

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение высшего
образования
«Тольяттинский государственный университет»

Гуманитарно-педагогический институт

(наименование института полностью)

Кафедра

Журналистика и социология

(наименование)

42.04.02 Журналистика

(код и наименование направления подготовки)

Журналистика данных

(направленность (профиль))

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
(МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ)**

на тему

Big data и журналистика данных: актуальный опыт взаимодействия

Обучающийся

М.В. Гончаренко

(Инициалы Фамилия)

(личная подпись)

Научный

руководитель

кандидат филологических наук, доцент, Л. В. Иванова

(ученая степень (при наличии), ученое звание (при наличии), Инициалы Фамилия)

Тольятти 2024

Оглавление

Введение.....	3
Глава 1 Большие данные как источник информации в журналистской деятельности.....	7
1.1 Большие данные как явление цифрового общества	7
1.2 Применение больших данных в журналистской деятельности	24
Глава 2 Использование технологий работы с большими данными в редакции сетевого издания.....	45
2.1 Использование инструмента веб-аналитики для работы с большими данными информационного ресурса	45
2.2 Разработка демонстрационной версии программы-парсера новостей	66
Заключение.....	83
Список используемой литературы и используемых источников.....	85

Введение

Развитие цифровых технологий и гаджетов в 20-х годах XXI века привело к накоплению огромных массивов данных. На текущий момент их общий объем, по некоторым оценкам [46], приближается к сотне зеттабайт. ООН прогнозирует их ежегодное увеличение на 40% [23].

В сети данные хранятся либо в виде структурированных баз данных, из которых даже обычный пользователь может извлечь необходимую ему информацию, либо в виде полуструктурированных или неструктурированных данных («больших данных»), для работы с которыми требуются специально подготовленные профессионалы.

Вопрос об использовании данных в журналистской деятельности решен, в XXI веке сформировалось направление профессиональной деятельности – журналистика данных. Однозначно признано, что в журналистике данных основным источником информации являются открытые базы данных, из них дата-журналисты черпают проблемы и истории для материалов, из которых извлекают смыслы, полезные для массовой аудитории. А вот вопрос использования больших данных в работе журналистов и СМИ остается дискуссионным. Большинство исследователей журналистики данных настаивает на том, что специалисты не работают с большими данными, так как они являются основным инструментом маркетинга и экономической аналитики. К тому же, использование больших данных подразумевает владение специфическими инструментами, и у многих журналистов отсутствуют компетенции, необходимые для качественной работы с массивами неструктурированных сведений.

Таким образом, актуальность настоящего исследования обусловлена двумя ключевыми факторами. С одной стороны, это экспоненциальный рост объемов данных, их регулярное пополнение новыми массивами, в которых

скрыто множество смыслов, полезных для социума, с другой – неоднозначное отношение журналистов и сотрудников редакций СМИ к такому информационному ресурсу как большие данные, и необходимость осмыслить аспекты взаимодействия редакций СМИ и big data.

Объектом исследования в магистерской диссертации являются большие данные как источник информации для журналистов и редакции СМИ.

Предметом исследования – способы использования больших данных в работе журналистов и редакции СМИ.

Цель магистерской диссертации: на основе систематизации теоретических представлений и изучения актуального состояния информационной сферы описать способы применения больших данных для решения профессиональных задач журналистов и редакций СМИ и предложить подходы к оптимизации использования big data при производстве контента.

Гипотеза исследования состоит в том, что грамотная «настройка» инструментов работы с большими данными способствует оптимизации деятельности редакции СМИ в части решения маркетинговых задач и производства новостного контента.

Для достижения поставленной цели необходимо решить следующие задачи:

- Систематизировать научные представления о феномене больших данных и возможностях их использования в работе редакции СМИ.
- Описать потенциально возможные способы использования больших данных в работе редакций СМИ и в журналистике.
- По заказу редакции сетевого издания оптимизировать работу с большими данными: произвести модификацию отчетов инструмента веб-аналитики; разработать демонстрационную версию программы-парсера новостей.

Теоретико-методологическую базу исследования составили: труды отечественных исследователей о применении и использовании больших данных в журналистике и работы зарубежных исследователей, посвященных журналистике данных.

Методы исследования:

- содержательный анализ – метод, ориентированный на выявление семантического потенциала данных в журналистских материалах;
- исторический метод – для изучения актуального мирового и отечественного опыта использования данных и больших данных в конкретных типах и видах массмедиа;
- метод моделирования – метод, позволяющий смоделировать технологический процесс использования больших данных для решения маркетинговых и профессионально-творческих задач редакций СМИ с учетом типологических характеристик издания и специфики проектного управления процессом.

В качестве эмпирической базы использовались материалы, написанные на основании больших данных и опубликованные в зарубежных и российских СМИ, таких как The New York Times, The Guardian, The Washington Post, Reuters, а также РИА Новости, РБК, ТАСС, Т-Ж. В эмпирическую базу также вошло большое количество чат-ботов информационных каналов социальных сетей «ВКонтакте» и Telegram.

Научная новизна исследования заключается в систематизации научных представлений и выработке рабочего определения понятия больших данных, применимое к деятельности журналистики и СМИ. Теоретическая значимость заключается в концептуальном обосновании технологий использования больших данных в журналистике при производстве новостного контента и в работе редакции СМИ для решения маркетинговых задач.

Практическая значимость исследования заключается в разработке способов работы редакции конкретного сетевого издания в направлении решения маркетинговых задач и подготовке новостного контента.

Положения, выносимые на защиту:

- Работа с большими данными в области производства контента для СМИ требует от журналистов наличия цифрового капитала и развитых цифровых компетенций.
- Применение больших данных для решения маркетинговых задач способствует оптимизации информационной политики конкретного СМИ.
- Использование программы-парсера новостей позволяет сократить количество времени на производство, повысить позиции публикуемой новости в рейтингах на сайтах-агрегаторах.

Структура магистерской диссертации состоит из: введения, двух глав, заключения, содержит 28 рисунков, список использованной литературы (66 источников).

Основной текст работы изложен на 95 страницах.

В первой главе «Большие данные как источник информации журналистской деятельности» комплексно изучается феномен больших данных, а также рассматриваются различные точки зрения на использование журналистами и сотрудниками редакции СМИ больших данных в своей профессиональной деятельности.

Во второй главе «Использование технологий работы с большими данными в редакции сетевого издания городского типа» изучаются возможности инструмента веб-аналитики «Яндекс Метрика» и производится модификация его отчетов. На основе результатов отчетов разрабатывается демонстрационная версия программы-парсера новостей с наиболее популярного источника трафика.

Глава 1 Большие данные как источник информации в журналистской деятельности

1.1 Большие данные как явление цифрового общества

Начало XXI века ознаменовалось стремительным и интенсивным внедрением во все сферы жизни общества информационных технологий. Это выразилось в бурном развитии вычислительной техники и средств связи, появлении новых и совершенствовании существующих информационных технологий, создании прикладных информационных систем. Достижения в области информационных технологий нашли применение в сфере организационного управления, промышленности, научных исследований и автоматизированного проектирования. Информатизация также затронула социальную сферу: образование, культуру, здравоохранение, науку и т.д.

В сфере массовых коммуникаций в это время набирают популярность социальные сети, мобильные платформы распространения контента – наступает эпоха интернета, которую Т. О'Рейли назвал Web 2.0. Развитие сферы оказания информационных услуг способствует формированию рынка, на котором информация рассматривается как коммерческий ресурс, обладающий определенной ценностью.

Пандемия COVID-19 в 2020 году, введение карантинных мер привели к очередному витку развития информационных технологий. Заметный рывок был сделан в сферах искусственного интеллекта и больших данных. Разработанные в них технологии стали использоваться в самых разных отраслях и сферах деятельности.

Увеличение объема данных, подлежащих обработке, и необходимость быстрой адаптации к новым методам работы и технологиям стало вызовом для цифрового общества, ответом на который стало внедрение новых технологий

обработки информации и рост уровня автоматизации различных процессов. «Процессы, методы и приемы, лежащие в основе использования ИКТ, продолжают экспоненциально развиваться и диверсифицироваться год за годом, и, по оценкам, темпы экспоненциального роста будут сохраняться, по крайней мере, в течение следующих 20 лет, вызывая новые специфические характеристики социальной среды, в которой мы будем жить» [12, с. 99], – отмечают Н. Гаврилюк и В. Сорочан. Одним из таких изменений, вызванных ускорением процессов работы с информацией, стало ускорение темпа человеческой жизни, что повлекло за собой негативные последствия: усиление нагрузки на психику человека, возникновение проблем со здоровьем, уменьшения времени, проводимого людьми в непосредственном контакте друг с другом. Н. Мамедова в работе «Человек в эпоху цифровизации: на грани реального и виртуального» сообщает: «Совершенно очевидно, что нынешнее поколение молодых людей растет в цифровой экосистеме, беспрецедентной по своей сложности и глобальности. Гораздо менее очевидны последствия для психологического здоровья, риски и преимущества этого нового цифрового мира» [27, с. 80].

Несмотря на неоднозначные эффекты развития информационных технологий современное общество уже невозможно представить без них. Они пронизывают все сферы жизнедеятельности общества: работу, коммуникацию и общение, досуг, развлечение и другие повседневные задачи. Интернет стал основной платформой коммуникации и обмена информацией: «в виртуальном пространстве нет препятствий и каких-либо ограничений для осуществления общения» [18, с. 100], – отмечают М. Заргарян, Н. Гришин и К. Садченко.

Благодаря развитию технологий информация и данные стали играть ключевую роль в формировании картины мира, они оказывают влияние на все аспекты жизнедеятельности человека и общества, начиная от науки и технологий и заканчивая социальными отношениями. Сферы их

ответственности в этом процессе распределились следующим образом: информация позволяет изучать и понимать окружающий мир, создавать новые технологии и решения; данные как оцифрованные сведения о мире применяются при анализе и интерпретации явлений и процессов.

Цитата «Данные – это новая нефть», приписываемая британскому математику и эксперту в области анализа данных К. Хамби, отражает роль данных в современном мире. Как и нефть, данные представляют собой ценный ресурс, однако в необработанном виде они непригодны для использования. Подобно тому, как нефть подвергается переработке и анализу для производства товаров, приносящих прибыль, данные также нуждаются в обработке и анализе для того, чтобы они могли принести пользу.

В последние годы объем данных, генерируемых человечеством, значительно увеличился. Помимо генерации и создания данных, человечество также достигло значительных успехов в области сбора и обработки данных. На сегодняшний день существуют и активно используются возможности по сбору широкого спектра данных из большого множества различных источников, как в структурированном, так и неструктурированном видах. Это стало возможно благодаря развитию информационных технологий и программного обеспечения. Использование специализированных инструментов и алгоритмов для обработки больших объемов данных, позволило собирать данные быстро и качественно, практически в режиме реального времени.

Совершенствуются способы хранения данных, собранных в течение многих лет. Для этого используются различные форматы и носители. Форматы хранения данных включают в себя текстовые файлы, базы данных, электронные таблицы, аудио, видео и многие другие типы файлов. Что касается носителей данных, то они могут быть физическими, такими как жесткие диски и флэш-накопители, или виртуальными, такими как облачные хранилища. Физические носители бывают разных размеров и типов, а

виртуальные носители обычно предоставляются интернет-сервисами. Оба типа носителей активно используются для хранения данных.

Десять лет назад использование данных осознавалось компаниями и организациями как полезное, но их наличие не было обязательным условием для успешной работы. В 20-х годах XXI века ситуация кардинально изменилась: успех компаний и организаций стал зависеть от качества работы с данными и использования их для принятия решений. Данные стали необходимы для успешного функционирования и получения конкурентных преимуществ на рынке. Появился подход data-driven, который фиксирует, что данные являются основой для принятия решений. О. Ударцева поясняет, что «обоснованность применения data-driven-подхода заключается прежде всего в том, что управленческие решения не принимаются интуитивно, а подкрепляются конкретными данными, что позволяет глубоко изучить текущее состояние субъекта» [50, с. 67].

Стремительное развитие технологий и непрерывное увеличение объемов генерируемой информации привели к возникновению такого феномена, как «большие данные». Определение этого термина вызывает ряд трудностей не только у обычных пользователей, но и у научного сообщества. Так, М. Корнев рекомендует: «... не замыкаться на ограниченном круге определений, а использовать различные характеристики понятия, обращаться к различным исследованиям и подходам в определении “больших данных”» [25, с. 84].

Во многих работах упоминается, что термин big data (большие данные) впервые был использован редактором журнала Nature К. Линчем в 2008 году, когда он отметил, что этот феномен вызывает значительный интерес для научных и исследовательских сообществ, поскольку он может помочь им лучше понять различные явления и процессы.

Однако использование этого термина датируется гораздо более ранними референциями. Кто-то связывает появление больших данных с началом 2000-х годов, когда широко стали доступны технологии и оборудования для работы с большими объемами данных. Кто-то относит появление больших данных к более поздним периодам истории.

Несмотря на различные точки зрения на происхождение феномена больших данных, рассматривать их только как результат технологической эволюции не вполне корректно, поскольку эта концепция возникла на стыке технологического развития и потребности в таком анализе. Фактически, о больших данных говорят как о социально-экономическом феномене, который возник в тот момент, когда появилась возможность обработки огромных объемов данных и возникла реальная потребность и возможность использования этих данных для различных целей.

Таким образом, большие данные – это социально-экономический феномен, который связан с обработкой и анализом огромных объемов данных, непрерывно генерируемых в современном мире.

Для лучшего понимания этого феномена используется их противопоставление с «малыми данными» (small data). Безусловно, первые сегодня по праву занимают центральное место в исследовательских работах и общественном внимании. Однако малые данные все еще остаются актуальными. Те же технические и социальные факторы, которые привели к появлению больших данных, также привели к образованию множества небольших наборов данных.

Малые данные относятся к контролируемым по размеру и объему наборам данных, зачастую измеряемым в мегабайтах, гигабайтах, а также терабайтах. Как правило, они состоят из хорошо структурированных однотипных данных. Такие данные ценятся за высокое качество и

непосредственную интерпретируемость, благодаря чему они хорошо подходят для определенных типов анализа.

Существует мнение, что чем больше данных, тем лучше. В работе «Технология big data в бизнесе – преимущества и пути совершенствования» исследователи О. Зиниша, Д. Кочаян и М. Мокосеева отмечают, что «перспективы внедрения Больших Данных связаны с неоспоримым конкурентным преимуществом, которое получают компании в части повышения операционной эффективности бизнеса, привлечения дополнительного потока клиентов, минимизации рисков и внедрения технологий анализа и прогнозирования данных» [19, с. 157]. При прочих равных это утверждение можно считать справедливым, но на практике увеличение объема данных влечет за собой различные дополнительные издержки. Особенно это может быть заметно в условиях фиксированного бюджета, требующего компромисса между качеством и количеством данных. Так, Л. Федорова поясняет, что «применение технологий больших данных относится к инновационным проектам, а как известно, они сложны с точки зрения оценки эффективности вложений и гарантированного результата, поэтому далеко не все компании стремятся внедрять их в свои операционные процессы» [52, с. 326]. Соответственно, в некоторых случаях малые данные могут быть более предпочтительными, чем большие, поскольку они позволяют делать правильные выводы быстрее, надежнее и с меньшими затратами.

Выявление и оценка различий между большими данными и малыми данными указывают на фундаментальную разницу в анализе. Оба этих типа данных обладают значительным потенциалом, однако они имеют свои особенности и требуют различных подходов при использовании.

Большие данные представляют собой масштабные и комплексные наборы данных, обработка которых выходит за рамки возможностей традиционных методов. Разные исследователи характеризуют их по разному

набору свойств. В тексте государственного стандарта «Информационные технологии. Большие данные. Обзор и словарь» [14] говорится о том, что большие данные обладают такими характеристиками, как: объем, разнообразие, скорость обработки и/или вариативность. Одной из распространенных концепций определения свойств больших данных является их схожая характеристика согласно 4V: объем (volume), скорость (velocity), многообразие (variety) и достоверность (veracity).

Через объем они определяются как огромные наборы данных, которые измеряются в экзабайтах и даже зеттабайтах. Подобный масштаб выходит далеко за границы того, что могут эффективно обработать традиционные системы управления данными. Стоит также отметить, что раньше большие данные измерялись в петабайтах и терабайтах [33]. Однако стремительное развитие информационных технологий и комплектующих к ним сделало последние обыденной единицей измерения информации для современного времени. Сегодня большое количество современных устройств имеют поддержку расширения памяти до таких объемов.

Значительные объемы вызывают серьезные проблемы, связанные с хранением данных, вычислительными мощностями и пропускной способностью сети. Эффективное управление и анализ данных требуют специализированных систем хранения, а также платформ для распределенных вычислений.

Такое свойство больших данных как скорость, трактуется и воспринимается исследователями по-разному. Одни говорят, что скорость характеризует большие данные с точки зрения скорости генерации этих данных. В этом плане для больших данных свойственен непрерывный поток производства сырых данных. Другие отмечают, что под скоростью определяется время, за которое большие данные обрабатываются [28]. Эта

позиция подразумевает, что обработка данных происходит в режиме реального времени, то есть практически в момент генерации.

Таким образом, в совокупности скорость в контексте данных обозначает экстремальную быстроту, с которой они генерируются, аккумулируются и обрабатываются. Быстрая генерация данных служит определяющим признаком, отличающим большие данные от малых.

Большая скорость передачи данных ведет к возникновению ряда проблем. Она требует способности обрабатывать данные в режиме реального времени или практически в режиме реального времени, чтобы извлечь полезную информацию по ходу развития событий. Анализ такого потока данных предполагает использование передовых технологий, таких как платформы для обработки потоков и сложных систем обработки событий. Принимая быстрые решения на основании данных и быстро реагируя на изменяющиеся условия, скорость позволяет добиться конкурентного преимущества.

Свойство многообразия больших данных заключается в том, в каком виде они собираются. Существуют три таких основных вида: структурированные, полуструктурированные, у которых есть схема, но она не постоянна, и неструктурированные, те у которых полностью отсутствует какая-либо структура. Как и в случае с традиционными базами данных, структурированные типы данных имеют четко определенную форму, что облегчает их организацию и анализ. Частично структурированные данные (или полуструктурированные), подобные файлам XML или JSON, обладают некоторой структурой, но при этом сохраняют гибкость. Неструктурированные типы данных требуют заранее определенного формата и могут включать текст, изображения, видео и содержимое социальных сетей. В работе «Информационные технологии обработки больших данных» П. Четырбок описывает структурированные и неструктурированные данные с

точки зрения их характера: «Структурированные данные часто называют количественными данными, что означает, что их объективный и заранее определенный характер позволяет нам легко подсчитывать, измерять и выражать данные в числах. Неструктурированные данные также называются качественными данными в том смысле, что они имеют субъективный и интерпретирующий характер» [56, с. 277].

Технологии обработки больших данных и аналитические платформы адаптированы для работы с таким широким спектром источников разного формата. Инновационные подходы, такие как анализ естественного языка и обработка изображений, позволяют извлекать полезную информацию из неструктурированных типов данных.

Достоверность больших данных относится к точности и надежности информации в наборе. Из-за большого объема и разнообразия источников часто возникают проблемы, связанных с качеством данных, такие как неточность, пропущенные значения и расхождения. Критерий достоверности акцентирует внимание на проблемах обеспечения доверия к огромным объемам собранных данных и их пригодности для значимого анализа. Как справедливо отмечает О. Петровская, «огромный информационный поток (Большие данные) составляет массив данных, с которым нужно работать и принимать необходимые меры для обеспечения (сохранения) достоверности информации» [32, с. 130]. Большие данные собираются из различных источников информации, включая пользовательский контент, данные датчиков и прочее, в связи с этим существует риск несоответствия, неточности, а также отсутствия значений.

Ошибки в процессе сбора данных из различных источников могут привести к отклонениям в наборе данных, что, в свою очередь, оказывает влияния на их предвзятость. Погрешности в методах сбора или выборки данных оказывают влияние на репрезентативность набора данных, а также это

приводит к искажению результата анализа. Решение проблемы достоверности требует скрупулезного подхода к процессам очистки, контроля и обеспечения качества информации.

Происхождение данных и прозрачность источников информации играют важную роль в поддержании целостности больших данных, так как они предоставляют сведения о том, как были собраны и обработаны данные. Такие процедуры регулируются техническими стандартами по работе с большими данными. А. Попов отмечает, что «существующие стандарты и регламенты играют важную роль в обеспечении безопасности и качества данных, связанных с Big Data» [35, с. 70].

Несмотря на то, что большие данные определяются разными концепциями nV , где n – это число свойств, а V – это первая буква названия каждого свойства, все эти концепции имеют общую основу – их истинная ценность заключается в потенциальном вкладе, который они могут принести предприятиям, организациям, исследователям и т.д. В этом плане ценность относится к способности извлекать значимую и полезную информацию из крупных и сложных массивов данных.

Аналитика больших данных способна обнаруживать скрытые паттерны и корреляции в массивах данных, которые могут оставаться незамеченными в небольших наборах. По мнению Н. Белодеда и Е. Хорошуна, «процессы анализа больших данных становятся основой для принятия стратегических решений» [8, с. 496].

Анализ исторических данных и данных в реальном времени обеспечивает прогностическое моделирование, позволяющее организациям предсказывать тренды, поведение клиентов и возможные проблемы. В ситуации использовании больших данных в энергетической отрасли М. Тимохин и В. Шаранин отмечают, что «они позволяют современным компаниям эффективно управлять энергетическими процессами,

прогнозировать спрос и производство энергии, оптимизировать распределение ресурсов, повышать энергетическую эффективность и интегрировать возобновляемые источники энергии в сети» [49, с. 33].

В работе «Big Data and advanced analytics в образовании» С. Бекназарова рассматривает использование больших данных в сфере образования для персонализации учебного процесса и обращает внимание на то, что «перспективы создания персонализированных обучающих программ и предоставления целенаправленной поддержки студентам делают эту область обещающей для будущего развития образования» [7, с. 76]. Также большие данные способствуют индивидуализации рекомендаций в области электронной коммерции, распространения контента и здравоохранения. Приспособление продуктов, услуг и контента к индивидуальным потребностям повышает удовлетворенность и уровень вовлеченности клиентов.

На основе анализа больших данных, предприятия могут повышать эффективность операций, цепочек поставок, распределения ресурсов, а также сократить ненужные затраты и увеличить эффективность. Большие данные стимулируют инновации, создавая фундамент для разработки новых товаров, сервисов и бизнес-моделей. Они поощряют экспериментирование и поиск новых решений. М. Комилов акцентирует внимание на том, что «компании, активно использующие анализ больших данных, получают конкурентные преимущества в виде более точного понимания рынка, предоставления персонализированных услуг и оперативного реагирования на изменения» [62, с. 43].

Использование больших данных подразумевает применение передовых методов анализа, алгоритмов машинного обучения и средств визуализации данных для получения полезной информации. Как в исследовательском сообществе, так и среди реальных практик не существует какого-либо

универсального инструментария и технологических подходов для работы с большими данными. Область аналитики больших данных опирается на широкий спектр инструментов и технологий, разработанных для решения задач, связанных с колоссальными объемами данных.

Все инструменты, для работы с большими данными направлены на трансформацию значительных объемов данных в более удобные для работы формы, то есть малые данные. Для работы с малыми объемами данных применяется свой набор инструментов. На этот счет у исследователей и практиков также нет общего мнения о том, из каких инструментов и технологических подходов состоит данный набор.

Таким образом, выбор инструментария для работы с данными должен определяться, исходя из специфики конкретной задачи и требований к результатам анализа. В частности, работа с большими данными предполагает использование специализированных платформ и алгоритмов, способных эффективно обрабатывать значительные объемы информации, а работа с малыми объемами данных может потребовать применения более гибких инструментов, позволяющих быстро анализировать и визуализировать данные для принятия оперативных решений.

Сегодня практически каждая платформа предлагает создать учетную запись для авторизации в системе при каждом посещении сайта – это позволяет компаниям сохранять данные о том, что просматривали, приобретали и изучали пользователи.

С этими же целями компании и организации используют карты лояльности, которые покупатели предъявляют каждый раз при оплате товаров. За счет них клиенты также получают определенные скидки и поощрения в виде каких-либо бонусов и персональных дополнительных акций. Основной целью программ лояльности является сбор личных данных пользователей, которые помогают компаниям наладить коммуникацию с покупателями.

«Успешность программы лояльности определяется ценностью собранных данных, получаемых в рамках программ лояльности» [21, с. 112], – поясняют О. Иванченко и Е. Барауля.

Раскрытие пользователем информации о себе при регистрации на какой-либо онлайн-платформе несет для него не только риски, но и преимущества. Поскольку, собирая данные о клиенте, онлайн-платформа и сервисы анализируют их, а затем используют полученную в результате анализа информацию для создания наиболее подходящего предложения для конкретного потребителя. Так, бизнесом обширные данные используются для исследования поведения пользователей и более индивидуального общения с ними. Производители предлагают потребителю продукт, который требуется именно ему, основываясь на его предыдущих покупках, поисковых запросах и прочей информации.

Помимо разработки предложений для пользователей, компании используют собранные данные о клиентах для улучшения собственной работы. Компании относятся к собственным данным с большим доверием, нежели к внешним, собранным сторонними организациями. С этой целью они делают упор на внутренние проекты в области анализа больших данных, создают внутренние отделы по их обработке и привлекают профильных специалистов.

Описанная выше ситуация говорит о том, что анонимное использование данных постепенно становится пережитком прошлого. Людям приходится мириться с тем, что практически все, с чем они взаимодействуют – оставляет за собой цифровой след, по оставленным данным формируется цифровой профиль человека.

Согласно исследованию об отношении граждан к утечкам персональных данных, проведенному аналитическим центром группы компаний «Гарда», 81% россиян обеспокоены тем, что может произойти утечка их персональных

данных [13]. Хотя зачастую проблемы, связанные с защитой персональных данных, вызваны действиями самих субъектов персональных данных, которые размещают их в открытых источниках.

В качестве субъекта данных может выступать физическое или юридическое лицо, которое обладает законными полномочиями для осуществления сбора, обработки и хранения данных. Такой субъект является вправе предоставить результаты обработки этих данных другим сторонам в соответствии с договором на предоставление информационных услуг. В научной литературе эту категорию называют информационными брокерами или брокеры данных. И. Аюшеева отмечает, что «при осуществлении такого рода деятельности недопустимо ущемлять личные неимущественные и иные права субъекта персональных данных: закрытая (конфиденциальная) информация может быть обработана только при условии получения согласия обладателя и обеспечения правомерного доступа к ней» [4, с. 129].

Если говорить о социальных сетях и других платформах, доступ к которым имеют физические лица, информация о пользователях и их персональных данных изначально размещается в базе данных самими пользователями этих социальных сетей и платформ с их согласия. Таким образом наполнение базы данных осуществляется не компаниями или организациями социальных сетей или платформа, а непосредственно самими пользователями. Обработка их персональных данных производится на основании согласия субъекта персональных данных, которое требуется для размещения личной информации в сети. Отношения между компаниями или организациями веб-ресурсов и пользователями регулируются договором, в соответствии с которым последние получают доступ к информационной системе и услугам сервиса. Такие договоры называются «пользовательским соглашением\согласием» или «правилами использования сайта\сервиса\ресурса» [11]. Их заключение происходит в форме электронного

документа, путем обычного клика, который подтверждает согласие лица с перечисленными в документе условиями. В. Шайдуллина подчеркивает, что «в эпоху big data концепция информированного согласия на обработку персональных сведений не обладает должной эффективностью, и тому есть несколько причин: невозможность предоставления исчерпывающего объема информации о целях и способах обработки данных; неспособность субъекта персональных данных к адекватному восприятию такой информации; невозможность индивидуально взаимодействовать с огромным числом компаний, занимающихся сбором и обработкой личных сведений» [57, с. 53].

Таким образом, часть информации, включая персональные данные, становится общедоступной после ее публикации на соответствующем ресурсе, что приводит к возникновению возможности ее обработки третьими лицами.

В сентябре 2022 года завершился почти шестилетний судебный спор между компаниями «ВКонтакте» и Double Data: социальная сеть пыталась защитить данные своих пользователей от их использования [16]. «ВКонтакте» просила суд признать базу данных своих пользователей интеллектуальной собственностью компании. Double Data же считала, что данные пользователей социальной сети не могут принадлежать самой «ВКонтакте», поскольку пользователи сами создали в ней свои страницы, тем самым дав согласия на то, что любой посетитель профиля может использовать личные сведения. Как итог, иск «ВКонтакте» так и не был удовлетворен.

Регулирование работы с большими данными в разных странах мира различается в зависимости от их законодательства и политики в области защиты персональных данных. Например, в странах Европейского Союза принят «Общий регламент по защите данных» (GDPR) [31], который устанавливает строгие правила для сбора и использования персональных данных. Этот закон требует, чтобы компании получали согласие от

пользователей на сбор и использование их данных, а также предоставляли им право на доступ, изменение и удаление своих данных.

В Китае регулирование работы с большими данными осуществляется через «Закон о защите личной информации» (Personal Information Protection Law) и «Закон о кибербезопасности» (Cybersecurity Law). Эти законы устанавливают требования по сбору, хранению и использованию персональных данных.

В России нет специальных законов, регулирующих использование больших данных. Однако есть законы, которые касаются обработки персональных данных и защиты конфиденциальной информации. К ним относятся: Федеральные законы «О персональных данных» от 27.07.2006 N 152-ФЗ [40] и «Об информации, информационных технологиях и о защите информации» от 27.07.2006 N 149-ФЗ [41], «Гражданский кодекс Российской Федерации» [38], «Трудовой кодекс Российской Федерации» от 30.12.2001 N 197-ФЗ [42], закон «О коммерческой тайне» от 29.07.2004 N 98-ФЗ [39] и другие.

Кроме того, в России действует система защиты персональных данных, которая включает в себя механизмы контроля и надзора за соблюдением требований законодательства. Главным органом по контролю за соблюдением законодательства в области персональных данных является «Роскомнадзор» [37].

Система обработки данных в энергетических сетях включает многочисленные приборы учета энергии, которые располагаются на значительном удалении от серверов. Для решения задач по обработке большого количества информации в энергетике используются инструменты из сферы больших данных – об этом пишет А. Аникеева в статье «Применение технологии «Большие данные» в энергетике» [2].

Технологии больших данных активно применяются и в сфере электронной коммерции и розничной торговли: компании и организации собирают разнообразные данные о клиентах (телефоны, электронные почты, даты рождения, поисковые запросы и т.д.) с целью последующего их анализа. Затем результаты используются для улучшения качества сервиса, более точного индивидуального предложения клиентам, оптимизации рабочих процессов компании, а также увеличения объем покупок. «Аналитика больших данных трансформирует отрасль розничной торговли, именно вопрос грамотной обработки и использования имеющегося массива данных становится главной задачей для ретейлеров» [29, с. 133], – отмечает М. Нигматуллина.

В аграрном секторе генерация больших объемов данных происходит за счет разнообразных датчиков на полях и фермах, а также прочих производственных участках, которые отслеживают финансово-экономические, организационные и производственно-технологические процессы [47]. На основе данных даются прогнозные оценки и применяются управленческие решения.

В мире больших данных традиционные границы между отраслями и дисциплинами размываются, а возможности для инновационного развития и экономического роста становятся безграничными.

Таким образом, большие данные являются определяющей чертой современного цифрового общества, трансформируя способы его существования, работы и взаимодействия с окружающей средой. Феномен больших данных представляет собой сдвиг парадигмы беспрецедентного масштаба, который открывает новую эру инноваций и открытий во всех сферах деятельности. Уже сегодня их применение производит революцию во множестве различных отраслей: оптимизация бизнес-стратегий, улучшение результатов образования, продвижение медицинских исследований – все это

возможно благодаря преобразующему потенциалу больших данных, который ограничивается только человеческим воображением и этическими соображениями. Учитывая продолжающуюся цифровизацию всех аспектов жизни, экспоненциальный рост данных порождает беспрецедентные возможности и вызовы. Многогранная природа больших данных позволяет характеризовать их по нескольким критериям, популярными из которых являются: огромный объем, высокая скорость как генерации, так и обработки, разнообразие источников данных и критическая значимость их достоверности. Сравнение больших данных с их меньшими аналогами (малыми данными) позволяет выделить их уникальную парадигму, которая заключается в преобразующей способности извлечения уникальной информации из огромных и динамичных наборов данных. Несмотря на инновационные преимущества больших данных, учеными и медиапрактиками признается существование проблемных юридических и этических аспектов их использования. В настоящее время действующее законодательство регулирует лишь вопросы конфиденциальности и ответственного управления данными организациями и отдельными лицами.

1.2 Применение больших данных в журналистской деятельности

Цифровая трансформация и развитие информационно-коммуникационных технологий в сфере медиа коренным образом трансформировали методы использования данных и подходы к их аналитической обработке. Следствием трансформации стало появление в журналистике такого направления, как журналистика данных, которая связана с процессами поиска, сбора, обработки и визуализации данных с целью их дальнейшего представления в журналистских материалах с использованием инфографики. «Основной особенностью дата-журналистики выступает то, что

в качестве источника используются не мнения экспертов, пресс-релизы или свидетельства очевидцев, а данные» [53, с. 65], – отмечает У. Хапов.

Деятельность данного направления может быть охарактеризована как процесс извлечения значимой информации из данных с целью последующего создания на их основе журналистских материалов и включения в них визуализации (в некоторых случаях интерактивных), наглядно демонстрирующих пользователям суть истории. Б. Херави и М. Лоренц определяют журналистику данных «как направление журналистики, в котором ключевую составляющую определяют данные, ... в котором объединены знания нескольких дисциплин, включая журналистику, социальные науки, информационные технологии, анализ данных, дизайн, и сторителлинг» [61, с. 26].

Несколько лет назад в научном дискурсе развернулась дискуссия об использовании больших данных в журналистике. В ней побеждало мнение о том, что журналисты не работают с большими данными, поскольку у них отсутствуют необходимые компетенции. Эта точка зрения основывалась на том, что журналистика – это гуманитарное направление, которое включает в себя сбор, анализ и интерпретацию информации для создания различных журналистских материалов и других форм медиаконтента.

Журналистика данных является одним из направлений журналистики, которое руководствуется теми же принципами создания качественного журналистского материала, что и другие ее направления. Ее отличительная черта состоит в том, что информация извлекается из данных, а не из традиционных источников, которые предоставляют информацию в готовом виде. М. Безденежных пишет: «Журналистика значительно трансформировалась вследствие новых технологий, однако ее суть осталась прежней, изменились лишь формы подачи информации» [6, с. 228].

Еще одной чертой журналистики данных является использование визуализаций и инфографики, которые используются для наглядного, доступного для широкой аудитории представления сложной, но значимой социальной информации.

Визуализация в дата-материалах может быть статичной или интерактивной. Статичная визуализация подразумевает представление только самих данных. Интерактивная визуализация позволяет аудитории самостоятельно исследовать данные. Как правило, такая визуализация первоначально представляет собой обзор данных, но в ней также присутствуют дополнительные инструменты для изучения деталей.

Визуализация данных способствует улучшению качества материала путем упрощения его представления в виде инфографики, предназначенной для освещения какой-либо сложной темы. Использование в журналистских материалах инфографики для визуализации данных – это новый стандарт представления оцифрованной информации. С. Симакова в монографии «Медиаэстетический код инфографического контента в журналистике» высказывает, что «инфографика занимает все более устойчивую позицию, переходит из разряда “новинок” ... в статус полноправной части новостного контента, а в отдельных случаях оказывается частью общего стандарта подачи новостей (то есть становится неизбежным элементом информационных потоков)» [44, с. 94]. Однако исследователи Э. Кларк и Д. Родригес отмечают, что «не каждая история поддается визуализации в том смысле, что некоторые истории лучше излагать традиционным письменным методом» [60, с. 82].

Развитие технологий и увеличение доступности данных в цифровом формате привело к трансформации журналистики в междисциплинарную дисциплину, которая включает в себя не только гуманитарные навыки, но и технические знания и цифровые навыки. В связи с этим спустя некоторое время появилась точка зрения, согласно которой журналистику данных

необходимо отделять от журналистики больших данных. Основное различие между этими двумя направлениями заключается в объеме данных, которые используются для анализа и интерпретации информации. Журналистика данных обычно использует небольшие объемы данных, которые могут быть извлечены из структурированных баз и проанализированы вручную или с помощью простых инструментов. Журналистика больших данных, напротив, использует большие объемы данных, которые требуют специализированных инструментов и методов для обработки и анализа.

Появление и распространение огромного количества инструментов для работы с большими массивами данных, основанных на искусственном интеллекте, постепенно стирает грани между журналистикой, журналистикой данных и журналистикой больших данных.

Множество приложений и программных продуктов позволяют обычным журналистам работать с большими данными. Тем не менее они все равно не владеют компетенциями на таком уровне, чтобы не испытывать трудностей при работе с неструктурированными большими данными. В. Файль и О. Кунгурова связывают это с низкой поддержкой направления журналистики данных в редакциях, а также с тем, что «у многих журналистов смутное представление о том, как необходимо создать дата-сети, как с ними работать и в чем их особенность» [51, с. 652].

Медиаорганизациям, использующим большие данные, приходится прибегать к специальному программному обеспечению, позволяющему собирать, расшифровывать неструктурированные массивы данных и организовать их в формат, подходящий для работы обычного журналиста. В последние годы на помощь им приходят инструменты интеллектуального анализа данных, такие как искусственный интеллект и нейронные сети. Так, И. Деева и А. Иляхина приводят примеры использования этих инструментов: «... журналисты используют голосовые помощники при написании новости с

диктофона, применяют текстовые расшифровщики аудиозаписей, используют автоматические переводчики текста, осуществляют с помощью специальных алгоритмов выжимку главной информации из больших по объему текстов и осуществляют поиск актуальных бэкграундов, применяют программы для генерации иллюстраций к новостям, проверки информации на достоверность и осуществления аналитики текстов» [22, с. 583].

Нейронные сети и искусственный интеллект применяются для извлечения из больших объемов данных новой, нестандартной информации, которая в дальнейшем может быть использована в качестве основы для создания журналистского материала. А. Арсентьева и А. Морозова отмечают, что «... внедрение ИИ при подготовке публикаций предполагает облегчение работы сотрудникам редакций, ускорение выхода материалов, связанного с анализом больших данных, фактчекингом, но при этом не лишает людей их работы» [3, с. 9].

Правильное использование этих инструментов открыло новые возможности для медиаиндустрии и облегчило работу журналистам, но, как справедливо отмечает Л. Циньян: «в то же время заставило их столкнуться с новыми вызовами и проблемами» [64, с. 539]. Одним из таких вызовов стала необходимость овладения новыми профессиональными компетенциями.

В научном сообществе у такого набора компетенций есть определенный термин – цифровой капитал – индивидуальные знания, навыки и способности каждого индивида по использованию цифровых технологий, платформ социальных сетей и других Интернет-инструментов для достижения как личных, так и профессиональных целей. Е. Варганова и А. Гладкова считают, что «чем больше объем цифрового капитала пользователей, тем соответственно большим количеством преимуществ – в экономической, политической, культурной и других сферах жизни – эти пользователи

обладают» [10]. Наряду с понятием цифрового капитала существует другой термин – цифровая компетенция.

В этой дефиниции требуется определения этого термина. В научном сообществе нет единого мнения по этому вопросу. В большинстве работ цифровая компетенция характеризуется как совокупность знаний, умений и навыков, необходимых для выполнения какой-либо работы, связанной информационно-коммуникационными технологиями. В работе «Цифровые компетенции: понятие, виды, оценка и развитие» исследователи И. Симарова, Ю. Алексеевичева, Д. Жигин делят цифровые компетенции на базовые и специальные, объясняя это тем, что «... приобретение специальных цифровых компетенций требует профессионального обучения, например, написание программного обеспечения с использованием языков программирования» [45, с. 942].

Таким образом, в контексте профессии журналиста под цифровым капиталом можно понимать набор специальных цифровых компетенций, которыми специалист должен обладать для успешного выполнения своих профессиональных обязанностей и достижения поставленных целей.

Н. Макарова в статье «Журналистика данных в системе профессиональных компетенций журналиста» производит сравнение двух направлений (коммуникационных технологий и средств массовой информации) со стандартами знаний и умений дата-журналиста и приходит к выводу, что «инструментарий, который нужен для работы с big data слишком специфичный для того, чтобы изучать его в процессе обучения в вузе, но при этом абсолютно необходимый, чтобы без него бывшего студента могли принять на работу» [26, с. 50]. В таком случае, к определению набора специальных цифровых компетенций журналиста следует подходить с позиции его базовых операции работы с данными:

- поиск и извлечение данных из различных источников, таких как официальные статистические данные, отчеты, исследования, интервью, социальные сети и другие;
- очистка данных, проверка на наличие ошибок, пропусков или дубликатов, удаление и фильтрация ненужных или некачественных данных фактчекинг;
- анализ данных, выявление трендов, связей, закономерностей;
- интерпретация данных, проведение сравнения результатов с другими исследованиями или данными, выводы на основе полученной информации;
- создание материала на основе результатов анализа данных, использование инструментов визуализации данных для представления информации в доступной и понятной форме.

Таким образом, цифровые компетенции современного журналиста целесообразно определять в соответствии с базовыми операциями по работе с данными.

Например, Е. Баранова и А. Шнайдер сообщают, что дата-журналисту «необходимо владеть навыками программирования (знание Python), работы с определенными техническими инструментами, такими как извлечение данных (Tabula, document cloud), очистка и анализ данных (Google spreadsheets, open refine), инструментами визуализации данных (Datawrapper, infogram, flourish)» [5]. Другой набор компетенций выводит М. Ким в работе «Инновационные практики в работе мультимедийных журналистов» на основе результатов опроса региональных журналистов сетевых изданий: «82% опрошенных нами респондентов отметили, что современному журналисту необходимо владеть навыками веб-верстки, видео- и аудиомонтажа, анализа и обработки больших данных, SMM, инфографики и мн. др.» [24, с. 77].

Еще одним фактором определения компетенций современного журналиста является его роль в рабочем коллективе. Для полноценного создания материалов на основе данных медиаорганизации создают собственные отделы по работе с большими данными, так называемые дата-отделы. В. Бондарчик, анализируя результаты опроса Европейского центра журналистики «Состояние журналистики данных 2021», приходит к выводам, что «только 22% дата-журналистов работают в специальных отделах обработки данных, при этом большее количество работает в редакциях, которые редко публикуют проекты, основанные на данных (25 %)» [9, с. 21]. Согласно этому же опросу и результатам других научных исследований, существуют различные модели работы отделов:

- совместная работа, в рамках которой журналисты в сотрудничестве с программистами, веб-разработчиками, дизайнерами, аналитиками данных и другими специалистами работают над созданием журналистских материалов;
- индивидуальная работа журналистов, владеющих компетенциями, позволяющими самостоятельно программировать, анализировать, и визуализировать данные, в соответствии с принципами медиадизайна.

Также существует еще одна модель работы медиаорганизации, которая занимается анализом данных не только для достижения своих целей, но и предлагает свои услуги другим медиакомпаниям. В работе «Специфика работы объединений дата-журналистов в Швеции» П. Давыдова объясняет это тем, что «дата-журналисты предлагают себя в качестве помощника как для своих коллег, так и для читателя, не только объясняя тенденции, но и предлагая технологичные решения проблем» [15, с. 69].

Таким образом, цифровой капитал современного журналиста включает комбинацию нескольких компонентов:

- навыки работы с данными, готовность использовать для этого различные инструменты и методы работы, умение правильно интерпретировать и представлять данные;
- знание технологий на основе искусственного интеллекта и нейросетей;
- понимание этических принципов работы с данными, осведомленность о конфиденциальности и защите персональных данных;
- навыки коммуникации, эффективного общения с различными источниками данных, такими как эксперты, ученые или представители организаций, со своими коллегами в коллективе, в том числе с применением цифровых устройств;
- умения создавать контент: интересные и информативные материалы на основе данных, в том числе с использованием инструментов визуализации данных;
- готовность к постоянному обучению и развитию компетенций, освоению новых технологий и методов для работы с данными.

Оценка цифрового капитала журналиста в рамках работы с большими массивами данных остается затруднительной. Прежде всего в аспекте навыков программирования. Программирование предоставляет возможность создавать индивидуальные решения, которые могут быть более эффективными и точными в решении определенных задач. Также программирование позволяет создавать интерактивные визуализации данных, которые позволяют аудитории взаимодействовать с данными и получать более глубокое понимание информации. Кроме того, владение навыками программирования позволяет более глубоко производить анализ и работу с данными.

Основным преимуществом программирования является возможность автоматизации различных процессов. Автоматизация подразумевает создание программных систем, которые выполняют определенные задачи без участия

человека. В этом плане владение программированием позволяет отойти от того, что уже предлагается в виде готовых решений и разработать свой уникальный продукт, который может быть адаптирован под конкретные потребности и требования. В сфере журналистики с помощью программирования автоматизировать можно практически любую операцию технологического цикла работы с данными, можно даже настраивать степень автоматизации той или иной операции. К примеру, можно автоматизировать процессы сбора и обработки.

Н. Самойленко обращает внимание на специфику продукта журналистского творчества: «Сегодня это гибридный информационный продукт, успех которого в равной степени зависит как от таланта журналиста, так и от технического исполнения» [43, с. 127]. Комбинацию журналистских и технических составляющих можно ярко наблюдать во время проведения специальных мероприятий, так называемых «хакатонов» (от английского «hackathon», где «hack» – взламывать, а «marathon» – марафон). Целью этих мероприятий является работа специалистов из разных областей над решением какой-либо проблемы. Так, на «хакатонах» журналистам предлагают придумать новые форматы подачи информации, инструменты для сбора, анализа и визуализации данных и т.д. Это оптимальное решение проблемы, о которой говорят исследователи Р. Ерженин, З. Бахвалова, Е. Волков и А. Абзаев: «Главная проблема для журналиста, знающего о возможностях дата-журналистики, заключается в сложности обработки и представлениях этих данных в привлекательном для читателя виде. В то же самое время в среде разработчиков программного обеспечения знают, как организовать и обеспечить информационные процессы обработки больших данных, но не знают, как представить результаты этой работы в значимом для читателя виде» [17, с. 113].

Несмотря на то, что навыки программирования могут быть полезны для журналиста, обладание ими пока что не является обязательным. Сегодня существует множество инструментов и сервисов, которые позволяют журналистам работать с данными, в том числе большими, без необходимости программирования.

Одним из таких инструментов стало появление и развитие такой технологии, как чат-боты. Существует мнение, что чат-боты появились сравнительно недавно, для решения коммерческих задач. Однако их история началась раньше, в прошлом веке с создания программы, способной естественным образом общаться с людьми.

Термин «чат-бот» в зависимости от области его применения определяется по-разному. Так, А. Шах и Е. Шапович в работе «Чат-боты как современный инструмент маркетинга» определяют его как «виртуального собеседника, который работает на основе установленных правил и алгоритмов» [58, с. 200]. Т. Цапина считает, что «чат-боты – это форма диалогового искусственного интеллекта, а также интерфейс, позволяющий наладить взаимодействие между человеком и компьютерной системой» [55, с. 347]. Сегодня чат-боты связывают с развитием интерактивных технологий.

В научной сфере концепция интерактивности трактуется по-разному в зависимости от принципов и способов её организации: для взаимодействия между пользователями, между пользователями и программой; для обеспечения возможности пользователям реагировать тем или иным способом на потребляемый контент и т.д. Наиболее исчерпывающее определение термина характеризует интерактивность как степень, в которой две или более стороны коммуникации могут воздействовать друг на друга, на среду коммуникации, на сообщения, а также на степень синхронизации таких воздействий. В таком виде дефиниция охватывает большинство аспектов интерактивности в том виде, как она представлена в журналистском дискурсе

и в СМИ как платформе распространения разнонаправленных потоков информации. «Именно благодаря интерактивности пользователь включается в цифровую медиасреду» [48, с. 91], – отмечает В. Соломин.

В журналистике и СМИ использование чат-ботов является относительно новым явлением. Оно стало возможным благодаря внедрению вычислительных процедур в практику журналистской профессии.

Расширение числа чат-ботов потребовали разработку их классификации. Авторы статьи «Разновидности чат-ботов и их роль в современной журналистике» [34, с. 158] Р. Погудина и А. Шубина предложили делить чат-боты на пять групп на основании их взаимодействия с пользователями. Так, они в работе выделяют: чат-боты рассылки, боты помощники, боты-игры, боты собиратели, боты-генераторы идей и контента. Иную классификацию разработала А. Хрущева [54]. В основу деления она положила степень сложности программы: простые – основываются на заранее запрограммированных алгоритмах и инструкциях, сложные – используют искусственный интеллект.

Существуют чат-боты, функционирующие на основе нейросетей и искусственного интеллекта, а также программы-парсеры данных, то есть программы автоматического сбора данных. Их использование позволяет журналистам осуществлять автоматическую сборку и обработку больших объемов данных, проводить фактчекинг (проверку достоверности информации), улучшать свой контент за счет возможности генерации текстовых и визуальных материалов.

Другие чат-боты разработаны с целью поддержания связи с аудиторией. Эти чат-боты создаются с целью вовлечения аудитории в создание разнообразного контента, а также для усиления интерактивности и повышения эффективности распространения новостей. Использование таких чат-ботов, в основном, приходится на социальные сети и мессенджеры. Медиаорганизации

расширяют тем самым информационное поле, которое они могут охватить своей повесткой. Сегодня практически каждая медиаорганизация использует подобных чат-ботов для создания каналов взаимодействия со своей аудиторией. Их использование также позволяет журналистам персонализировать контент, организовать рассылку уведомлений и обеспечить обратную связь, а также вовлекать пользователей в создание контента.

Персонализация контента заключается в предоставлении возможности аудитории формировать личную ленту новостей, получать интересную и полезную им информацию. «Все медиатренды ведут к тому, что журналистика станет максимально персонализированным сервисом, и чат-боты на данный момент являются частным примером развития этого процесса» [20, с. 131], – отмечает А. Иванов.

В предоставлении возможности для аудитории внести вклад в создание контента реализуется принцип «партисипативной» или гражданской журналистики. В статье «Гражданская журналистика как объект развития цифровой журналистики» К. Решетникова отмечает, что «в результате развития современных технологий гражданская журналистика обрела новую жизнь, поскольку любой человек, обладающий техническим оснащением, может записывать новости и распространять их по всему миру» [36, с. 118]. Дополнительным стимулом вовлечения аудитории в создание контента служит материальное поощрение, которое медиаорганизации готовы предоставить за предложенную информацию в зависимости от ее уникальности и новизны.

Таким образом, применение чат-ботов в создании журналистских материалов свидетельствует о том, что развитие интерактивных технологий достигло нового уровня. Чат-боты превратились в эффективный инструмент не только для работников СМИ, но и для пользователей, которым предоставляется удобный канал для взаимодействия с медиаорганизациями.

Современный человек, взаимодействуя со множеством информационных источников и проводя много времени в интернете, производит большие данные с помощью различных устройств (смартфонов, планшетов, ноутбуков и т.д.). Согласно десятому отчету цифровых новостей Reuters Institute, использование смартфонов для просмотра новостей росло самыми быстрыми темпами в течение многих последних лет, особенно во время карантина, объявленного в связи с эпидемией коронавируса во всем мире [63]. Стремительный рост использования мобильных устройств для просмотра новостных ресурсов открыл доступ к индивидуализированным информационным средствам.

Журналистика, следуя примеру других областей деятельности, начала использовать возможности, открываемые большими данными, для извлечения из них ценной информации, которую можно применить при разработке успешных маркетинговых стратегий и повышения эффективности публикаций, а также для формирования новых источников дохода. Так, А. Шилина отмечает, что «Цифровизация и датафикация в современном мире производят изменения параметров журналистики: происходит увеличение объемов цифровой информации, изменяются ее форматы и появляются новые, контент становится интерактивным, в результате подобных трансформаций и дальнейшего распространения новых технологий источники больших объемов данных и сервисов становятся доступными для изучения и работы журналистов» [59, с. 13]. Таким образом, «большие данные как социальный, культурный и технологический феномен служат концептуальным «микроскопом» в понимании журналистики и как профессиональной сферы, и как коммерческого предприятия» [1, с. 109], – отмечают Г. Шаймерген и С. Абдиназар. Т. Новосильцева указывает на роль человека в работе с большими данными, называя ее центральной «в группировке информации с помощью технологий» [30, с. 87].

Использование больших данных в журналистике влияет не только на производственно-творческий процесс – оптимизацию журналистского повествования и создание историй на основе скрытых данных, но и на редакционные, и на маркетинговые процессы.

Следуя современным трендам, медиаредакции подстраивают свою работу под аналитику, полученную на основании анализа пользователей, использующих приложения этих медиа в качестве источника новостей. Даже простейшие данные, такие как «тапы» и «смахивания», являющиеся обычными действиями пользователей в приложениях, могут существенно помочь медиаорганизациям в улучшении качества и визуального наполнения предлагаемого ими контента.

Современному пользователю, помимо приложений, просмотр различного медиаконтента доступен на множестве онлайн-источников: сетевых изданий, социальных сетей, электронных рассылок и т.д. Средства массовой информации используют аналитический подход для улучшения взаимодействия с пользователями и рекламодателями и для обеспечения своевременной доставки контента на нужное устройство для нужной аудитории.

С помощью аналитики больших данных медиаорганизации могут получить представление о предпочтениях аудитории и моделях потребления контента. Достигается это путем отслеживания и анализа поведенческих показателей таких как, клики, время, проведенное на странице, совместное использование контента и взаимодействие с мультимедийными элементами. Поведенческий анализ помогает выявить какие темы, форматы и типы контента больше всего резонируют с аудиторией. Это позволяет медиаорганизациям соответствующим образом оптимизировать свою контент-стратегию и создавать более привлекательный и актуальный контент.

Анализ настроений представляет собой использование алгоритмов обработки естественного языка (NLP) и методов машинного обучения для анализа текстовых данных, включая комментарии, отзывы и сообщения в социальных сетях, с целью определения настроений или эмоционального тона аудитории. Медиаорганизации используют анализ настроений для оценки общественного мнения по конкретным вопросам, продуктам и или событиям, а также для понимания отношения аудитории к их контенту и бренду. Полученные в ходе этого результаты помогают организациям определить области для улучшения, решить проблемы аудитории и адаптировать свои сообщения таким образом, чтобы они лучше соответствовали с запросами аудитории.

Современные средства аналитики позволяют медиаорганизациям в режиме реального времени осуществлять мониторинг показателей активности аудитории и эффективности своей работы. Это позволяет определить, в какое время суток пользователи чаще всего просматривают, делятся и оценивают тот или иной контент, а также позволяет установить, с каких типов и моделей устройств это осуществляется. Непрерывный анализ этих показателей позволяет своевременно принимать обоснованные решения. Например, такие как выявление актуальных тенденций, определение новых информационных поводов и реагирование на мнения аудитории. В целом, это касается внесения корректировок в редакционную политику организации, выбора каналов распространения информации и определения наиболее эффективных способов взаимодействия с аудиторией с целью достижения ее максимальной вовлеченности.

Социальные сети играют центральную роль в потреблении новостей, а также в персонализации их потоков. Большинство пользователей регулярно посещает одну или несколько социальных сетей или приложений для коммуникации, а также используют их для получения, распространения и

обсуждения новостей. Данные о пользователях, собираемые на основе их активности в социальных сетях, зачастую выявляют ранее неизвестные факторы, которые могут быть использованы в интересах медиаорганизаций.

Например, большие данные позволяют им определить спрос на различные виды новостей и категории контента для разных возрастных групп. Так, они используются для сегментации аудитории на основе различных критериев, в числе которых могут быть такие, как демографические данные, географическое положение, интересы и т.д. Анализируя данные, поступающие из таких источников, как трафики веб-сайтов, взаимодействия в социальных сетях (данные о подписках) и т.п., медиаорганизации способны определять отдельные сегменты аудитории и адаптировать контент и маркетинговые стратегии к предпочтениям каждой группы.

Применение медиаорганизациями больших данных для решения маркетинговых задач способствует укреплению связей производителей информации с аудиторией, в том числе через понимание ее предпочтений, поведенческих моделей и настроений. Так, основываясь на персональных предпочтениях каждого пользователя: истории его просмотров, лайков и другой его активности, медиаорганизации формируют индивидуальную информационную повестку, а также готовят системы рекомендаций, функционирующие на базе алгоритмов машинного обучения, анализируют большие объемы информации для прогнозирования популярности тех или иных материалов. Персонализация контента (или «таргетирование» аудитории) обеспечивает повышение вовлеченности и лояльности аудитории.

Основываясь на больших данных как источнике информации, журналисты могут создавать оригинальный контент. Однако существует объективное ограничение данного вида деятельности – уровень цифровой компетенции производителя. Работа с большими данными требует

использования инструментов искусственного интеллекта и языков программирования.

Языки программирования позволяют работать как уже с готовыми дата-сетями с целью извлечения ценной информации из них, которая может быть полезна при написании материала, так и с полу-структурированными и неструктурированными видами данных. Преимущество языков программирования заключается в возможности разработки специальной программы по сбору необходимых данных с большого множества различных источников. Это позволяет не только отслеживать какие-либо изменения, но и формировать собственные наборы данных.

Собранные и обработанные данные можно визуализировать с помощью тех же языков программирования. Это касается не только создания базовых диаграмм и графиков, таких как гистограмм, линейных графиков, круговых диаграмм, но и динамических с элементами всплывающих подсказок, масштабирования, фильтрацией и т.д. Все это дает возможность аудитории детально изучать представленные данные, что помогает ей обрести более глубокое понимание.

Алгоритмы генерации естественного языка, анализируя большие данные, способны автоматически создавать письменный контент, включая новостные статьи, выжимки и отчеты. Использование таких инструментов не только экономит время журналистов, освобождая им время для более творческих задач, но и позволяет обрабатывать большое количество информации с относительно малой потерей качества. Для создания мультимедийного контента используются инструменты на базе искусственного интеллекта для создания фото и видео изображений, что позволяет не просто дополнить нарратив, но и обогатить материалы визуальными элементами.

Взаимодействие с инструментами создания контента происходит через веб-интерфейс или специальное программное обеспечение. В некоторых случаях доступ к этим инструментам может быть ограничен либо требовать каких-либо действий со стороны пользователя (например, регистрации, подтверждения своих учетных данных и т.д.). При получении доступа к инструменту, работа с ним заключается в написании так называемых «промптов» (от английского *prompt* – запрос) или по-другому «команд\инструкций», на основе которых происходит генерация ожидаемого результата. Так, в формировании запроса для генерации текста, в качестве уточняющих параметров, указываются желаемая длина, стиль, ключевые слова и другие возможные характеристики. Затем происходит генерация первоначального варианта запрашиваемого продукта. В зависимости от того, насколько полученный результат будет соответствовать изначальным ожиданиями, может потребоваться его дополнительная редакция. После коррекции или в случае неудовлетворения результата, производятся следующие итерации до достижения наиболее подходящего результата.

Таким образом, применение больших данных в журналистике направлено на решение маркетинговых и профессионально-творческих задач. В первом случае – это изучение аудитории и персонализация контента. Во втором – получение достоверной социальной информации и создание оригинального контента.

Выводы первой главы.

Цифровизация, развитие информационно-коммуникационных технологий и повсеместное применение технологий больших данных оказали значительные изменения в сфере медиа. Одним из значимых изменений, произошедших в журналистике, стало возникновение направления журналистики данных, основной деятельностью которого является сбор, обработка и визуализация данных.

Сущность данного направления раскрывается через взаимодействие с данными, однако внутри направления ученые предлагают журналистику данных и журналистику больших данных. Журналистики данных обычно оперирует небольшими объемами, известными как «малые данные». В противоположность этому, журналистика больших данных использует большие объемы данных, требующие применения специализированных инструментов и методов для обработки и анализа. Дифференциация журналистики данных в условиях повышения цифровой культуры специалистов массмедиа и распространения большого числа программных инструментов, основанных на искусственном интеллекте, нецелесообразна. В настоящее время можно говорить о стирании границ между журналистикой, журналистикой данных и журналистикой больших данных. Множество готовых приложений и программных продуктов позволяют журналистам любых направлений работать с большими данными. Корректное применение этих инструментов открывает новые перспективы для медиасферы и упрощает работу журналистам, однако ставит перед профессионалами задачу освоения новых профессиональных навыков.

Определение комплекса компетенций современного журналиста зависит от его роли в рабочем коллективе. Некоторые медиаорганизации имеют возможность создавать собственные отделы по работе с данными, в то время как другие полагаются на сторонних специалистов. В связи с этим, объективная оценка цифрового капитала современного журналиста в контексте его работы с большими данными представляет собой сложную задачу. В попытках конкретизации необходимого набора компетенций современного журналиста, мы пришли к выводу, что следует исходить из его основных операций работы с данными: поиск и извлечение, очистка, анализ, интерпретация данных и создание материала на их основе.

Многочисленные исследования свидетельствуют о том, что современный журналист должен обладать тремя ключевыми навыками: анализом данных, программированием и работой с нейросетями. Из этого списка следует, что анализ данных поможет журналисту не только в создании информативных материалов, но и в генерировании уникальных идей. Владение программированием позволит специалисту автоматизировать рабочие процессы с данными. Наконец, работа с нейросетевыми моделями даст возможность в кратчайшие сроки создавать уникальный контент. В редакции СМИ применение больших данных сводится к достижению двух основных целей. Первая заключается в оптимизации решения маркетинговых задач, которая подразумевает собой более точное изучение аудитории. Вторая направлена на реализацию профессионально-творческих задач, которая помогает сотрудникам редакций СМИ и журналистам добиваться эффективных результатов, экономя при этом человеческие и временные ресурсы, а также успешно конкурировать с другими СМИ благодаря созданию оперативного и оригинального контента.

Глава 2 Использование технологий работы с большими данными в редакции сетевого издания

2.1 Использование инструмента веб-аналитики для работы с большими данными информационного ресурса

В настоящее время для решения маркетинговых задач активно используются различные инструменты для сбора, анализа и интерпретации данных о поведении пользователей на сайтах. Такие инструменты предоставляют ценную информацию о том, какие страницы являются наиболее популярными, с каких типов устройств посетители заходят на сайты, из каких источников они приходят и многое другое. Одними из таких готовых решений являются инструменты веб-аналитики.

Существует большое множество таких инструментов от разных компаний и с разным функционалом. Среди наиболее популярных инструментов веб-аналитики выделяются «Яндекс Метрика» и Google Analytics. Кроме этих двух сервисов, на рынке представлены и другие инструменты веб-аналитики, к ним относятся Piwik, Kissmetrics, Adobe Analytics и многие другие. Каждое из этих решений обладает своими особенностями и преимуществами, но всех их объединяет то, что они направлены на отслеживание и работу с поступающим на сайты пользовательским трафиком. Выбор конкретного инструмента зависит от целей, потребностей и возможностей организации.

В роли заказчика исследования выступила редакция информационного портала «Площадь Свободы», которая предоставила доступ к изучению используемого ими инструмента веб-аналитики. Данная возможность позволила подробно изучить текущую ситуацию с использованием инструмента и, в соответствии с задачами исследования предложить способы

оптимизации использования веб-аналитики в маркетинговых целях для изучения целевой аудитории. Изначально сайт служил платформой, на которой публиковались материалы выпускаемой печатной версии газеты, при этом он не индексировался в новостных агрегаторах и имел низкую поисковую оптимизацию. После смены руководства редакции сайт был зарегистрирован как СМИ, что открыло дополнительные возможности для отслеживания результатов его работы. В июле 2023 года была произведена модернизация сайта, а в сентябре – получено свидетельство о регистрации СМИ. Таким образом, сайт превратился из платформы для дублирования газетных материалов в самостоятельное СМИ, в котором новости о событиях города и региона стали обновляться ежедневно.

«Яндекс Метрика» является бесплатным инструментом веб-аналитики, разработанным компанией «Яндекс», основная цель которого заключается в предоставлении наглядных отчетов, записей действий посетителей, отслеживании источников трафика и оценке эффективности онлайн- и офлайн-рекламы. Применение этого сервиса предоставляет возможность получить не только общую картину о пользователях на сайте, но и ряд других возможностей, в числе которых:

- анализ аудитории – процесс изучения характеристик посетителей сайта;
- анализ поведения – процесс изучения действий пользователей на сайте;
- анализ источников – процесс изучения каналов, через которые пользователи попадают на сайт;
- оптимизация рекламы – это процесс улучшения эффективности рекламных кампаний;
- электронная коммерция – процесс продажи товаров и услуг через Интернет;

- сквозная аналитика – это процесс объединения данных о посетителях сайта со всей цепочкой событий, которые произошли после их первого контакта с компанией.

Отслеживание всех этих процессов достигается за счет специального JavaScript-кода, который устанавливается на каждой странице сайта. Этот код позволяет собирать информацию о посетителях сайта для ее дальнейшей обработки на серверах «Яндекса». Таким образом, когда пользователь заходит на сайт, который использует код «Яндекс Метрики», его браузер автоматически загружает этот скрипт, который отправляет данные о визите пользователя на сервера «Яндекса». Например, эти данные могут включать: IP-адрес пользователя, дату и время его визита, URL-адрес страницы, которую он просматривает, информацию о браузере и операционной системе пользователя и т.д.

В добавок к этому, сервис использует файлы «cookies» (небольшие текстовые файлы, в которых браузер записывает данные с просматриваемых сайтов) для идентификации уникальных посетителей и отслеживании их действий на сайте. Такой подход направлен на получение сведений о поведении пользователей: сколько времени они проводили на сайте, какие страницы просматривали, с каким чаще контентом взаимодействовали и т.д.

После сбора данных об активности пользователей на сайте, «Яндекс Метрика» предоставляет отчеты в удобных формах: как в табличном виде, так и в виде различных готовых графиков и диаграмм. Это дает возможность владельцам сайтов или нанятым специалистам детально изучить активность пользователей на их ресурсе, а также понять, как они взаимодействуют с ним.

Вся работа с «Яндекс Метрикой» строится на отчетах, которые сервис автоматически формирует на основе собранных данных за выбранный промежуток времени. Каждый отчет содержит как графическую информацию

в виде графиков и диаграмм, так и текстовую статистику, представленную в виде таблицы. В зависимости от выбранного типа отчета можно получить данные о разных метриках (параметрах). Также, для углубленного изучения в отчетах «Яндекс Метрики» можно создавать сегменты аудитории для более точного анализа поведения определенных групп пользователей.

Поскольку «Яндекс Метрика» является многофункциональным сервисом веб-аналитики, позволяющим решать большой спектр задач, в рамках исследования мы фокусируемся преимущественно на работе с ее готовыми отчетами. В качестве примера подробно разберем отчет «Поисковые системы». На рисунке 1 приведен пример того, как выглядит отчет «Яндекс Метрики».

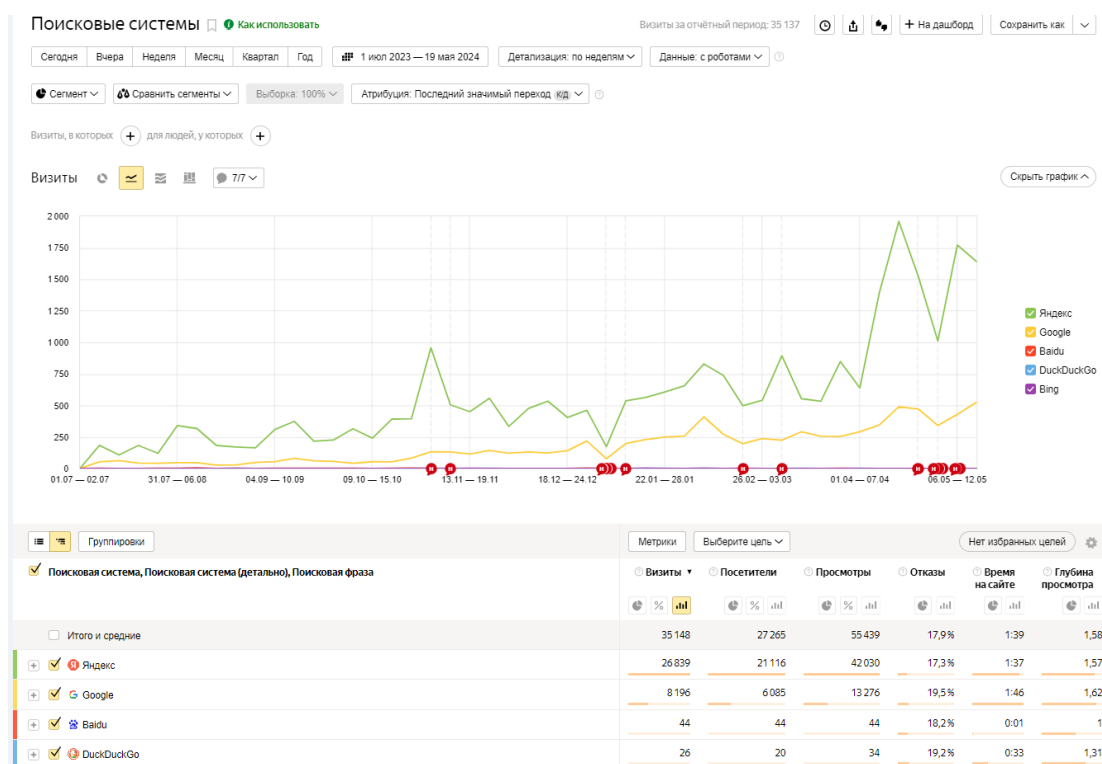


Рисунок 1 – Отчет «Поисковые системы»

Изначально, для отображения данных необходимо задать временной интервал, по которому система подгрузит соответствующие данные. В рассматриваемом примере, мы используем весь промежуток времени с

момента установки счетчика «Яндекс Метрики» на сайт до текущего момента. Ниже параметров отображения данных расположена визуальная составляющая отчета, которая может быть отображена в виде различных графиков и диаграмм в зависимости от цели изучения, удобства восприятия и т.д. В нижней части отчета находится таблица, с подробными данными по каждой записи в ней. Таблица, в свою очередь, разделена на две части, а именно это:

- группировки – характеризуют атрибут визита, по которому можно сгруппировать данные (отображены в таблице слева);
- метрики – являются числовыми величинами, которые рассчитываются на основе атрибута визита (отображены в таблице справа).

В рассматриваемом примере в качестве группировок выбраны те, которые непосредственно относятся к поисковым системам. Однако, при необходимости, они могут быть изменены на другие. Например, это может быть возраст посетителей или устройства, с которых они посещают вебсайт.

Обратим внимание на заголовки столбцов этой таблицы, которые относятся к метрикам. В этом примере отображены все базовые количественные и качественные метрики, которые являются основой практически каждого отчета «Яндекс Метрики». Подробнее рассмотрим каждый из них:

- визиты – последовательность действий (т.е. активность) одного посетителя на сайте, в счет которой идут: просмотры страниц, переходы по внешним ссылкам, загрузки файлов и т.д.;
- посетители – пользователи интернета, которые посещали сайт в течение определенного промежутка времени;
- просмотры – перезагрузка страницы сайта и загрузка при переходе посетителя на нее;

- отказы – количество визитов, продолжительность которых была меньше заданного для расчета отказов времени (по умолчанию это 15 секунд), а также во время которых было зафиксировано не больше одного просмотра страницы;
- время на сайте, которое принимается за разницу во времени между первым и последним событием в визите;
- глубина просмотра характеризуется количеством просмотров страниц сайта во время одного визита.

По аналогии с группировками, метрики также можно изменять, убирать и добавлять необходимые. Также, в таблице с метриками можно выбрать «Цель» (рисунок 2).

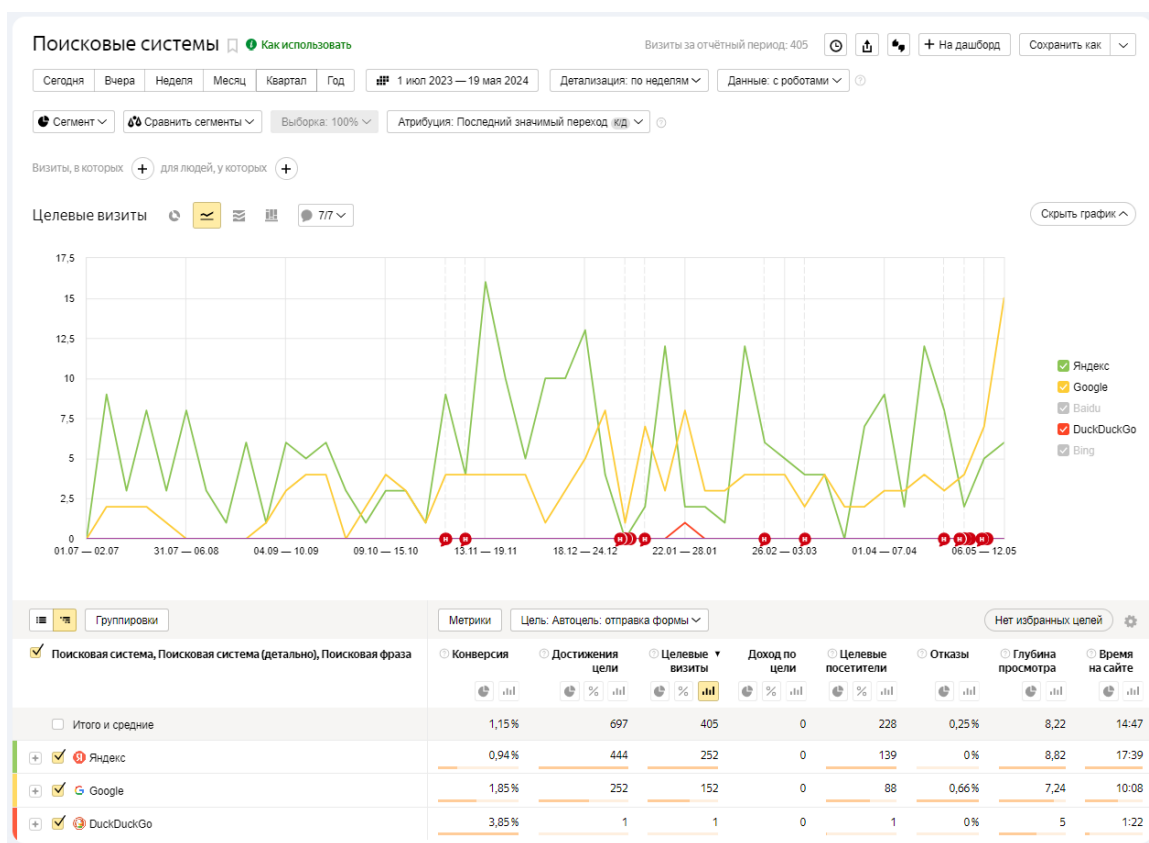


Рисунок 2 – Отчет «Поисковые системы». Цель – отправка формы

При выборе цели появляются новые метрики в таблице, которые помогают оценивать эффективность сайта и его конверсию, т.е. количество посетителей, совершивших целевое действие. В рассматриваемом примере выбрана цель, которая наряду еще с несколькими была автоматически сгенерирована в начале работы с «Яндекс Метрикой». Эта цель отображает данные отправки всех форм обратной связи на сайте. Таким образом с помощью ее в этом отчете мы можем узнать, посетители из какой поисковой системы чаще всего взаимодействуют с формами на сайте.

Помимо рассмотренных основных возможностей работы с отчетами «Яндекс Метрики», в них присутствуют более тонкие инструменты для изучения собранных данных. Одним из таких инструментов является сегментация, которая работает по принципу фильтрации отображаемых данных. Так, с ее помощью можно отследить активность какого-либо определённого сегмента пользователей.

Таким образом, с помощью всех рассмотренных инструментов в отчете можно получать необходимую информацию и на ее основе делать определенные выводы. Для рассматриваемого примера можно сделать некоторые выводы, например:

- за все время большее число посетителей приходят на сайт через поисковую систему «Яндекса»;
- большинство целевых визитов совершается посетителями, пришедшими на сайт через поисковую систему «Яндекса».

Следующим полезным инструментом «Яндекс Метрики», который позволяет объединять сразу несколько отчетов, является «Дашборды». Они представляют собой интерактивные панели, которые отображают ключевые показатели эффективности сайта в удобном и наглядном формате. С их помощью можно комплексно взглянуть и оперативно оценить общую картину

работы сайта, выявить проблемные места и определить направления для улучшения.

В начале работы с «Яндекс Метрикой» автоматически создается дашборд с названием «Сводка», который является примером того, как можно организовывать вместе несколько отчетов. Этот дашборд включает в себя 13 виджетов, каждый из которых отображает конкретные отчеты по данным за выбранный период времени в виде показателей, таблиц и графиков (рисунки 3-6). Далее подробно разберем представленные отчеты на каждом из рисунков, попутно совершая анализ имеющихся данных.

На рисунке 3, в качестве общих показателей, представлены три базовые метрики (просмотры, визиты и посетители) в виде виджетов показателей и посещаемость сайта в виде виджета столбчатого графика, на котором в пропорциях отображены базовые метрики за весь выбранный период.

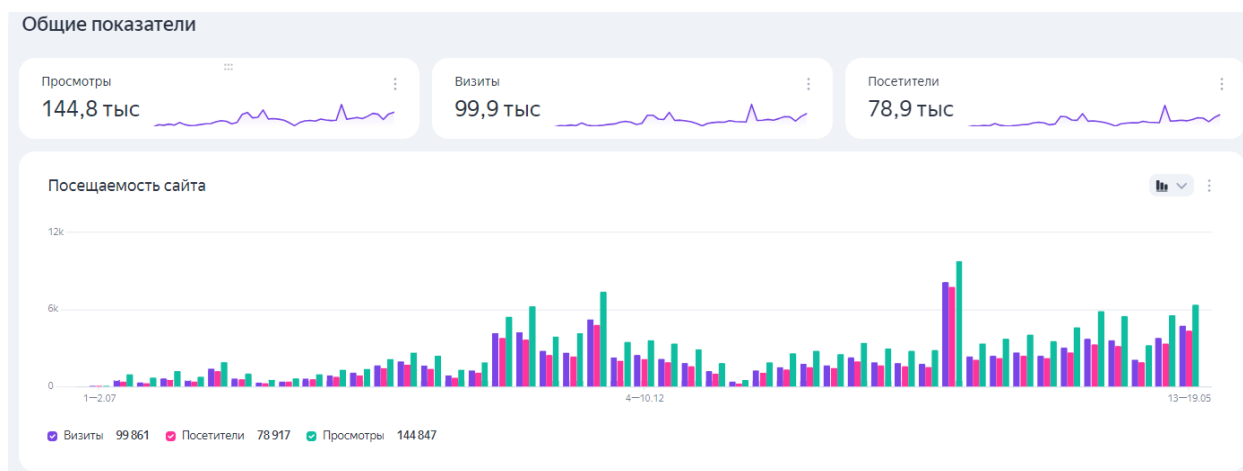


Рисунок 3 – Дашборд «Сводка». Общие показатели

Виджеты показателей сообщают общее число, в то время как виджет графика позволяет пронаблюдать динамику посещения сайта, а также дать оценку эффективности действий, которые предпринимались в определенный период времени. Также на этом графике можно заметить, что количество просмотров превалирует над количеством визитов и посетителей во всех

временных отрезках. Это считается нормальным поскольку визиты и посетители формируются на основе какой-то последовательности просмотров.

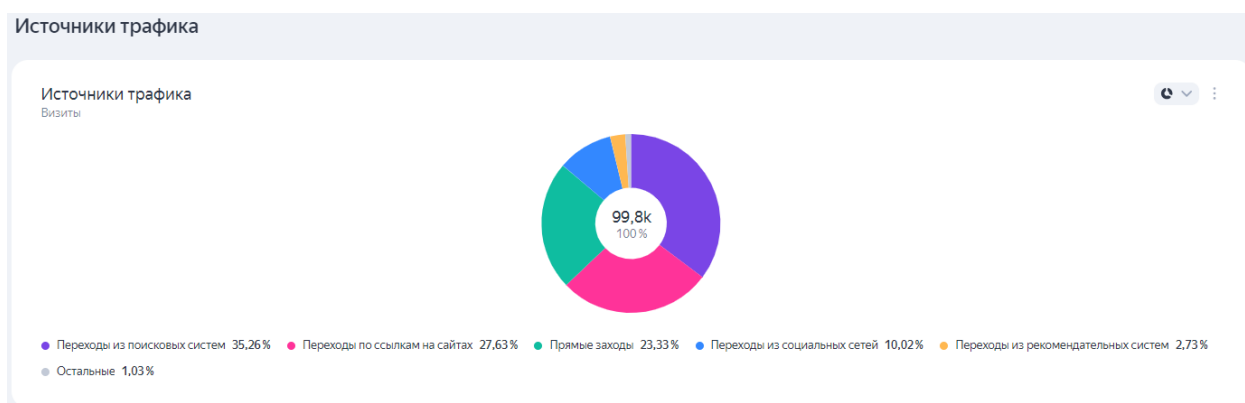


Рисунок 4 – Дашборд «Сводка». Источники трафика

Следующим в дашборде «Сводка», представленным на рисунке 4 является виджет диаграмма «Источники трафика», которая в процентном соотношении показывает, из каких источников пользователи попадают на сайт. Значение источников трафика помогает определить эффективность различных каналов привлечения пользователей и принять меры для улучшения работы тех, которые приносят меньше всего пользы. Круговой вид диаграммы позволяет рассмотреть, какие из источников трафика являются лидерами за выбранный ранее промежуток времени.

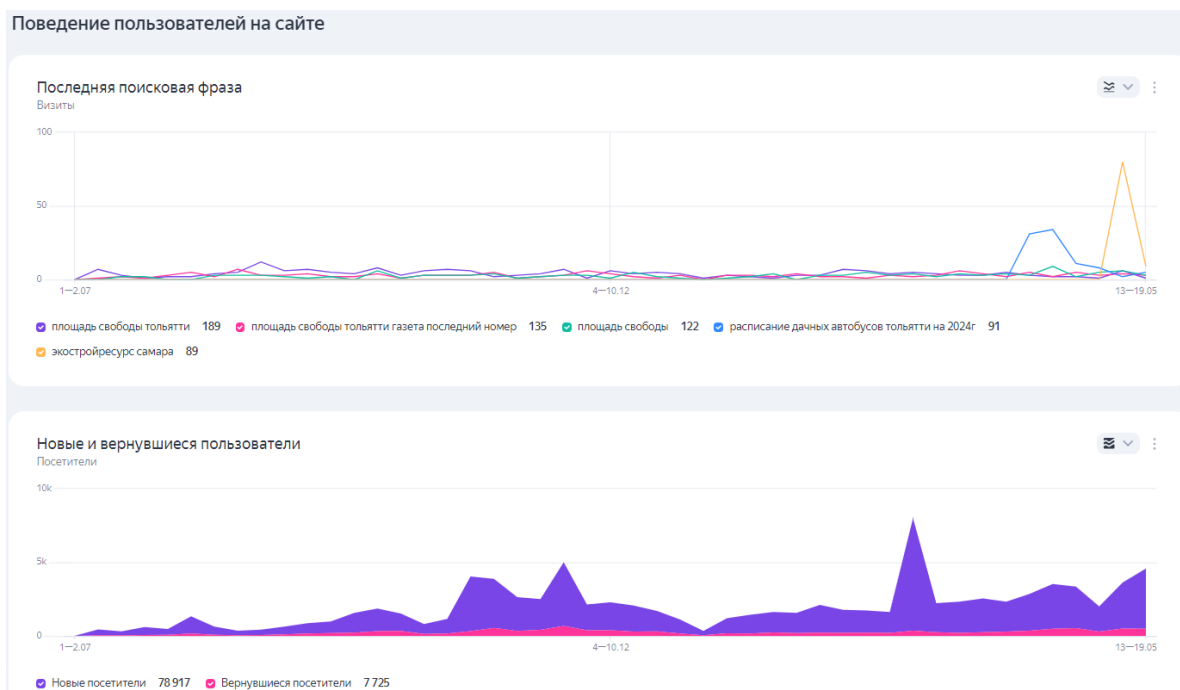


Рисунок 5 – Дашборд «Сводка». Поведение пользователей на сайте 1

На рисунке 5 изображена первая часть виджетов последнего блока под названием «Поведение пользователей на сайте» дашборда «Сводка». Виджет график «Последняя поисковая фраза» отображает, по какому поисковому запросу посетители приходят на сайт. С помощью данного отчета и его способа отображения можно отследить, какие темы являются наиболее актуальными и запрашиваемыми пользователями в разные промежутки времени.

Далее представлен виджет линейного графика с накоплением, который отображает количество новых и вернувшихся посетителей за все время использование инструмента. С помощью этого графика можно отследить уровень удовлетворенности пользователей, а также оценить качество создаваемого контента. Высокое число вернувшихся пользователей, может означать больший уровень удовлетворенности и сигнализировать о хорошем качестве контента поскольку эти показатели побуждают пользователей возвращаться в дальнейшем на ресурс. Рисунок 6 отображает последние виджеты, которые включены в дашборд «Сводка».

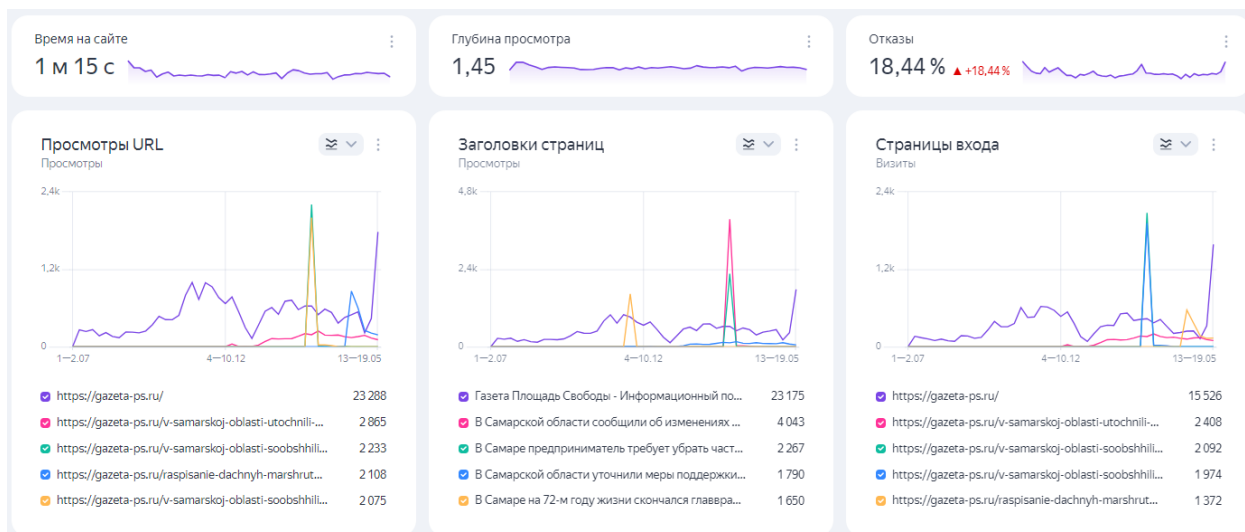


Рисунок 6 – Дашборд «Сводка». Поведение пользователей на сайте 2

Три сверху (время на сайте, глубина просмотра и отказы) – виджеты показателей, которые отображают среднее значение за весь выбранный промежуток времени. Также рядом можно наблюдать схематичный график изменения этого значения.

Большое количество проведенного времени на сайте может говорить о том, что его контент является полезным для пользователей. Аналогично с глубиной просмотра, чем выше показатель, тем интереснее и разнообразнее информация на сайте. Что касается отказов, то их высокий процент является признаком того, что сайт не предоставляет пользователю нужную ему информацию.

Последние три виджета графика являются похожими, то есть отображают схожие данные, но имеют разные характеристики. Так, например, виджет «Просмотры URL» отображает наиболее просматриваемые адреса страниц на сайте. Виджет «Заголовки страниц» практически схож с предыдущим, разница заключается только в том, что он характеризует наиболее просматриваемые заголовки страниц сайта. Последним является виджет «Страницы входа», который показывает какие страницы чаще всего

пользователи открывают первыми перед тем, как продолжить свое взаимодействие с ресурсом. Иными словами, данная метрика отвечает на вопрос «Какую страницу пользователь открывает первым делом?».

На этом обзор дашборда «Сводка» завершен. Стоит отметить, что данную страницу можно легко видоизменять под конкретные нужды. Так, можно добавлять новые виджеты или удалять ненужные, менять как их порядок, так и расположение, а также каждую из них можно отображать в виде: графика, таблицы и показателя.

Помимо представленных отчетов в виде виджетов в дашборде «Сводка» у инструмента «Яндекс Метрики» есть еще и другие отчеты, и их группы, которые невозможно поместить на панель отслеживания (т.е. дашборд), но с ними можно работать как по отдельности, так и в совокупности с информацией из других отчетов. Поэтому, перед тем как перейти к созданию нового дашборда и наполнению его доступными отчетами, сначала рассмотрим такие группы.

Первой группой является «Мониторинг», отчеты которого позволяют отслеживать работу сайта в реальном времени. Это позволяет своевременно предпринимать какие-либо действия при возникновении внезапных ошибок со стороны работы сайта. Также с помощью этой группы отчетов можно проводить тестирование при внесении изменений и отслеживать его результаты. Данный сегмент отчетов будет полезен разработчикам или их команде при обслуживании сайта.

Еще одной группой отчетов, которые невозможно отображать на дашбордах, является группа «Контент». Отчеты этой группы специально предназначены для информационных ресурсов, которые публикуют разного рода контент на своем сайте. На рисунке 7 представлена общая информация по отчетам группы «Контент».

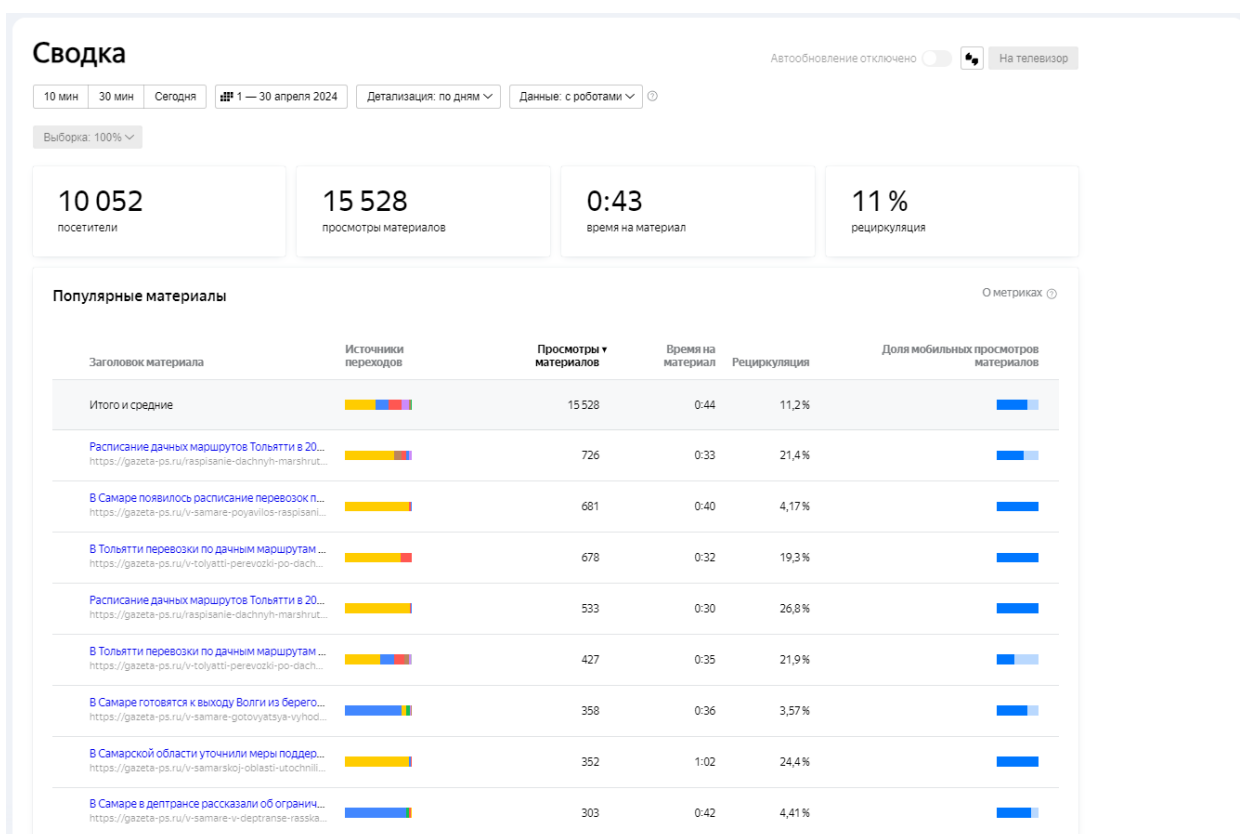


Рисунок 7 – Отчет «Сводка»

Были выбраны данные за апрель 2024 года. Здесь в удобном видео вначале представлены базовые метрики, по которым можно изучить то, какой публикуемый контент был наиболее интересен посетителям за выбранный период. В случае рисунка 7 можно сказать, что в апреле большое количество посетителей интересовалось вопросом расписания дачных маршрутов. Затем, детально можно просмотреть с каких источников они приходили, как долго в среднем они проводили времени за чтением материала, доля мобильных просмотров и показатель рециркуляции (т.е. то, какой процент посетителей продолжил смотреть другие новости на сайте).

Также у «Яндекс Метрики» есть инструменты, такие как вебвизор, карта скроллинга, карта кликов, карта ссылок и т.д. Все эти инструменты позволяют отследить активность посетителей на сайте вплоть до каждого действия, совершенного в течении одного визита пользователем. Такого рода

информация может быть как для каких-то технических целей, которые помогут совершенствовать навигацию сайта, так и для маркетинговых целей, которые подскажут, что чаще всего просматривают посетители.

Теперь для удобства работы с данными, создадим новый дашборд, в который будут добавлены необходимые отчеты для того, чтобы получить более обширную картинку работы сайта и активности посетителей на нем. Первым делом определим отчеты, которые могут охарактеризовать общую активность посетителей сайта. На рисунке 8 отображены те же виджеты отчетов с одинаковыми значениями, которые ранее были представленные в дашборде «Сводка», но в различных его местах.

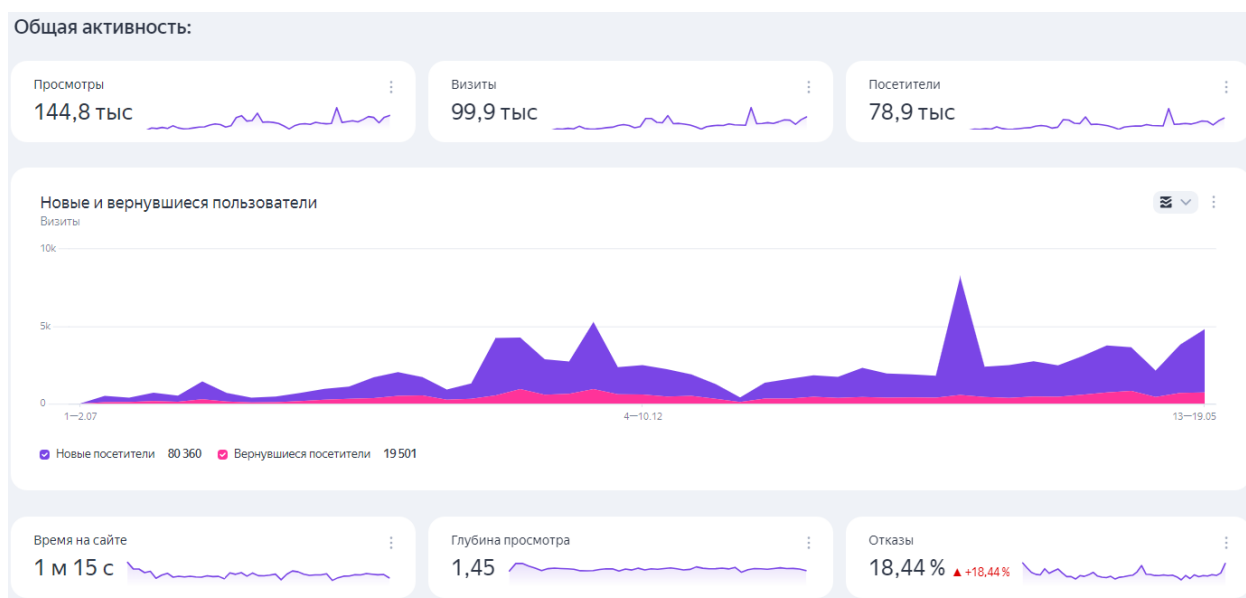


Рисунок 8 – Дашборд «Дата сводка». Общая активность

В новом дашборде «Дата сводка» мы посчитали целесообразным переместить их в одно место, поскольку, на наш взгляд, из всех других отчетов они отражают базовую активность посетителей сайта. На основании имеющихся данных, мы сделали следующие выводы:

- несмотря на умеренный общий рост количественных показателей посещаемости («Визиты», «Просмотры» и «Посетители»), число

качественных показателей («Отказы», «Глубина просмотра» и «Время просмотра») не демонстрируют каких-либо направленных движений; – не смотря на относительно малое количество вернувшихся пользователей к количеству новых, наблюдается пропорциональный рост числа первых во времена прироста новых посетителей.

Первый вывод может означать то, что аудитория приходит на сайт для того, чтобы посмотреть буквально одну какую-то конкретную новость целиком, либо бегло пролистать ленту последних новостей, которая располагается на главной странице сайта. Поскольку у большинства выкладываемых на сайте новостей время на чтение варьируется от одной до трех минут, значения показателей «Время на сайте» и «Глубина просмотра» являются вполне соответствующими.

Касательно второго вывода, график показывает большую разницу между потоками новых и вернувшихся пользователей. Однако, прежде чем составлять однозначные выводы, необходимо учесть несколько моментов. Первый заключается в том, что «Яндекс Метрика» определяет нового пользователя по его уникальному идентификатору, который помещается и хранится у пользователя в браузере (файлы cookie) на одном устройстве. Соответственно, если в реальности один и тот же пользователь зайдет посетит сайт с двух разных устройств, то «Яндекс Метрика» посчитает его не как один новый пользователь, а как уже как два новых пользователя. Еще одним фактором, на который стоит обратить внимание является то, что, поскольку данные о первом визите пользователя хранятся у него же в браузере в виде файлов cookie, пользователь может периодически очищать эти данные вручную либо с помощью какого-либо программного обеспечения. Таким образом, в следующий раз после чистки этих файлов, когда пользователь посетит сайт, «Яндекс Метрика» посчитает его за нового посетителя.

Следующим блоком, который мы сочли нужным отобразить, является группа отчетов, касающаяся характеристик аудитории сайта. На рисунке 9 изображен отчет «География», который позволяет узнать количество посетители из разных стран и их долевое распределение по округам, а также пронаблюдать динамику этих показателей и вовлеченность посетителей в разных регионах и городах.

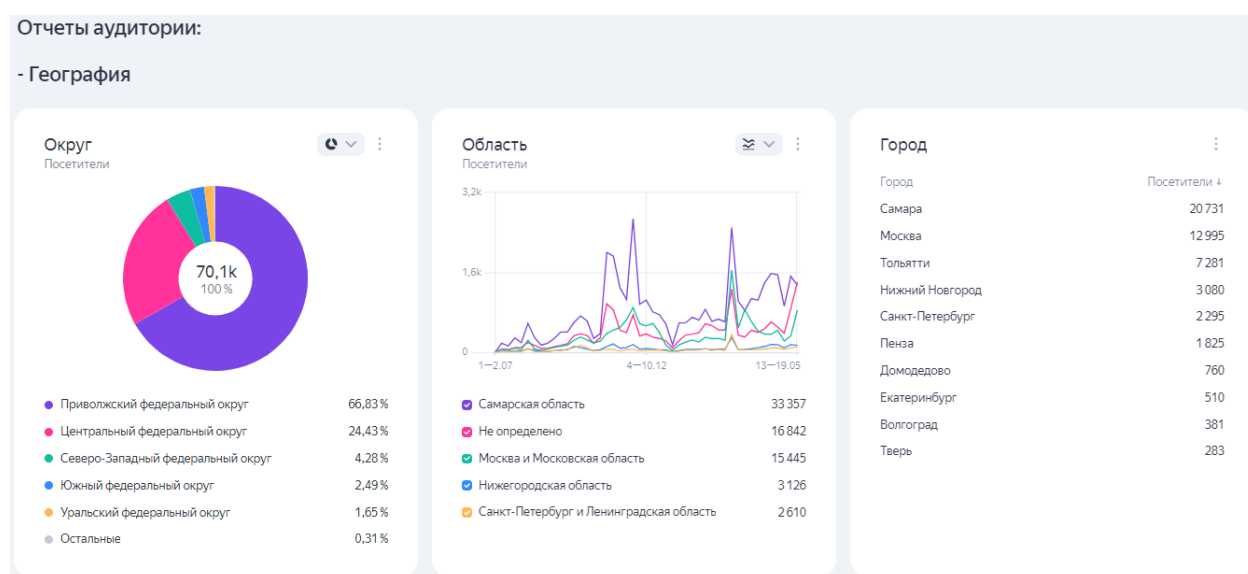


Рисунок 9 – Дашборд «Дата сводка». Отчеты по аудитории: география

Поскольку абсолютное большинство посетителей проживают на территории Российской Федерации, все выбранные виджеты отображают данные посетителей только этой страны. Так, согласно отображаемым данным:

- больше половины посетителей находятся в Приволжском федеральном округе;
- большее количество посетителей из Самарской области;
- город Самара является лидером среди других городов, Тольятти только на третьем месте после Москвы.

Рисунок 10 отображает пару отчетов, относящихся к демографическим показателям, характеризующим особенности аудитории.

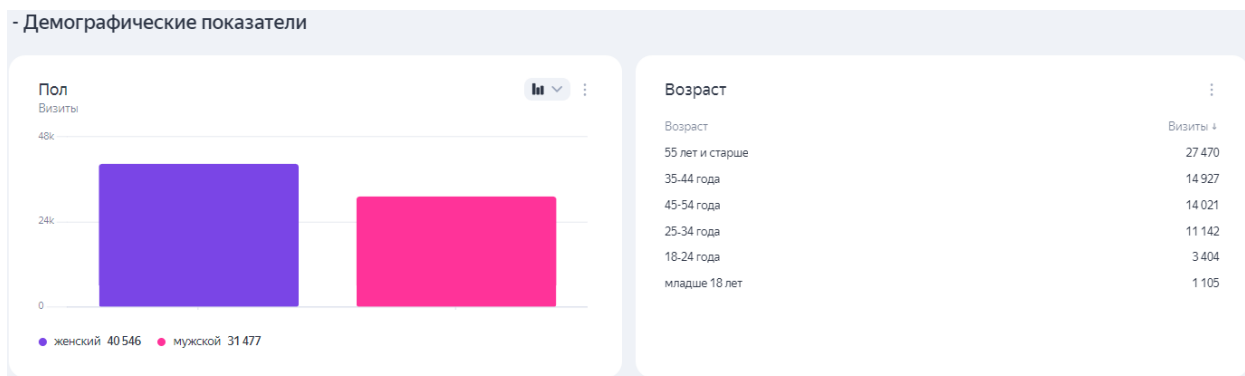


Рисунок 10 – Дашборд «Дата сводка». Отчеты по аудитории: демографические показатели

Так, слева расположен виджет столбчатого графика по отчету «Пол», в котором количество посетителей женского пола составляет большую часть. Справа отображен виджет таблица по отчету «Возраст», в которой основная часть аудитории является взрослым поколением, преимущественно группы «55 лет и старше». Последним отчетом, характеризующим интересы посетителей, является «Долгосрочные интересы» (рисунок 11).

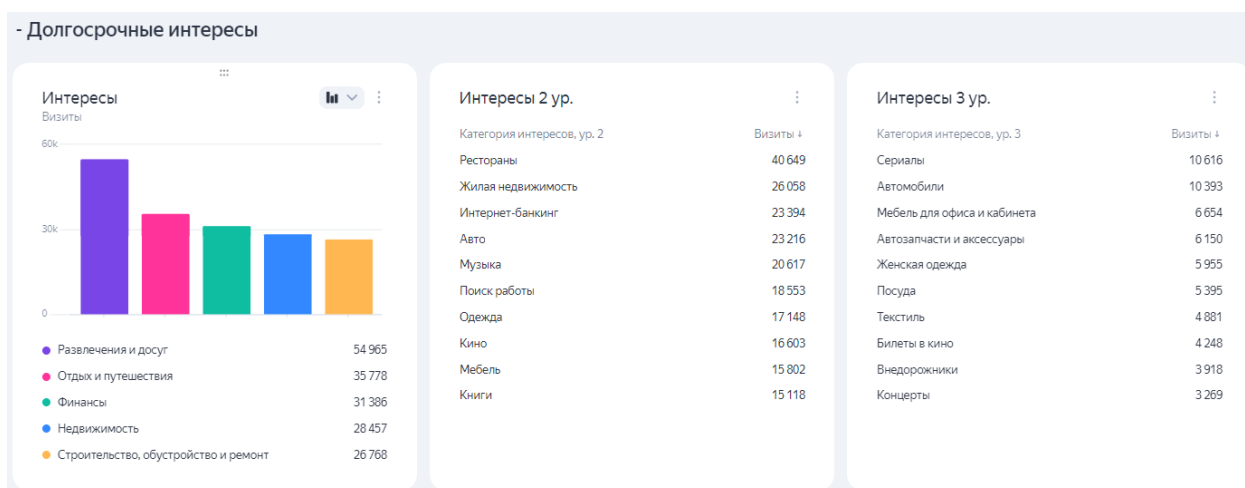


Рисунок 11 – Дашборд «Дата сводка». Отчеты по аудитории: долгосрочные интересы

Данный отчет показывает, насколько какая-либо тематика интересна посетителям сайта. Для определения интересов посетителей используется специальная технология «Крипта» [65], которая делает это путем анализа их

поведения в интернете. На рисунке 11 слева на право отображены уровни интересов посетителей. Виджет столбчатого графика слева представляет совокупные интересы аудитории, которые затем декомпозируются на более детальные в виджетах таблиц второго и третьего уровней.

На рисунке 12 изображены виджеты отчетов, которые отображают источники трафика (то, откуда посетители приходят на сайт).

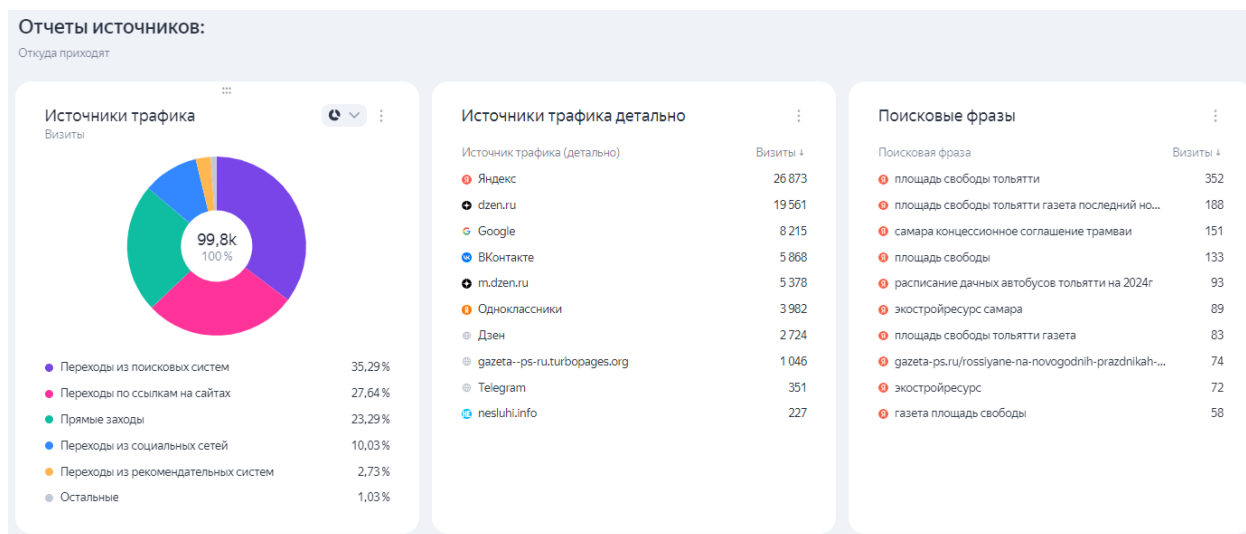


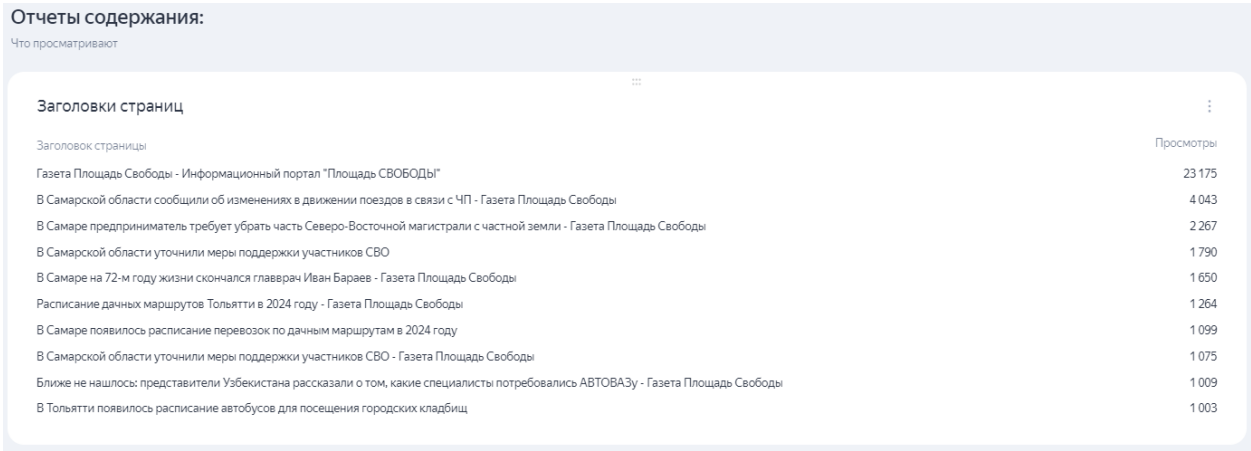
Рисунок 12 – Дашборд «Дата сводка». Отчеты источников

Слева расположен виджет круговой диаграммы для изучения общей картины того, из каких источников посетители попадают на сайт. Так, можно увидеть, что большое количество процентов приходится на «Переходы из поисковых систем» и «Переходы по ссылкам на сайтах», по 35,29% и 27,64% соответственно. Что еще здесь примечательно – это то, что доля «Прямых заходов» составляет практически четверть (23,29%). Прямые заходы обозначают то, что пользователи перешли на сайт напрямую, вбив URL-адрес в строку браузера или сохранив страницу в закладках. Посещения такого типа не связаны с поисковыми системами, социальными сетями или другими источниками трафика. Высокий показатель прямых заходов указывает на то,

что сайт имеет высокую узнаваемость, а также это говорит о том, что пользователи часто возвращаются на сайт.

Посередине находится детальная виджет таблица источников трафика, которая показывает, какие конкретно каналы приводят большее всего пользователей. На первом месте лидирует поисковая система «Яндекс», затем с практически одинаковым количеством идет блог-платформа «Дзен» (ее компьютерная и мобильная версии вместе). Также через нее пользователи попадают на сайт благодаря «Переходы из рекомендательных систем».

Последним виджетом таблицей справа является отчет о поисковых запросах в поисковике «Яндекса». Если открыть полностью данный отчет, то в нем отображено огромное количество уникальных запросов, но самым популярным из них является название информационного ресурса. На рисунке 13 отображена первая часть блока отчетов содержания.



Отчеты содержания:
Что просматривают

Заголовки страниц	Просмотры
Заголовок страницы	
Газета Площадь Свободы - Информационный портал "Площадь СВОБОДЫ"	23 175
В Самарской области сообщили об изменениях в движении поездов в связи с ЧП - Газета Площадь Свободы	4 043
В Самаре предприниматель требует убрать часть Северо-Восточной магистрали с частной земли - Газета Площадь Свободы	2 267
В Самарской области уточнили меры поддержки участников СВО	1 790
В Самаре на 72-м году жизни скончался главврач Иван Баравеев - Газета Площадь Свободы	1 650
Расписание дачных маршрутов Тольятти в 2024 году - Газета Площадь Свободы	1 264
В Самаре появилось расписание перевозок по дачным маршрутам в 2024 году	1 099
В Самарской области уточнили меры поддержки участников СВО - Газета Площадь Свободы	1 075
Ближе не нашлось: представители Узбекистана рассказали о том, какие специалисты потребовались АВТОВАЗу - Газета Площадь Свободы	1 009
В Тольятти появилось расписание автобусов для посещения городских кладбищ	1 003

Рисунок 13 – Дашборд «Дата сводка». Отчеты содержания 1

Виджет таблица заголовков страниц бы подробнее разобран ранее в дашборде «Сводка», поэтому здесь он сохранен без изменений. Однако, что здесь примечательно, что данный виджет подтверждает ранее сделанный вывод о том, что большинство посетителей просматривают главную страницу, где отображаются все последние новости.

На рисунке 14 в следующей части блока отчетов содержания рядом друг с другом находятся виджеты столбчатых графиков с накоплением.

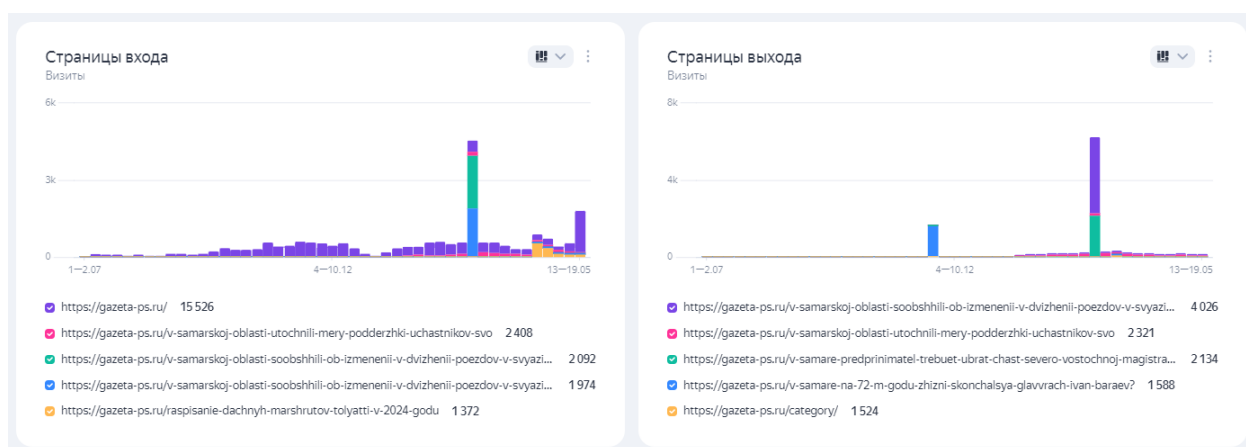


Рисунок 14 – Дашборд «Дата сводка». Отчеты содержания 2

Такое расположение рядом друг с другом позволяет удобнее проанализировать входные страницы посетителей на сайт и те, с которых они с него уходят. По страницам входа можно заметить, что большинство пользователей вначале посещают главную страницу сайта, в то время как в страницах выхода она отсутствует.

Последним блоком дашборда «Дата сводка» являются отчеты, касающиеся технологий, которые используют посетители для визитов сайта. Данная группа отчетов позволяет понять, на какие технологии необходимо ориентировать контент. В изображенном на рисунке 15 случае, видно, что большинство посетителей используют смартфоны для просмотра сайта.

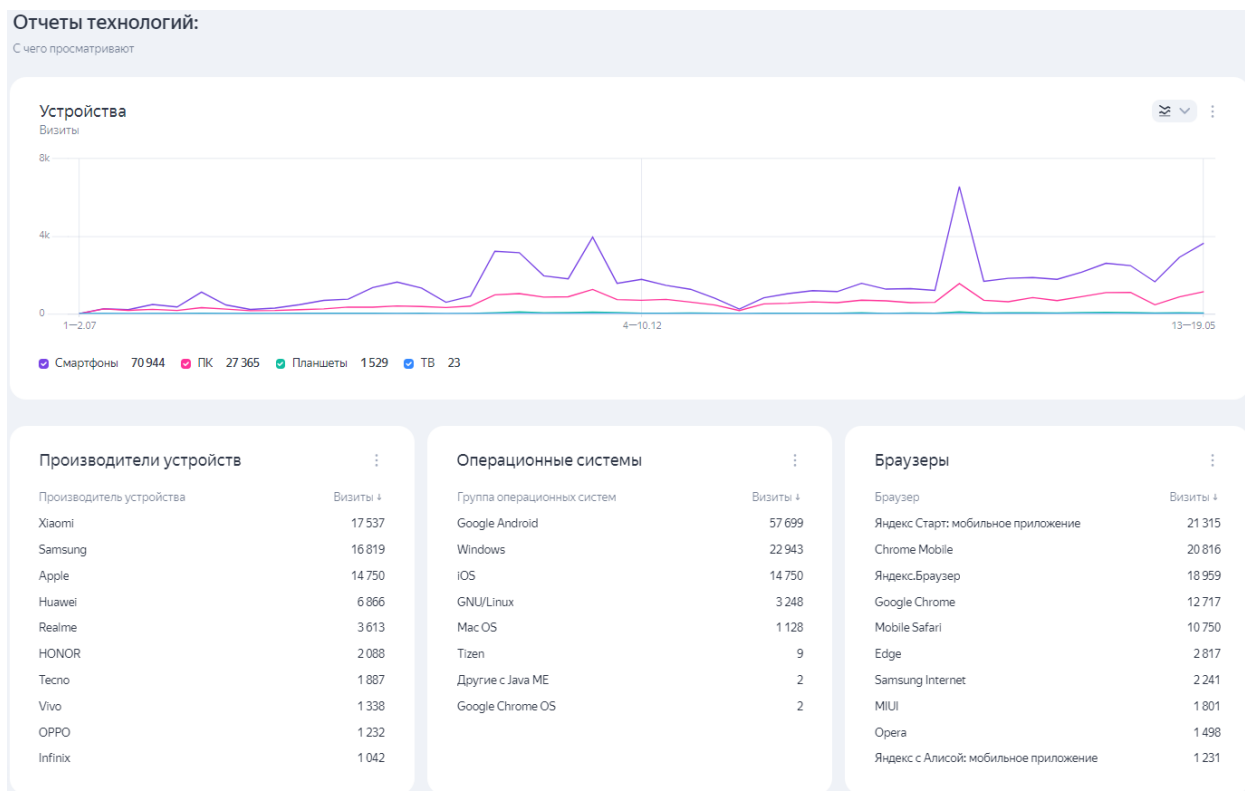


Рисунок 15 – Дашборд «Дата сводка». Отчеты технологий

На рисунке 15 отображены наиболее распространенные производители устройств, операционные системы и браузеры. Такая детальная информация может дать понять, на каких устройствах работа сайта осуществляется корректно, а на каких возникают неопределенные проблемы.

На этом работа с новым дашбордом «Дата сводка» завершена. В последствии его также можно будет модифицировать и видоизменять под любые необходимые цели и нужды. Он может быть взят в качестве основы для создания другого дашборда.

Таким образом, использование сервисов веб-аналитики медиаорганизациями, в частности, «Яндекс Метрики», подчеркивает важную роль этих инструментов в изучении аудитории и оптимизации работы сайта. Обзор автоматически сгенерированного дашборда подчеркивает способность сервиса с самого начала работы с ним собирать и визуализировать важные показатели. Эта первоначальная панель выступила в качестве

основополагающего примера, части которого впоследствии были позаимствованы в новом дашборде. Помимо этого, создание новой пользовательской панели позволило включить в нее другие доступные отчеты сервиса. Они позволили получить более глубокое представление: об общих сведениях активности посетителей на сайте (количество просмотров, визитов, новых и вернувшихся посетителей, времени на сайте, глубине просмотра и отказов) характеристиках аудитории (географию, демографические показатели, а также долгосрочные интересы), детальной информации об источниках, с которых пользователи попадают на сайт, и содержании того, какие действия они там совершают и что просматривают, используемых технологиях, при помощи которых посетители просматривают сайт. Выводы, сделанные на основе этих отчетов, послужили примером практического применения инструментов веб-аналитики для информационных ресурсов. Кроме того, один из них позволил определить дальнейшую направленность работы. Так, в ходе подробного изучения полученных результатов источников трафика и обсуждения их с заказчиком, было принято решение о разработке демонстрационной версии программы-парсера новостей с наиболее популярного источника «Дзен» для увеличения популярности информационного ресурса заказчика. Еще одной причиной принятия данного решения стало то, что платформа «Дзен» выступает не только в качестве источников перехода по ссылкам на сайтах, но также она выступает как рекомендательная система.

2.2 Разработка демонстрационной версии программы-парсера новостей

Разработка программы начинается с определения конкретных целей и задач. Согласно обговоренному с заказчиком техническому заданию,

программа должна собирать актуальные новости с указанных ресурсов (главные новости и новости, относящиеся к Самарской области). Стоит отметить, что программа не будет полностью копировать все содержимое страницы, на которой присутствует новость, а будет лишь заимствовать необходимые элементы. Так, это подразумевает сбор новостных заголовков, текстов, а также их первоисточников. На этапе сбора одновременно происходит очистка данных от ненужной разметки. Затем структурированные данные помещаются в локальную базу данных для их временного хранения. По истечению определенного времени собранные ранее данные постепенно зачищаются. Во время хранения записей в базе данных у пользователя есть доступ к получению актуальных на текущий момент новостей по обеим категориям. Также у пользователя есть возможность воспользоваться специальным интегрированным инструментом на базе искусственного интеллекта для создания уникального контента на основе выбранной из базы данных новости. Взаимодействие пользователя с программой-парсером будет происходить посредством чат-бота.

Первым делом необходимо обозначить набор инструментов, которые будут использованы для разработки программы, а также уточнить некоторые моменты касательно основных процессов работы программы.

В качестве основного языка программирования выбран Python, поскольку он обладает рядом преимуществ при написании программ парсеров. Это достигается не только за счет богатого набора готовых библиотек, но также за счет его высокой производительности. Несмотря на то, что Python не является самым быстрым из языков программирования, его производительности достаточно для решения поставленной задачи в рамках исследования. Так, предполагается, что основной алгоритм сбора данных будет с некоторой периодичностью посещать определенные страницы сайта, на которых размещаются отслеживаемые новости.

За счет простоты использования, а также совместимости со многими языками программирования, для хранения собираемых данных будет использована база данных SQLite. Еще одним преимуществом данного выбора является его масштабируемость. В случае если возникнет необходимость собирать больше данных с большего количества источников, можно будет легко перейти на другую систему управления базами данных (СУБД).

Для организации взаимодействия системы и пользователя был выбран чат-бот социальной сети «ВКонтакте». С помощью данного чат-бота у пользователя будет возможность получать информацию о последних актуальных новостях в двух режимах. Первый режим характеризуется тем, что чат-бот сам будет оповещать пользователя, когда основной алгоритм по сбору данных будет обнаруживать новые новости на сайте. У пользователя будет возможность приостановки и возобновления исполнения данной функции по отношению к себе с помощью специальных соответствующих команд. Вторым режимом предполагается отправку чат-ботом новостей исключительно по конкретному запросу пользователя. В этом случае чат-бот не осуществляет никакой рассылки новостей самостоятельно. Также, для создания уникального контента на основе собранных данных в чат-бот будет интегрирована нейросетевая модель GigaChat от «Сбера».

Перед тем, как перейти к написанию программы необходимо обратиться на сайт-источник и детально изучить его содержание.

На рисунке 16 изображена главная страница ресурса, с которой предполагается собирать данные.

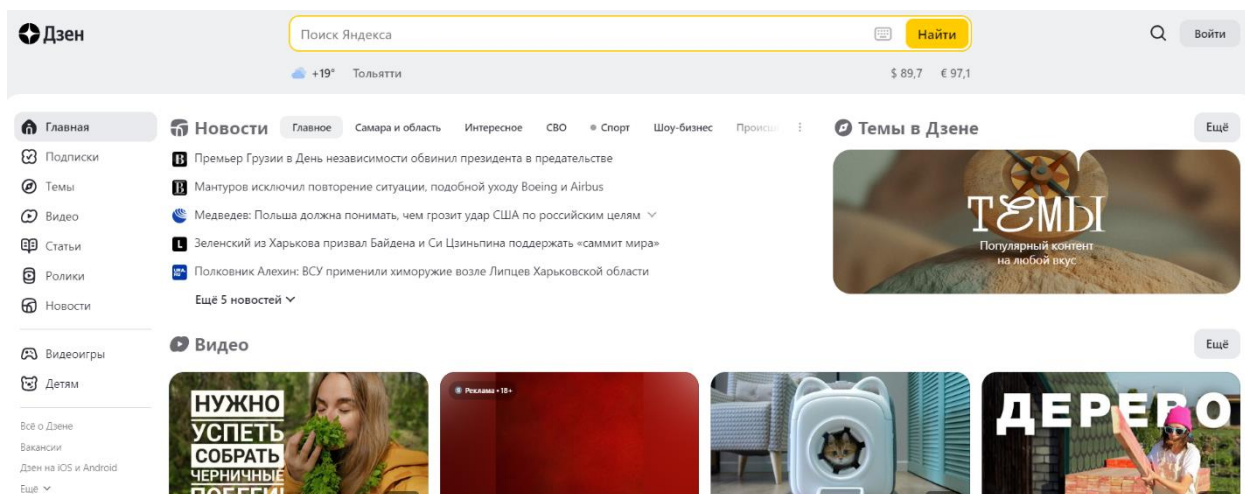


Рисунок 16 – Главная страница «Дзен»

Можно заметить, что изначально открывшаяся вкладка не полностью отображает интересующий список главных новостей, а лишь его небольшую часть. Для более подробного раскрытия этого списка необходимо выполнить два одинаковых действия – совершить клик на элемент «Еще 5 новостей», после чего появится их максимально расширенный список. Аналогичную процедуру необходимо проделать для получения такого же списка новостей по Самарской области, только перед этим потребуется переключиться с раздела главных новостей.


В целях избежания совершения этих дополнительных действий, из-за которых в последствии могут возникнуть некоторые неудобства, целесообразно продолжить изучение содержимого сервиса на предмет наиболее подходящих путей к сбору данных. Так, подробнее изучив структуру сайта, было найдено две соответствующих его страницы.

На рисунке 17 изображен блок страницы, в котором находятся все последние главные новости.

Главная
Подписки
Темы
Видео
Статьи
Ролики
Новости
Видеоигры
Детям

Всё о Дзене
Вакансии
Дзен на iOS и Android
Ещё ▾


Главное



URA.Ru Сегодня

Грузия собирается вернуть в свой состав Абхазию и Южную Осетию

Грузия планирует вернуть в состав своего государства Абхазию и Южную Осетию, а также стать членом Евросоюза к 2030 году.



Коммерсантъ 2 часа назад

Spiegel: страны Балтии введут на Украину войска в случае прорыва ВС России

Как пишет Der Spiegel, в таком случае страны Балтии и Польша введут свои войска на Украину.

\$89,70

- В Госдуме предложили установить минимальный уровень доходов для повышения НДФЛ
- Президент России Владимир Путин прибыл с государственным визитом в Узбекистан
- Полковник Алёхин: ВСУ применили химоружие рядом с Липцами Харьковской области
- У берегов Сочи в Черном море заметили дрон-разведчик США
- Мантуров: Россия не допустит повторения ситуации, подобной уходу Boeing и Airbus
- Два украинских снаряда сбиты в Белгородской области
- В июне 80-летние пенсионеры получат повышенную в два раза пенсию
- Задержан советник и близкий друг губернатора Орловской области
- Лавров назвал разговоры об ударах по России оружием США агонией
- Захарова заявила, что Россия опереживает народу Армении из-за наводнения
- Минобороны сообщило о взятии села Берестовое в Харьковской области
- Зеленский попросил Байдена и Си Цзиньпина приехать на саммит по Украине
- Макрон заявил, что отреставрированный после пожара Нотр-Дам откроется 8 декабря
- Оппозиция Грузии намерена выдвинуть в премьеры антироссийского радикала
- В подконтрольном ВСУ городе Запорожье произошел взрыв на фоне воздушной тревоги
- Эвакуированные из Белгорода подростки разгромили башкирский лагерь «Орленок»
- Медведев: Польше нужно понимать, что удар США по российским целям грозит войной
- Подполье: ВС России нанесли удар по аэродрому Староконстантинов

Рисунок 17 – Страница главных новостей «Дзен»

Данный блок, в свою очередь, разделен на две части, которые содержат в себе иконки источников и заголовки новостей. По наличию иллюстраций и кратких описаний у новостей в левой части, можно предположить, что они являются наиболее актуальными. Несмотря на разделенное расположение новостей в блоке, ссылки на их страницы находятся под одинаковым тегом в разметке страницы. Таким образом, это в дальнейшем при написании алгоритма сбора позволит получать все новости за совершение одного действия.

На этой же странице располагается ряд аналогичных блоков по разным категориям. Так, следующим по порядку идет блок о местных новостях. Аналогично блоку главных новостей, можно было повторить те же самые действия для сбора новостей в блоке Самарской области. Однако, забегаю немного вперед, необходимо учесть один момент. Когда пользователь совершает переход по какой-либо ссылке у себя в браузере, он отправляет

запрос на сервер, который впоследствии его обрабатывает и выдает ему то, что пользователь видит у себя на странице. В этом запросе пользователь также передает некоторые параметры, среди которых присутствует его IP-адрес, по которому сервер способен определить приблизительное местоположение пользователя. Сейчас, поскольку мы отправляем запросы из Самарской области, сервер возвращает нам местные новости, но в случае дальнейшей загрузки программы на удаленный сервер для автономной работы, за место них могут быть отражены новости той области, в которой будет расположен дата-центр. Следственно, такой вариант не подходит, необходимо продолжить изучение сайта.

Рисунок 18 отображает страницу новостей, посвященную исключительно новостям из Самарской области.

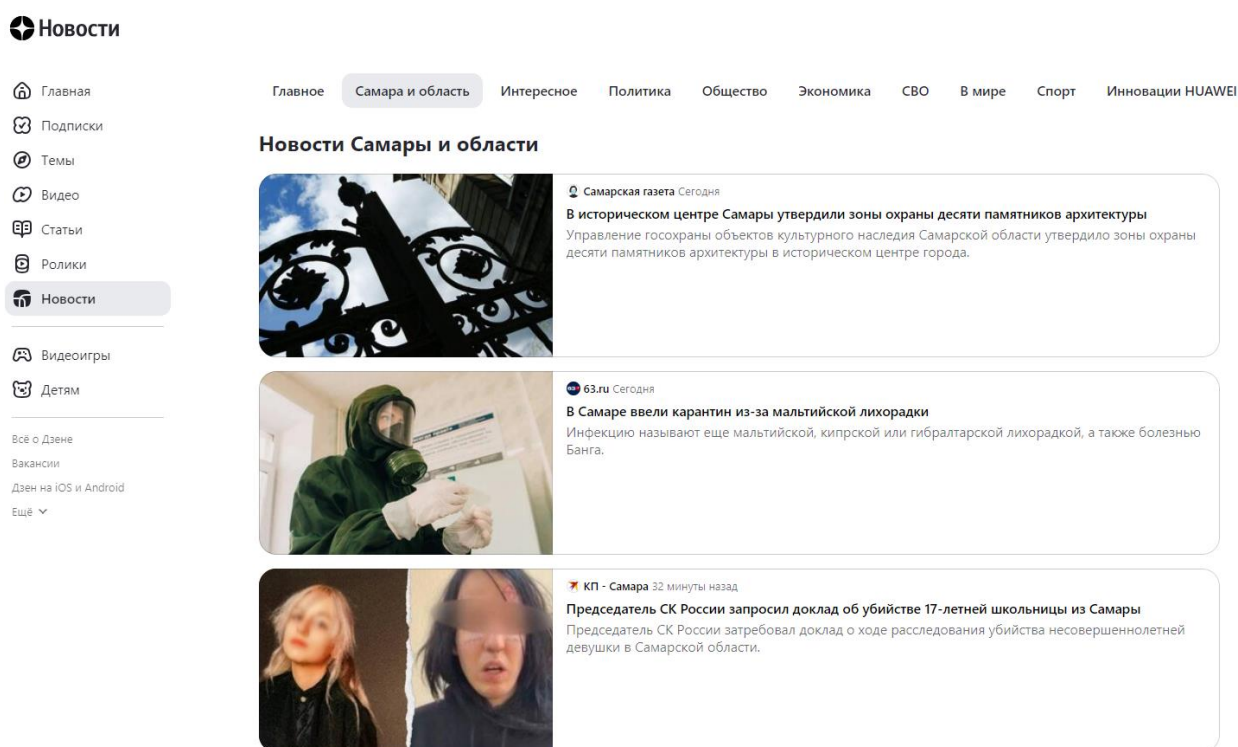


Рисунок 18 – Страница новостей Самарской области

В отличие от предыдущей страницы, здесь все новости расположены в более развернутом виде. Аналогично, ссылки на материалы находятся под одним и тем же тэгом в разметке страницы.

Далее перейдем к изучению структуры страницы единичной новости. Это необходимо для того, чтобы понять какие элементы нужно будет собирать. На рисунке 19 отображена страница, на которой расположена необходимая для сбора информация.



Рисунок 19 – Страница конкретного новостного сообщения

Примечательно здесь то, что новости «Дзен» формируются по большей части на основе четырех ключевых фраз, взятых из разных новостных источников. Немаловажным здесь является то, что после фраз указываются не только названия источников, а ссылка на полную версию материала.

Далее перейдем к проектированию базы данных, в которой будут помещаться собранные новости на временное хранение. Создадим таблицу для новостей в базе данных согласно следующей структуре столбцов:

- id – является уникальным идентификатором каждой записи;

- category – обозначает категорию новостей (главное либо новости Самарской области);
- url – ссылка на собранную новость в «Дзен»;
- title – заголовок собранной новости;
- text – текст собранной новости;
- sources – источники, которые упоминаются в тексте собранной новости;
- relevance – данный столбец необходим для обозначения актуальности новости на текущий момент;
- parsed_datetime – дата и время, в которые была собрана новость.

Разработка вышеописанной части программы. Для написания кода мы используем среду разработки PyCharm. Начнем с создания нового проекта и установки следующих библиотек:

- aioredis позволяет асинхронно выполнять операции с базой данных SQLite, что способствует улучшению производительности и уменьшению использования ресурсов;
- sqlalchemy является удобной библиотекой для работы с базами данных, которая позволяет сосредоточиться на логике приложения, а не на деталях реализации SQL-запросов;
- requests – это популярная библиотека для работы с HTTP запросами, которая предоставляет простой интерфейс для сетевых запросов;
- BeautifulSoup4 и lxml необходимы для разбора и обработки XML и HTML документов, а также для поиска, разбора и изменения элементов структуры веб-страницы;
- selenium представляет собой инструмент автоматизации браузера, который позволяет автоматизировать операции, выполняемые различными браузерами;

- `gigachain` и `gigachat` являются библиотеками, которые позволяют упростить и автоматизировать работу с нейросетевыми моделями GigaChat.

Реализация функции сбора и очистки данных с ресурса включает осмысление основных процессов работы с базой данных. Поскольку алгоритм по сбору данных будет имитировать действия пользователя посредством автоматизированного управления браузером, как это было подробно описано ранее, в начале кода необходимо задать некоторые параметры для его оптимальной работы. В их число входят настройки браузера, такие как `User-Agent`, `headless` опция и другие, а также адреса целевых страниц, с которых будет производиться сбор данных.

На рисунке 20 схематично отображена логика работы алгоритма по сбору данных, которая в точности повторяет описанные ранее действия по выборке новостей. Единственным отличием здесь является то, что на схеме присутствуют компоненты, которые отвечают за отправку пользователям сообщений о последних новостях. Для этого алгоритм обращается к базе данных с запросом на получение списка пользователей, которых необходимо оповестить об изменении актуальных новостей. Данное действие происходит в конце цикла, перед тем как алгоритм перейдет в режим ожидания для того, чтобы повторить процесс снова через заданное время.

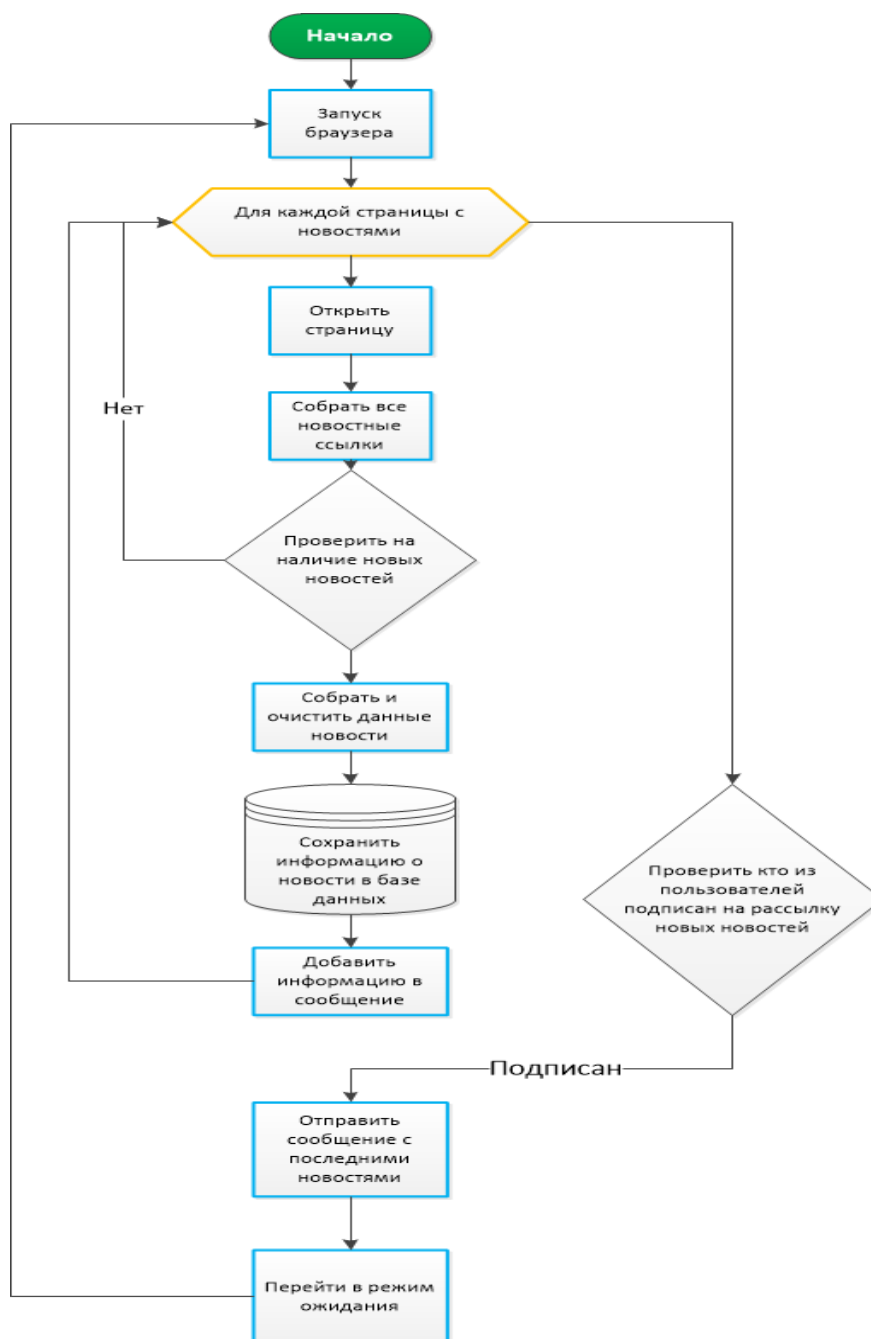


Рисунок 20 – Схема работы алгоритма сбора новостей

Следующий этап – разработка алгоритма взаимодействия системы с пользователем посредством чат-бота – представлен в схеме на рисунке 21.

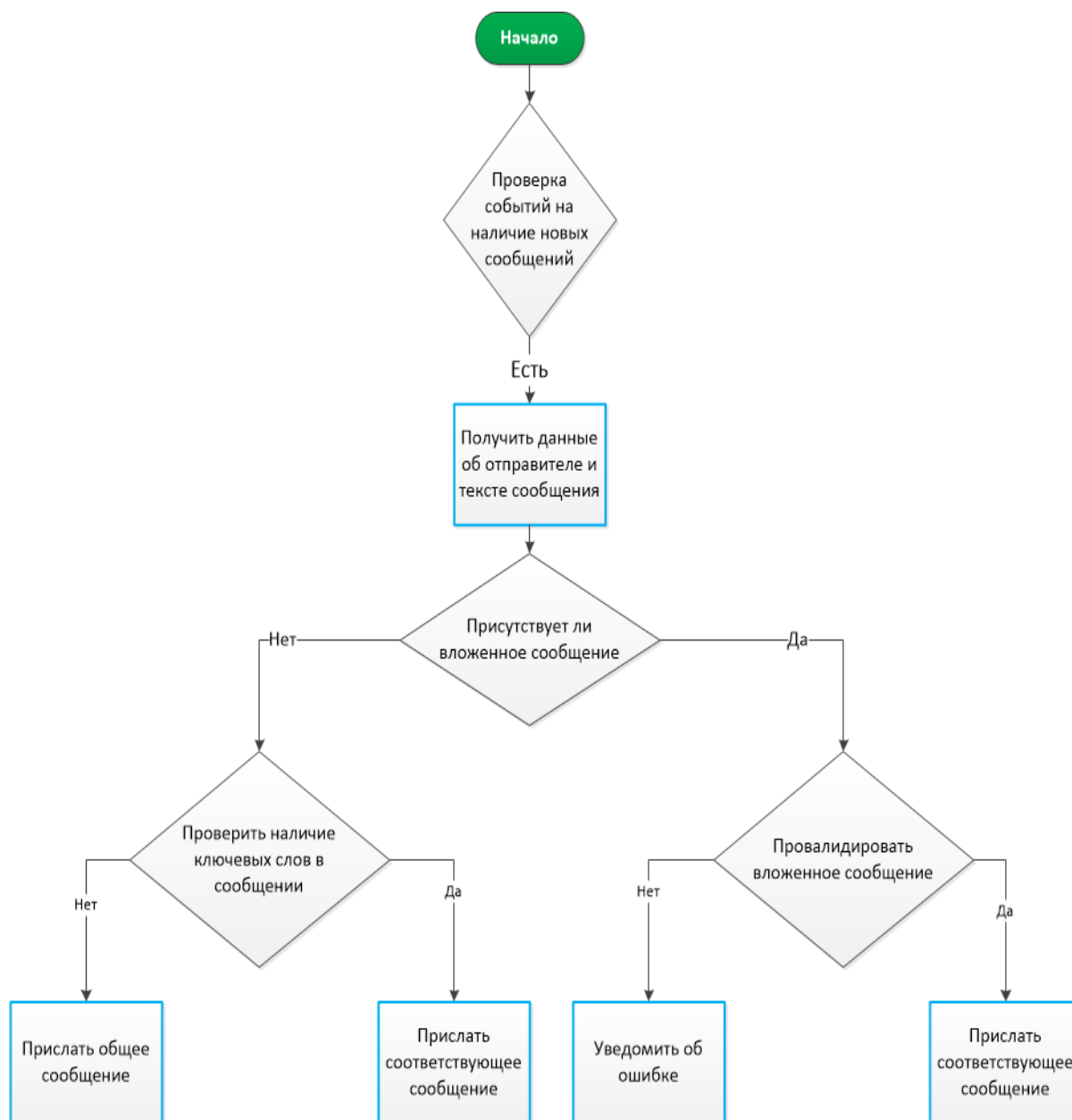


Рисунок 21 – Схема работы алгоритма чат-бота

Основной принцип, по которому работает чат-бот, определяется наличием вложенных сообщений. Вне зависимости от наличия или отсутствия вложенных сообщений, алгоритм производит проверку сообщения в обоих случаях. Так, в случае отсутствия вложенного сообщения, проверка происходит за счет совпадения с ключевыми фразами, которые пользователь отправляет через клавиатуру беседы. В случае присутствия вложенного сообщения, алгоритм в первую очередь проверяет совпадение ключевой

фразы, с которой оно было отправлено, а затем переходит к валидации вложенного сообщения.

Для того, чтобы наглядно продемонстрировать взаимодействия пользователя с системой, далее будет детально разобрана каждая доступная операция этого процесса.

На рисунке 22 изображена главная клавиатура беседы, которая содержит кнопки «Запустить» \ «Остановить».

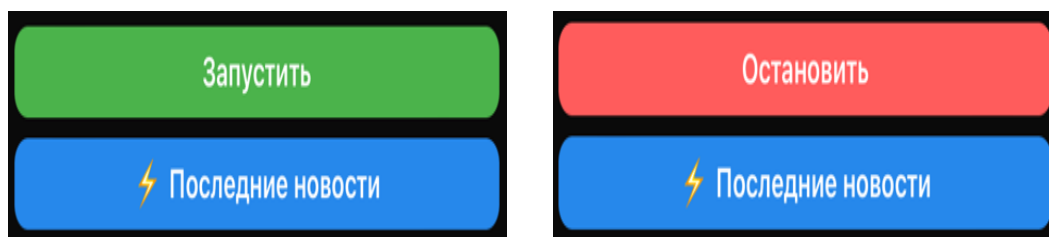


Рисунок 22 – Кнопки «Запустить» \ «Остановить»

С их помощью пользователь способен переключаться между режимами, в которых чат-бот оповещает о последних новостях в режиме реального времени. Когда отображается зеленая кнопка, то чат бот находится в режиме пассивного мониторинга новостей, а когда отображается красная кнопка – происходит своевременная отправка новостей после процесса их сбора и обработки. Пример такого взаимодействия изображен ниже на рисунке 23.

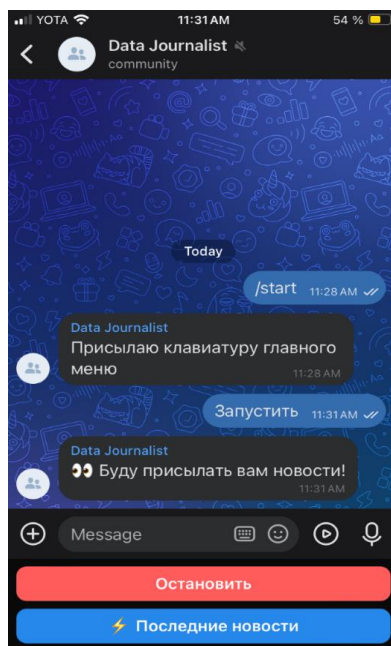


Рисунок 23 – Пример использования кнопки «Запустить»

Следующей фиксированной кнопкой главной клавиатуры является «Последние новости», которая инициирует появление дополнительной клавиатуры, предоставляющей возможность выбора категории для получения последних актуальных новостей (рисунок 24).

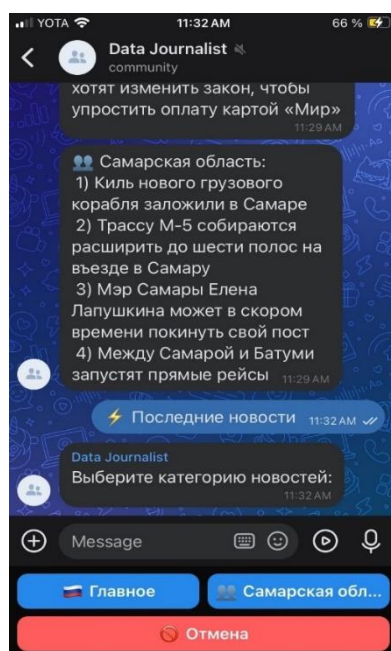


Рисунок 24 – Пример использования кнопки «Последние новости»

В присланной клавиатуре присутствуют кнопки «Главное» / «Самарская область», нажатие на которых пришлет актуальные новости выбранной категории на текущий момент (рисунок 25).

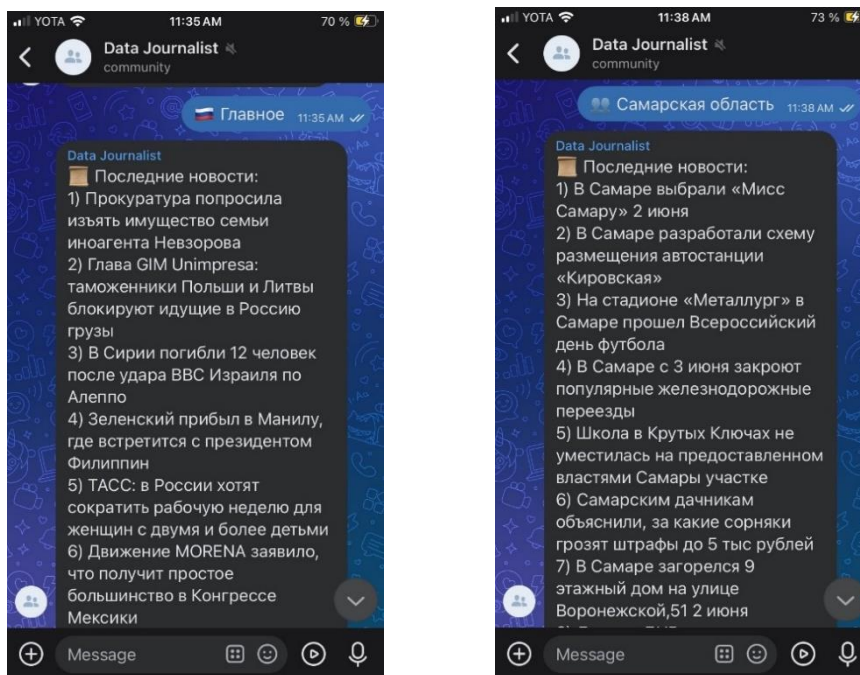


Рисунок 25 – Пример использования кнопок категорий

Стоит также отметить, что во время выбора категории для получения полного списка актуальных новостей у пользователя временно отключается возможность получать последние новости, т.е. чат-бот перестает присылать новости своевременно. Это необходимо для того, чтобы пользователь мог в этот короткий промежуток времени сделать выбор без отвлечения внимания, последние новости какой категории ему необходимо получить. Для выхода из этого состояния, а также в случае случайного нажатия кнопки «Последние новости», можно использовать кнопку «Отмена», которая возвращает пользователя к главной клавиатуре (рисунок 26).

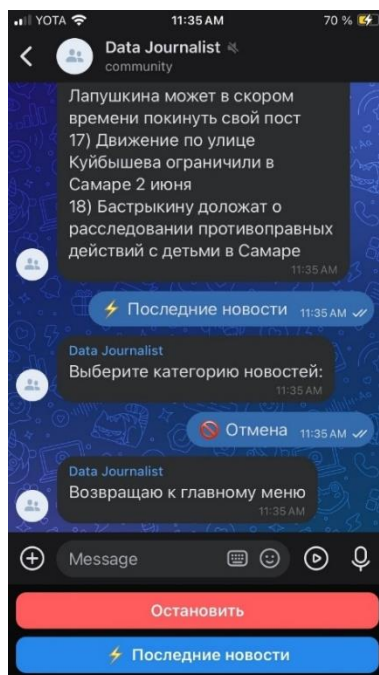


Рисунок 26 – Пример использования кнопки «Отмена»

Далее рассмотрим процесс взаимодействия с чат-ботом посредством вложенных сообщений. Так, для того чтобы получить подробную информацию о интересующей новости, необходимо ответить на соответствующее шаблону сообщение одной из входящих в него цифр (в нем должен присутствовать нумерованный список новостей от 1 до 18).

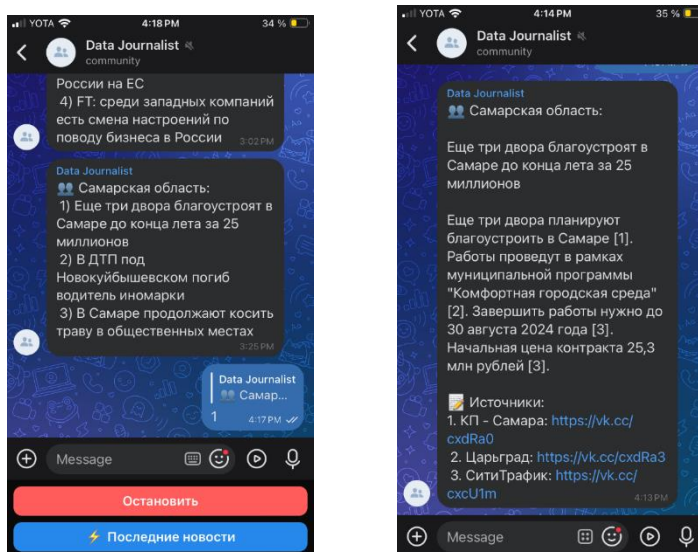


Рисунок 27 – Пример использования кнопок категорий

Для того, чтобы совершить рерайт необходимой новости на месте, достаточно ответить на соответствующее шаблону сообщение (в нем должны присутствовать: категория, заголовок, текст и список источников) сообщением «.» или «..». Пример изображен ниже на рисунке 28.

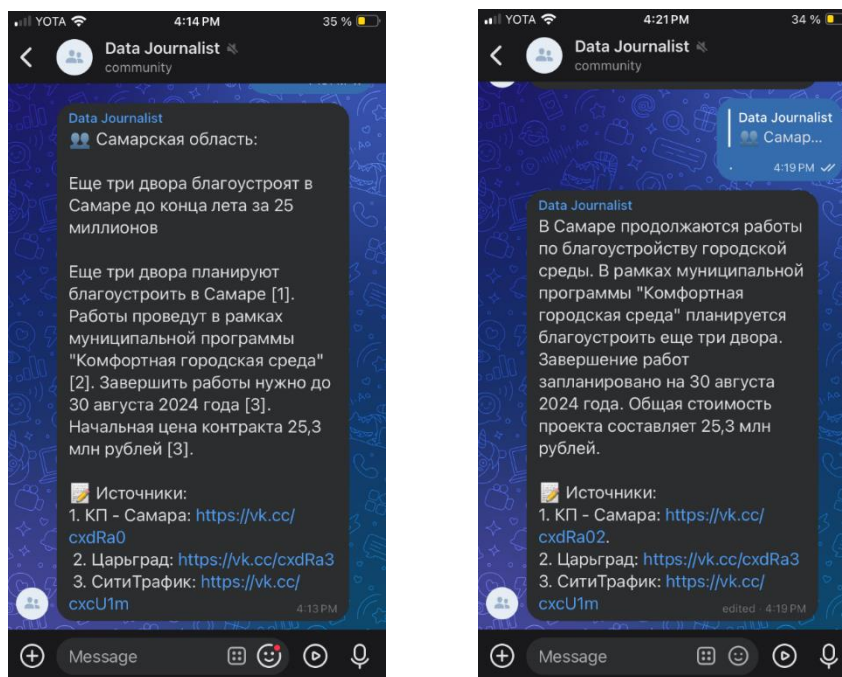


Рисунок 28 – Пример рерайта новости

Сообщение «.» (одиночная точка) задействует обыкновенную языковую модель GigaChat, а сообщение «..» использует ее улучшенную Pro-версию.

Дальнейшие действия в процесс разработки демонстрационной версии программы-парсера новостей направлены на локальное тестирование системы с последующим ее размещением на подходящем хостинг-сервисе, а также ее настройку и оптимизацию для эффективной работы на нем.

Таким образом, разработанная в рамках магистерской диссертации программа-парсер для сбора новостей позволила наглядно продемонстрировать потенциал применения языков программирования для организации автоматизированного цикла работы с данными, представленными

в новостных сообщениях, находящихся на высоких позициях рейтингов. Также, интегрирование нейросети в программу-парсер способствует оптимизации работы редакции сетевого издания с новостным контентом, позволяя затрачивать меньшее количество действия на создание контента.

Выводы второй главы.

Основанная на применении научного подхода настройка инструментов веб-аналитики, используемой медиаорганизацией, включает несколько этапов и операций. После получения доступа к веб-аналитике и тщательного изучения текущего состояния инструмента в ее работу были внесены определенные модификации, направленные на получение более подробной информации о функционировании сайта и активности аудитории на нем. На основании полученных данных за весь период использования инструмента веб-аналитики, были даны рекомендации по улучшению взаимодействия издания с аудиторией, они касались демографических и поведенческих параметров. Данные веб-аналитики легли в основу технического задания по разработке демонстрационной версии программы-парсера новостей, в которую интегрирован искусственный интеллект. Технология работы программы-парсера представляет собой следующий цикл: сбор и очистка популярных новостей; помещение их в локальную базу данных; оповещение пользователя о появлении свежих новостей посредством чат-бота. У пользователя есть возможность совершить рерайт выбранной новости на месте, используя интегрированную нейросеть. После завершения работы над программой-парсером она была загружена на удаленный сервер для автономного функционирования, что дало возможность использовать ее в любое время независимо от состояния компьютера, на котором осуществлялась разработка.

Заключение

Бурное развитие технологий привело к увеличению объема данных, генерируемых человечеством. Сегодня огромное количество людей пользуются на постоянной основе различными устройствами, сервисами, социальными сетями, платформами и другими возможностями информационных технологий. В ответ на этот вызов происходит внедрение новых технологий обработки информации и повышение уровня автоматизации различных процессов. Критически важными это стало для сферы массовых коммуникаций, в которой объемы сведений, требующих изучения, обработки и использования, приобрели неуправляемые масштабы.

Возможность сделать взаимодействие журналистов и редакций СМИ с большими данными более эффективным была изучена на примере деятельности редакционного коллектива информационного портала «Площадь Свободы» г. о. Тольятти Самарской области. Обнаружилось, что ресурсы больших данных не используются редакцией в полной мере. Со стороны руководства редакции был проявлен интерес к оптимизации процесса и предоставлен доступ к имеющемуся у них инструменту веб-аналитики. После изучения инструмента была осуществлена модернизация его главной панели. Это позволило получить значимые для информационной политики издания характеристики аудитории (пол, возрастную группу, местоположение (на уровне региона, области и города)), долгосрочные интересы, а также синтезировать сведения о ее поведенческой активности на сайте (источники трафика (откуда пользователи приходят), источники содержания (какие страницы чаще всего просматривают)). На основе этих данных у редакции появилась возможность создавать и распространять контент, более четко соответствующий особенностям и запросам аудитории.

Результаты работы с большими данными с помощью веб-аналитики были учтены при разработке демонстрационной версии программы-парсера популярных новостей для информационного портала «Площадь свободы». Использование разработанного программного обеспечения в работе редакции, во-первых, помогло сократить время, которое сотрудники затрачивали на подготовку новостного контента – это было обеспечено за счет автоматизированного мониторинга последних новостей, а также возможности совершить рерайт выбранной новости на месте; во-вторых, способствовало повышению популярности информационного портала «Площадь Свободы» за счет своевременного встраивания его ленты новостей в актуальную повестку и попадания в ведущие позиции рейтингов. Результаты проведенного исследования и эксперимента доказывают актуальность и практическую значимость обращения к большим данным и их использования для оптимизации работы конкретной редакции. Следующим этапом работы может стать модернизация программы-парсера за счет расширения количества источников по сбору данных. Это, в свою очередь, позволит редакции еще больше экономить временные и человеческие ресурсы.

Список используемой литературы и используемых источников

1. Абдиназар С., Шаймерген Г. Big data как феномен инновационной журналистики // Вестник Кабардино-Балкарского государственного университета: Журналистика. Образование. Словесность. 2021. Т. 1. № 3. С. 99–121.
2. Аникеева А. Е. Применение технологии «Большие данные» в энергетике // Современные проблемы телекоммуникаций : Материалы Всероссийской научно-технической конференции с международным участием, Новосибирск, 19–20 апреля 2023 года / Под редакцией А.В. Ефимова, Т.И. Монастырской. Новосибирск : Сибирский государственный университет телекоммуникаций и информатики, 2023. С. 158–163.
3. Арсентьева А. Д., Морозова А. А. Проблемы внедрения алгоритмов искусственного интеллекта в российскую журналистику // Огарёв-Online. 2021. № 2 (155). С. 8–15.
4. Аюшеева И. З. Большие данные: проблемы определения гражданско-правового режима // Lex russica. 2023. Т. 76. № 10 (203). С. 125–134.
5. Баранова Е.А., Шнайдер А.А. Формы подачи материалов в дата-журналистике // Litera. 2022. № 3. С. 98–107. URL: https://nbpublish.com/library_read_article.php?id=37556 (дата обращения: 07.11.2023).
6. Безденежных М. А. Способы агрегации новостей в журналистике // Журналистика, мультимедиа: информационный и социокультурный потенциал : сборник научных трудов / Кубанский государственный университет. Краснодар : Кубанский государственный университет, 2021. С. 227–231

7. Бекназарова С. С. Big Data and advanced analytics в образовании // BIG DATA and Advanced Analytics. BIG DATA и анализ высокого уровня. 2024. № 2. С. 67–77.
8. Белодед Н. И., Хорошун Е. С. Экономика данных: от Big Data к искусственному интеллекту // Технологическая независимость и конкурентоспособность Союзного государства, стран СНГ, ЕАЭС. 2023. Т. 1. С. 495–498.
9. Бондарчик В. В. Журналистика данных в 2021 году: обзор результатов глобального опроса // Журналистыка – 2022: стан, праблемы і перспектывы : матэрыялы 24-й Міжнар. навук.-практ. канф., Мінск, 3 лістап. 2022 г. / Беларус. дзярж. ун-т ; рэдкал.: Л. Р. Хмель (гал. рэд.) [і інш.]. – Мінск : БДУ, 2022. С. 18–22.
10. Вартанова Е.Л., Гладкова А.А. Цифровой капитал в контексте концепции нематериальных капиталов // Медиаскоп. 2020. Вып. 1. URL: <http://www.mediascope.ru/2614> (дата обращения: 17.03.2023).
11. Гаврилов В. Н., Чиркаев С. А., Рассказов А. А. Правовое регулирование пользовательского соглашения в России // Вестник Саратовской государственной юридической академии. 2023. №. 3 (152). С. 155–161.
12. Гаврилюк Н. П., Сорочан В. В. Социум цифровой эпохи–вызовы, риски или новые возможности? // Этносоциум и межнациональная культура. 2022. № 12. С. 98–107.
13. Гарда: 81% россиян обеспокоены вопросами утечки персональных данных // Гарда Технологии - системы информационной безопасности. URL: <https://gardatech.ru/news/garda-81-rossiyan-obespokoeny-voprosami-utechki-personalnykh-dannykh/> (дата обращения: 17.02.2024).
14. Российская Федерация. Национальный стандарт. ГОСТ Р ИСО/МЭК 20546-2021 Информационные технологии. Большие данные. Обзор и словарь от 01.11.2021 // Федеральное агентство по техническому

регулированию и метрологии: офиц. сайт. URL: <https://protect.gost.ru/document1.aspx?control=31&baseC=6&page=4&month=4&year=-1&search=&id=240981> (дата обращения 14.09.2023).

15. Давыдова П. Специфика работы объединений дата-журналистов в Швеции // Слово в науке. 2022. Спецвыпуск. С. 67–70.

16. Дело «ВКонтакте» vs «Дабл Дата» // Legal Insight. URL: <https://legalinsight.ru/articles/delo-vkontakte-vs-dabl-data/> (дата обращения: 01.12.2023).

17. Ерженин Р. В. и др. Проектное обучение: создание инструментария обработки больших данных для использования в дата-журналистике // Информационные и математические технологии в науке и управлении. 2021. № 4 (24). С. 111–124.

18. Заргарян М. Т., Гришин Н. А., Садченко К. Э. Современные коммуникации в цифровой среде // Цифровая трансформация: тенденции и перспективы. I Международная научнопрактическая конференция (Москва, 21 декабря 2022 г.) / под ред. Н.Л. Кетовой и М.Т. Заргарян // Сборник трудов конференции М. : Мир науки, 2022. С. 97–103.

19. Зиниша О. С., Кочаян Д. Г., Мокосеева М. А. Технология Big Data в бизнесе – преимущества и пути совершенствования // Colloquium–Journal. 2020. № 11 (63). С. 154–158.

20. Иванов А. Д. Чат-бот в Telegram и ВКонтакте как новый канал распространения новостей // Вестник Волжского университета им. В. Н. Татищева. 2016. Т. 1. № 3. С. 126–132.

21. Иванченко О. В., Барауля Е. В. Развитие программ лояльности в условиях цифровизации маркетинговой деятельности // Вестник Ростовского государственного экономического университета (РИНХ). 2021. № 2 (74). С. 109–115.

22. Иляхина А. А., Деева И. В. Перспективы применения технологий искусственного интеллекта в журналистике // Вестник науки. 2024. Т. 3. № 1 (70). С. 580–588.
23. Инновации в сфере данных для целей развития // Организация Объединенных Наций | Мир, достоинство и равенство на здоровой планете : офиц. сайт. URL: <https://www.un.org/ru/global-issues/big-data-for-sustainable-development> (дата обращения 18.11.2022).
24. Ким М. Н. Инновационные практики в работе мультимедийных журналистов // Управленческое консультирование. 2023. № 4 (172). С. 72–80.
25. Корнев М. С. История понятия «большие данные» (Big Data): словари, научная и деловая периодика // Вестник РГГУ. Серия: История. Филология. Культурология. Востоковедение. 2018. № 1. С. 81–85.
26. Макарова Н. Я. Журналистика данных в системе профессиональных компетенций журналиста // Знак: проблемное поле медиаобразования. 2020. № 4 (38). С. 44–52.
27. Мамедова Н. М. Человек в эпоху цифровизации: на грани реального и виртуального // Век глобализации. 2021. № 3. С. 74–85.
28. Маркеева А. В., Гавриленко О. В. Большие данные как исследовательская технология: возможности и ограничения применения в современной управленческой практике // Общество: социология, психология, педагогика. 2021. № 12 (92). С. 94–103.
29. Нигматуллина М. Как розничные сети и интернет-магазины используют собственные данные для принятия решений в бизнесе // Инновации и инвестиции. 2023. № 5. С. 131–134.
30. Новосильцева, Т. Н. Обработка больших данных в дата-журналистике // Майские чтения (язык и репрезентация культурных кодов) : XI Всерос. с междунар. участием науч. конф. молодых ученых (Самара, 13-15 мая 2021 г.) : материалы и доклады / М-во науки и высш. образования Рос.

Федерации, Самар. нац. исслед. ун-т им. С. П. Королева (Самар. ун-т), Соц.-гуманитар. ин-т, Фак. филологии и журналистики ; под общ. ред. А. А. Безруковой. - Самара : Инсома Пресс, Ч. 1. 2021. С. 86–90.

31. Общий регламент защиты персональных данных (GDPR) Европейского союза // Текст GDPR на русском с комментариями и ссылками | GDPR-Text.com. URL: <https://gdpr-text.com/ru/> (дата обращения: 04.05.2024).

32. Петровская О. В. Цифровая трансформация и проблемы обеспечения достоверности информации // Аграрное и земельное право. 2020. № 3 (183). С. 130–132.

33. Пискорская С. Ю., Гончаров А. Е. Феномен больших данных в социальной философии и профессиональном образовании // Профессиональное образование в современном мире. 2018. Т. 8. № 4. С. 2178–2185.

34. Погудина Р. М., Шубина А. С. Разновидности чат-ботов и их роль в современной журналистике // Медиа как фактор адаптации человека к социальным, экономическим и политическим изменениям : сборник материалов Международной научно-практической конференции (Екатеринбург, 20–22 апреля 2023 г.). Екатеринбург : Издательство Уральского университета, 2023. С. 156–160.

35. Попов А. А. Технические и этические стандарты в сфере больших данных // Гуманитарные и политико-правовые исследования. 2023. № 4 (23). С. 67–77.

36. Решетникова К. В. Гражданская журналистика как объект развития цифровой журналистики // Передовые научно-технические и социально-гуманитарные проекты в современной науке. Сборник статей VI международной научно-практической конференции. Москва: «Научно-издательский центр «Актуальность.РФ», 2022. С. 118–119.

37. Роскомнадзор: Федеральная служба по надзору в сфере связи, информационных технологий и массовых коммуникаций: офиц. сайт. URL: <https://rkn.gov.ru/> (дата обращения 17.02.2024).

38. Российская Федерация. Законы. Гражданский кодекс Российской Федерации : ГК // КонсультантПлюс : сайт. URL: https://www.consultant.ru/document/cons_doc_LAW_5142/ (дата обращения: 17.02.2024).

39. Российская Федерация. Законы. О коммерческой тайне : Федер. закон № 98-ФЗ : принят Государственной Думой 9 июля 2004 г. : одобрен Советом Федерации 15 июля 2004 г. : послед. ред. // КонсультантПлюс : сайт. URL: https://www.consultant.ru/document/cons_doc_LAW_48699/ (дата обращения: 17.02.2024).

40. Российская Федерация. Законы. О персональных данных: Федер. закон № 152-ФЗ : принят Государственной Думой 8 июля 2006 г. : одобрен Советом Федерации 14 июля 2006 г. : послед. ред. // КонсультантПлюс URL: https://www.consultant.ru/document/cons_doc_LAW_61801/ (дата обращения: 17.02.2024).

41. Российская Федерация. Законы. Об информации, информационных технологиях и о защите информации : Федер. закон № 149-ФЗ : принят Государственной Думой 8 июля 2006 г. : одобрен Советом Федерации 14 июля 2006 г. : послед. ред. // КонсультантПлюс URL: https://www.consultant.ru/document/cons_doc_LAW_61798/ (дата обращения: 17.02.2024).

42. Российская Федерация. Законы. Трудовой кодекс Российской Федерации : ТК // КонсультантПлюс : сайт. URL: https://www.consultant.ru/document/cons_doc_LAW_34683/ (дата обращения: 17.02.2024).

43. Самойленко Н. С. Конвергенция журналистики, медиакоммуникаций и IT // Актуальные вопросы современной филологии и журналистики. 2023. №. 1 (48). С. 121–128.

44. Симакова С. И. Медиаэстетический код инфографического контента в журналистике. Челябинск: Челябинский государственный университет. 2022. С. 160.

45. Симарова И. С., Алексеевичева Ю. В., Жигин Д. В. Цифровые компетенции: понятие, виды, оценка и развитие // Вопросы инновационной экономики. 2022. Т. 12. № 2. С. 935–948.

46. Сколько «весит» вся информация в интернете? // Вокруг света - первый познавательный портал | Вокруг Света. URL: <https://www.vokrugsveta.ru/quiz/329408/> (дата обращения 18.11.2022).

47. Смирнова О. Б., Ламонина Л. В., Дунцов А. Н. О технологии больших данных // Инновационные технологии в АПК, как фактор развития науки в современных условиях : Сборник международной научно-исследовательской конференции, посвященной 70-летию создания факультета ТС в АПК (Мех ФАК), Омск, 26 ноября 2020 года. Омск: Омский государственный аграрный университет имени П.А. Столыпина, 2020. С. 724–727.

48. Соломин, В. Е. Тренды развития мультимедийной (цифровой) журналистики и новых медиа // Медиа и коммуникации: состояние, проблемы, перспективы : Сборник статей Национальной научно-практической конференции (к 30-летию кафедры журналистики и русской литературы XX века), Кемерово, 23 октября 2021 года / Под общей редакцией А.В. Чепкасова, Ф.С. Рагимовой . Кемерово: Кемеровский государственный университет, 2022. С. 89–93.

49. Тимохин М. Ю., Шаранин В. Ю. Искусственный интеллект и теория принятия решений: современные тенденции // Инженерный вестник Дона. 2023. № 10 (106). С. 33–43.
50. Ударцева О. М. Эффективный библиотечный сайт: data-driven-подход к управлению сайтом с применением аналитических инструментов // Библиосфера, 2021. № 2. 65–76.
51. Файль В. П., Кунгурова О. Г. Тенденции развития дата-журналистики в СНГ // Поколение независимости: ориентиры и перспективы. 2021. С. 651–656. URL: https://repo.kspi.kz/bitstream/handle/123456789/4111/2021-konf-stud-magistr_652-657.pdf?sequence=1&isAllowed=y (дата обращения: 04.01.2023).
52. Федорова Л. А., Ху Гуйюй, Хуан Сяоянь, Землякова С. А. Применение технологий Big Data в деятельности современных предприятий // Вестник Алтайской академии экономики и права. 2020. № 9. С. 322–329.
53. Хапов, У. А. Дата-журналистика: становление и особенности // EurasiaScience : Сборник статей XXXIX международной научно-практической конференции, Москва, 15 августа 2021 года. Москва: Общество с ограниченной ответственностью "Актуальность.РФ", 2021. С. 65–66.
54. Хрущева А. А. Чат-боты в бизнес-коммуникации: виды и функции // Медиасреда. 2022. № 1. С. 154–159.
55. Цапина Т. Н. Чат-боты и возможности их использования в практике работы современных компаний // Научные дискуссии в условиях мирового кризиса: новые вызовы, взгляд в будущее. 2022. Т. 1. С. 347–351.
56. Четырбок П. В. Информационные технологии обработки больших данных // Теория и практика экономики и предпринимательства / Труды XX T338 Международной научно-практической конференции. Симферополь-Гурзуф, 20-22 апреля 2023 год / Под ред. проф. Н. В. Апатовой. Симферополь: Издательский дом КФУ им. В. И. Вернадского, 2023. С. 276–278.

57. Шайдуллина В. К. Большие данные и защита персональных данных: основные проблемы теории и практики правового регулирования // Общество: политика, экономика, право. 2019. № 1 (66). С. 51–55.
58. Шах А. В., Шапович Е. Г. Чат-боты как современный инструмент маркетинга // Стратегия и тактика развития производственно-хозяйственных систем : сб. науч. тр. / М-во образования Респ. Беларусь, Гомел. гос. техн. ун-т им. П. О. Сухого, Гомел. обл. орг. о-ва «Знание» ; под ред. В. В. Кириенко. Гомель : ГГТУ им. П. О. Сухого, 2019. С. 200–203.
59. Шилина А. Г. Основные характеристики материалов журналистики данных в зарубежной и российской качественной прессе: автореф. дис. канд. филол. наук: 10.01.10. М., 2019. С. 33.
60. Clark A. M., Rodríguez J. Big Data and Journalism: How American Journalism is Adopting the Use of Big Data // *Novum jus*. 2021. No. 1. Pp. 69–89.
61. Heravi B. R., Lorenz M. Data Journalism Practices Globally: Skills, Education, Opportunities, and Values // *Journalism and Media*. 2020. No 1. Pp. 26–40.
62. Komilov M. Analysis of big data and its practical use in business // *Science and technology in the modern world*. 2024. No. 2. Pp. 43–45.
63. Newman N., Fletcher R., Schulz A., Andi S., Robertson C. T., Nielsen R. K. Reuters Institute digital news report 2021 // Reuters Institute for the study of Journalism. 2021.
64. Qingyang L. Журналистика в цифровом медиaprостранстве // *Global science: prospects and innovations*. 2023. Pp. 539–542.