонения собранные данные проанализированы повторно. Обнаружено, что часть токенов («понятно», «скрытый», «условие») были отнесены при машинной обработке к кластеру «Недоверие», в то время как аналитик при описании данных отнес указанные ответы в кластер «Навязчивость» в связи с контекстом.



Составлено авторами по материалам исследования

Рис. 3. Распределение мнений респондентов о трудностях, возникающих при получении или использовании финансовых продуктов (услуг)

В целом результаты признаны приемлемыми, однако на незначительном объеме данных сам процесс токенизации занимает больше времени, чем ручная обработка, а также даже небольшой процент погрешности, связанный с отсутствием анализа контекста ответа, может давать значительные отклонения, когда общее число ответов невелико.

ЗАКЛЮЧЕНИЕ

Проведенный эксперимент показал хороший уровень пригодности использования алгоритма для анализа неструктурированных данных, касающихся мнения потребителей товаров и услуг о качестве конкуренции на товарных рынках. При этом очевидной стала возможность сбора данных из других источников, например, социальных сетей. Таким образом, осуществление мониторинга становится возможным на постоянной основе в режиме реального времени, что значительно сокращает разрыв между выявлением отклонения и принятием соответствующих мер. Это может быть особенно важным для социальной инженерии в области государственной политики и реформирования, устанавливающих правила, в том числе для массового потребления.

Аналогичным образом целесообразно реструктурировать всю систему мониторинга, так как на практике респонденты затрудняются ответить на большую часть закрытых вопросов, не обнаруживая подходящего ответа или попросту отмечая «затрудняюсь ответить». Открытые вопросы дают более глубокое понимание проблем и болей потребителей, вопрос стоит лишь в обработке полученных данных. Одновременно с этим остро встает вопрос о том, как обойти нежелание респондентов писать ответы на открытые вопросы, так как по всем исследованиям их доля не превышает 20 %, а сами по себе открытые вопросы могут стать причиной прекращения заполнения анкеты в целом.

Наличие возможности голосовых ответов на такие вопросы, может стать решением проблемы. Кроме того, анализ голосовых интонаций может помочь в проведении более глубокого исследования. При этом автоматическая расшифровка голосовых сообщений (перевод в текст) становится необходимым встроенным инструментом форм для проведения опросов.

Методы машинного обучения требуют качественных данных, то есть предварительно обработанных под существующие методы. Качество данных в этом случае является многомерной конструкцией, охватывающей не только актуальность, своевременность и доступность, но и интерпретируемость и пригодность для обработки. Для повышения точности и полноты алгоритма, а значит и ускорения обработки данных возможно создание некой аннотированной базы данных, созданной под конкретные

нужды и выступающей в качестве стандарта (например, с уже отобранными и проверенными токенами). В этом случае качество стандарта будет по большей мере отражать качество обработанных данных. При этом создание стандарта аннотированной базы данных может происходить как машинными методами, так и вручную.

В целом необходимо отметить, что обработка незначительных объемов данных не требует автоматизации. Более того, ручная обработка в таком случае может стать более качественной, так как исследователь будет иметь возможность рассматривать данные в контексте, чего лишена токенизация: предварительная обработка текста исключает неинформативные элементы, например, частицы, а частица «не» может диаметрально менять значение информации.

Большие исследования, такие как перепись населения, требуют предельно структурированных опросных листов без открытых вопросов, что значительно сужает горизонт возможностей сбора информации при использовании колоссальных ресурсов на организацию самого процесса сбора. Обученные и протестированные на небольшой выборке аннотированные базы дают огромные возможности для обработки неструктурированной информации, что при тех же затратах на проведение опроса увеличивает информативность в разы.

Список литературы

1. *Андреев А.В.* Искусственный интеллект и его роль в обработке больших данных. Умная цифровая экономика. 2023;1(3):65–69.
2. *Sarker I.H.* Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science. 2021;2:160.
3. *Шлыков С.В.* Применение методов машинного обучения для автоматизации процессов в нефтегазовой отрасли. Транспорт и хранение нефтепродуктов. 2023;2:46–53. https://doi.org/10.24412/0131-4270-2023-2-46-53
4. *Бобков С.П., Суворов С.В., Орлов А.П., Пивнев Е.А.* Использование методов машинного обучения для оценки рисков при внедрении нового кредитного продукта. Известия ВУЗов. Серия «Экономика, финансы и управление производством». 2020;4(46):59–63. https://doi.org/10.6060/ivecofin.2020464.509
5. *Осипова Т.А., Зайцев К.С., Бифферт В.О.* Применение алгоритмов машинного обучения к задаче выявления мошенничества при использовании пластиковых карт. International Journal of Open Information Technologies. 2021;8:23–29.
6. *Ходашинский И.А., Сарин К.С., Бардамова М.Б., Светлаков М.О., Слезкин А.О., Корышев Н.П.* Биометрические данные и методы машинного обучения в диагностике и мониторинге нейродегенеративных заболеваний: обзор. Computer Optics. 2022;6(46):988–1020. https://doi.org/10.18287/2412-6179-CO-1134
7. *Tautan A., Ionescu B., Santarnecchi E.* Artificial intelligence in neurodegenerative diseases: A review of available tools with a focus on machine learning techniques. Artificial Intelligence in Medicine. 2021;117:102081. https://doi.org/10.1016/j.artmed.2021.102081
8. *Chaudhary K., Alam M., Al-Rakhami M.S. et al.* Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics. Journal of Big Data. 2021;1(8):73. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00466-2
9. *Amrit C., van Hillenberg C., van der Spoel S.* Predictive Analytics for Truck Arrival Time Estimation: A Field Study at a European Distribution Center. International Journal of Production Research. 2016;55(17):1–21. http://dx.doi.org/10.1080/00207543.2015.1064183
10. *Logemann J., Gross G., Köhler I.* Consumer Engineering, 1920s–1970s: Marketing between Expert Planning and Consumer Responsiveness. London: Palgrave Macmillan; 2019. 296 p. http://dx.doi.org/10.1007/978-3-030-14564-4
11. *Положихина М.А.* Маркетинг и общество потребления: Рецензия на коллективную монографию «Consumer Engineering, 1920s–1970s: Marketing between Expert Planning and Consumer Responsiveness» / ed. By J. Logemann, G. Cross, I. Köhler. – London: Palgrave Macmillan, 2019. – 296 p. Социальные новации и социальные науки. 2021;3:133–155. https://doi.org/10.31249/snsn/2021.03.08
12. *Шкурин Д.В.* Сравнительная оценка качества данных офлайн- и онлайн-опросов. Дискуссия. 2015;8(60):101–104.
13. *Subramaniyaswamy V., Harshaa S., Padma J.M., Prabhalammbeka B.S.* Sentiment analysis string token classification algorithm. International Journal of Pure and Applied mathematics. 2018;12(119):13287–13294.
14. *Денисова О.Ю., Мухутдинова Э.А.* Большие данные – это не только размер данных. Вестник Казанского технологического университета. 2015;4(18):226–230.
15. *Friedman R.* Tokenization in the Theory of Knowledge. Encyclopedia. 2023;1(3):380–386. http://dx.doi.org/10.3390/encyclopedia3010024

References

1. *Andreev A.V.* Artificial intelligence and its role in big data processing. Smart Digital Economy. 2023;1(3):65–69. (In Russian).
2. *Sarker I.H.* Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science. 2021;2:160.
3. *Shlykov S.V.* Application of machine learning methods to automate processes in the oil and gas industry. Transport and storage of oil products and hydrocarbons. 2023;2:46–53. (In Russian). https://doi.org/10.24412/0131-4270-2023-2-46-53
4. *Bobkov S.P., Suvorov S.V., Orlov A.I., Pivnev E.A.* Using machine learning methods to assess risks when implementing a new credit product. News of Higher Educational Institutions. The Series "Economics, Finance and Production Management". 2020;4(46):59–63. (In Russian). https://doi.org/10.6060/ivecofin.2020464.509
5. *Osipova T.A., Zaytsev K.S., Bifert V.O.* The use of ML algorithms in the task of detecting fraud when using plastic cards. International Journal of Open Information Technologies. 2021;8:23–29. (In Russian).
6. *Khodashinskyi I.A., Sarin K.S., Bardamova M.B., Svetlakov M.O., Slezkin A.O., Koryshev N.P.* Biometric data and machine learning methods in the diagnosis and monitoring of neurodegenerative diseases: a review. Computer Optics. 2022;6(46):988–1020. (In Russian). https://doi.org/10.18287/2412-6179-CO-1134
7. *Tautan A., Ionescu B., Santarnecchi E.* Artificial intelligence in neurodegenerative diseases: A review of available tools with a focus on machine learning techniques. Artificial Intelligence in Medicine. 2021;117:102081. https://doi.org/10.1016/j.artmed.2021.102081
8. *Chaudhary K., Alam M., Al-Rakhami M.S. et al.* Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics. Journal of Big Data. 2021;1(8):73. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00466-2
9. *Amrit C., van Hillenberg C., van der Spoel S.* Predictive Analytics for Truck Arrival Time Estimation: A Field Study at a European Distribution Center. International Journal of Production Research. 2016;55(17):1–21. http://dx.doi.org/10.1080/00207543.2015.1064183
10. *Logemann J., Gross G., Köhler I.* Consumer Engineering, 1920s–1970s: Marketing between Expert Planning and Consumer Responsiveness. London: Palgrave Macmillan; 2019. 296 p. http://dx.doi.org/10.1007/978-3-030-14564-4
11. *Polozhikhina M.A.* Marketing and consumer society: a review of a collective monograph "Consumer Engineering, 1920s–1970s: Marketing between Expert Planning and Consumer Responsiveness". Social Novelties and Social Sciences. 2021;3:133–155. (In Russian). https://doi.org/10.31249/snsn/2021.03.08
12. *Shkurin D.V.* Comparative evaluation of data quality of online and offline surveys. Discussion. 2015;8(60):101–104. (In Russian).
13. *Subramaniyaswamy V., Harshaa S., Padma J.M., Prabhalammbeka B.S.* Sentiment analysis string token classification algorithm. International Journal of Pure and Applied mathematics. 2018;12(119):13287–13294.
14. *Denisova O.Yu., Muhutdinova E.A.* Big data is not just about the size of the data. Herald of Technological University. 2015;4(18):226–230. (In Russian).
15. *Friedman R.* Tokenization in the Theory of Knowledge. Encyclopedia. 2023;1(3):380–386. http://dx.doi.org/10.3390/encyclopedia3010024