

DOI: [10.14515/monitoring.2021.5.1970](https://doi.org/10.14515/monitoring.2021.5.1970)



Д. А. Лесневский, И. Г. Рекун

РЫНОЧНЫЙ ОПЫТ СОЗДАНИЯ ПЛАТФОРМЫ ДЛЯ АНАЛИЗА АККАУНТОВ БЛОГЕРОВ И ИХ АУДИТОРИИ

Правильная ссылка на статью:

Лесневский Д. А., Рекун И. Г. Рыночный опыт создания платформы для анализа аккаунтов блогеров и их аудитории // Мониторинг общественного мнения: экономические и социальные перемены. 2021. № 5. С. 178—184. <https://doi.org/10.14515/monitoring.2021.5.1970>.

For citation:

Lesnevsky D. A., Rekun I. G. (2021) Market Experience of Creating a Platform for Analysing Blogger Accounts and Their Audience. *Monitoring of Public Opinion: Economic and Social Changes*. No. 5. P. 178–184. <https://doi.org/10.14515/monitoring.2021.5.1970>. (In Russ.)

РЫНОЧНЫЙ ОПЫТ СОЗДАНИЯ ПЛАТФОРМЫ ДЛЯ АНАЛИЗА АККАУНТОВ БЛОГЕРОВ И ИХ АУДИТОРИИ

ЛЕСНЕВСКИЙ Дмитрий Александрович — Бакалавр социальных наук, ко-фаундер, Optia.ai, Москва, Россия
E-MAIL: d.lesnevsky@optia.ai
<https://orcid.org/0000-0002-2721-1831>

РЕКУН Игорь Геннадьевич — магистр прикладной математики и физики, ко-фаундер, Optia.ai, Москва, Россия
E-MAIL: i.rekun@optia.ai
<https://orcid.org/0000-0003-3912-9891>

Аннотация. Цель работы — структурировать опыт разработки инструмента для анализа блогеров и их аудиторий в социальных сетях. В статье представлены основные этапы создания этого инструмента: от интервью с потенциальными клиентами для планирования функционала к концептуальной модели, методам обработки данных и перспективам. Серия интервью ($N=20$) показала, что подходы к анализу аудитории блогеров на рекламном рынке варьируются по глубине проработки в зависимости от задач. Аккаунты блогеров оценивали как по количественным показателям (объем аудитории), так и по качественным — с точки зрения соответствия имиджу брендов. Интерес для развития бизнеса представили именно аудиторные показатели. Модель для описания пользователей социальной сети (а следовательно — для аудиторий) сложилась из трех составляющих: контента самого аккаунта, контента его аудитории и контента других аккаунтов, с которыми он взаимодействует. Обработка массива данных потребовала технических решений на основе машинного обучения

MARKET EXPERIENCE OF CREATING A PLATFORM FOR ANALYSING BLOGGER ACCOUNTS AND THEIR AUDIENCE

*Dimitry A. LESNEVSKY*¹ — BA in Social Sciences, Co-founder
E-MAIL: d.lesnevsky@optia.ai
<https://orcid.org/0000-0002-2721-1831>

*Igor G. REKUN*¹ — MA in Applied Mathematics and Physics, Co-founder
E-MAIL: i.rekun@optia.ai
<https://orcid.org/0000-0003-3912-9891>

¹ Optia.ai, Moscow, Russia

Abstract. The purpose of this study is to structure the experience of developing a tool for analysing bloggers and their audiences on social media. The paper presents the main stages of creating this tool: starting with interviews with potential clients for planning functionality and going up to a conceptual model, data processing methods, and its perspectives. A series of interviews ($N=20$) showed that approaches to analysing the audience of bloggers in the advertising market vary in their depth depending on the tasks. Bloggers' accounts were assessed both in terms of quantitative indicators (size of the audience) and qualitative ones, i.e., in terms of compliance with the brand image. Audience indicators were of interest for business development. The model for describing users of a social media (and therefore for audiences) has developed from three components: the content of the account itself, the content of its audience, and the content of other accounts with which it interacts. The processing of the data set required technical solutions based on machine learning for the analysis of

для анализа текстов и изображений. В результате мы получили как наборы классических индикаторов — пол, возраст, географическое расположение, так и векторное представление поведенческих характеристик аудитории. Вероятно, дальнейшее развитие технологий приведет к уменьшению возможностей по независимому анализу социальных медиа. Социальные сети ограничивают доступ к API, повышая стоимость разработки решений, а законодательство в области персональных данных увеличивает риски для компаний, создающих новые продукты в этой области.

Ключевые слова: блогеры, социальные сети, анализ контента, графы, анализ аудиторий

texts and images. As a result, the authors obtained both sets of classic indicators (gender, age, geographic location) and a vector representation of the behavioral characteristics of the audience. The further development of technology will probably lead to a decrease in the ability to independently analyse social media. Social media restricts access to APIs, increasing the cost of developing solutions, and legislation in the field of personal data increases the risks for companies creating new products in this sector.

Keywords: bloggers, audience analysis, social media, content analysis, graphs

Постановка проблемы и методология исследования

В 2018 г. число уникальных пользователей Instagram достигло одного миллиарда в месяц по всему миру (37 млн в РФ), благодаря чему платформа превратилась в одного из основных игроков рекламного рынка нашей страны. Одновременно с ростом аудитории увеличивался интерес к размещениям рекламы у блогеров, где, по оценкам компаний, работающих на этом рынке, прогнозы роста составляли от 13% до 100% в год¹.

Поверхностный обзор рынка и происходящих на нем процессов показал, что сложившиеся подходы к анализу и подбору блогеров можно улучшить. Мотивацией для создания нашей компании послужило то, что предложений для точного анализа блогеров недостаточно, а спрос на подобные услуги есть.

Для проверки этой гипотезы в июне — августе 2018 г. мы провели в Москве серию интервью с потенциальными заказчиками из 20 компаний: восемь были представителями брендов, и еще 12 представляли рекламные агентства. Целью интервью была проверка спроса на инструмент для анализа блогеров и предварительные продажи, а основными темами стали: ожидания от работы с блогерами; критерии выбора их кандидатур; виды и содержание рабочих процессов; и, наконец, потенциальные запросы к инструменту. 18 интервью прошли в формате групповых встреч представителей нашей компании с потенциальными заказчиками, два — в формате телефонных разговоров. Информантов подбирали методом снежного кома.

¹ Николаева Ю. Объем российского рынка рекламы у блогеров приблизился к 5 млрд рублей в год // Sostav. 2018. 15 марта. URL: <https://www.sostav.ru/publication/obem-rossijskogo-rynka-reklamy-u-blogerov-priblizilsya-k-5-mlrd-rublej-v-god-30795.html> (дата обращения: 09.10.2021).

Форматы проектов

По результатам интервью мы условно разделили все рекламные проекты с блогерами на три основных формата в зависимости от их фокуса:

- 1) охватные кампании;
- 2) performance;
- 3) спецпроекты.

Цель *охватных кампаний* — донести ключевые сообщения до целевой аудитории. Показатели эффективности, по которым оцениваются рекламные размещения, — просмотры публикаций и социальные действия: лайки и комментарии.

Performance-проекты ориентированы на трафик — переходы на сайты или в другие профили Instagram и дальнейшие продажи. Постановка такой цели возможна для проектов, где эти метрики можно и целесообразно отслеживать. Примером может служить кейс продвижения каршеринга Anytime².

Спецпроекты — разнообразные варианты нестандартного сотрудничества блогеров и брендов: участие в мероприятиях, разработка совместных продуктовых линеек, например — коллекция Reebok Crossfit X Pokras Lampas³.

Процессы подбора и анализа аккаунтов блогеров

Некоторые компании сотрудничали с блогерами с помощью подрядчиков, другие решали задачу внутренними ресурсами. Вне зависимости от формата проектов, процесс состоял из трех этапов:

- 1) анализ и формирование списка кандидатур;
- 2) согласование или доработка списка;
- 3) размещение и оценка результатов.

По словам информантов, основные трудозатраты уходили на первые два этапа. Практики оценки варьировались от субъективных ручных оценок до попыток оцифровать и спрогнозировать эффект от размещения.

Стандартная методология ручной оценки заключалась в просмотре постов, изучении профилей комментаторов и пользователей, подписанных на блогера. Несмотря на очевидные методологические проблемы, такой способ позволял быстро оценить качество аудитории и имиджевое соответствие бренду.

Развитие и продолжение этого способа — скоринговые таблицы, позволяющие формализовать и ускорить процесс подбора и согласования кандидатур блогеров. Алгоритм работы с таблицами оказывается следующим:

- 1) формируется список блогеров;
- 2) каждый блогер оценивается по набору критериев;
- 3) критерии суммируются с установленными весами;
- 4) список сортируется по убыванию, идут переговоры с блогерами.

Критерии и их веса формируются индивидуально под рекламную кампанию. Примерами критериев могут быть: 1) доля целевой аудитории, 2) имиджевое соответствие бренду, 3) наличие популярных блогеров в подписчиках, 4) уровень вовлеченности, 5) отсутствие рекламы конкурентов и др.

² Пыжов К. Кейс Perfluence: продвижение каршеринга Anytime в Белоруссии и 500+ постов от блогеров // VC.ru. 2021. 22 июня. URL: <https://vc.ru/marketing/261671-keys-perfluence-prodvizhenie-karsheringa-anytime-v-belorussii-500-postov-ot-bloggerov> (дата обращения: 09.10.2021).

³ Подробнее см.: URL: https://www.instagram.com/p/Bgn1_HMnuqF/ (дата обращения: 09.10.2021).

Часть агентств, которая занималась блогерами продолжительное время, использовала для оценки исторические данные по эффективности размещений. Они помогали спрогнозировать границы показателей эффективности, которые мог ожидать клиент в рамках кампании, и, следовательно, позволяли выбрать оптимальные кандидатуры блогеров для сотрудничества.

Потребности потенциальных клиентов

Все участники интервью соглашались, что процессы изучения и подбора блогеров можно улучшить, в частности, повысить скорость и точность анализа аккаунтов. Чаще всего называли следующие запросы к инструменту для анализа:

- Возможность оценить пересечения аудиторий между аккаунтами и посчитать долю эксклюзивной аудитории для каждого блогера из произвольного списка;
- Возможность оценить долю «некачественной» аудитории — ботов и коммерческих аккаунтов;
- Возможность оценить долю целевой аудитории по сочетанию пола, возраста, географии и профиля интересов.

Важно отметить, что мы не выявили интереса потенциальных заказчиков к оценке имиджевого соответствия блогеров бренду, хотя все участники интервью обозначали важность этого критерия для рабочих процессов. Отсутствие интереса объяснялось тем, что люди с такой задачей априори справляются лучше.

Постановка бизнес-задачи

Мы поставили своей целью разработать решение, которое бы помогало планировать *охватные* и *performance* кампании с блогерами. Эффективность таких кампаний, как показали интервью, зависит от трех факторов:

- 1) объема целевой аудитории у блогера;
- 2) креативной составляющей;
- 3) бюджета.

Стоимость сотрудничества и креативная составляющая могут меняться в зависимости от проекта. Соответственно, нашей задачей стала оценка целевой аудитории для каждого аккаунта, чтобы клиенты могли формировать оптимальный список кандидатур для сотрудничества, минимизируя стоимость контакта с представителем целевой аудитории.

Чтобы проект был экономически целесообразен, нам нужно было создать решение, способное с достаточной для рынка точностью находить любые аудиторные сегменты. Для определения целевой аудитории мы построили концептуальную модель из трех характеристик:

- 1) география;
- 2) демография (пол и возраст);
- 3) профиль интересов и поведенческих факторов.

Основная методологическая проблема, с которой мы столкнулись при техническом воплощении этой модели, — отсутствие бенчмарков и размеченных наборов данных, которыми можно было бы проверить ее качество. Как следствие, тестирование подходов проводили вручную, проверяя выборку профилей на соответствие результатам.

Техническая реализация

Исходные данные для анализа открыты и доступны в социальной сети. Правильное их использование позволяет оценить все характеристики из концептуальной модели.

География определяется по наиболее распространенным геотегамам к публикациям. Такой подход работает с открытыми аккаунтами, в которых есть достаточное количество данных — 15 % аудитории. В случае, если данных недостаточно, дополнительным источником информации становятся социальные связи пользователя.

Наиболее прямой подход для определения демографических характеристик — распознавание пола и возраста всех людей, присутствующих на фотографиях в профиле, и усреднение оценки. Источником ошибок в данном случае будут профили, в которых пол и возраст людей на фотографиях смещены относительно владельца профиля. Для профилей без фотографий людей оценка пола и возраста невозможна.

Более корректный подход — задействовать метаданные, описание профиля и связи с другими аккаунтами. Для получения обучающей выборки можно использовать профили, где указан возраст или присутствуют фотографии с дня рождения. На данной выборке обучается регрессионная модель, которая ранжирует вес подписок на маркерные аккаунты, вклад возраста стандартных аккаунтов, для которых существуют взаимные связи, и вклад якорных аккаунтов, для которых известно истинное значение возраста на основе биографии и/или данных о возрасте из описаний к фотографиям.

Профиль интересов и поведенческих факторов оценить сложнее, для каждого пользователя существует три источника информации:

- 1) метаданные к контенту;
- 2) социальный граф;
- 3) контент профиля (тексты и изображения).

Для оценки некоторых характеристик, например, частоты путешествий или интереса к определенным тематикам, достаточно только одного источника. Большая часть других характеристик требует анализа крупных объемов текстового и визуального контента.

Без ограничения общности группировку визуального и текстового контента выполняют одним из методов кластеризации. На качество результата напрямую влияет метод векторизации данных, так как этап кластеризации агностичен к содержанию и оперирует в векторном пространстве. Для визуального контента могут применяться сверточные нейронные сети, обученные на задаче классификации по базе данных ImageNet или под задачу метрического обучения (metric learning) на целевом содержимом.

Для представлений текстового содержимого можно использовать как классические подходы (LDA), так и подходы с применением нейронных сетей. LDA показал неудовлетворительные результаты (silhouette score на тестовом «датасете» составил 0,43) при работе с контентом из Twitter и Instagram ввиду разреженности представлений модели «Сумка слов» (Bag of Words).

Модели класса «трансформер» (transformer), обученные на задаче «перефразирование» (paraphrasing) или «семантическое сходство» (semantic similarity), пока-

зывают лучшие результаты, однако требуют размеченных данных. Получение хороших представлений в отсутствие размеченного «датасета» возможно, но требует значительно больших вычислительных ресурсов. Доминирующим подходом здесь оказывается обучение авторегрессионных моделей. При достаточном размере «датасета» и модели (20 млрд слов и 2 млрд параметров) проявляются свойства мета-обучения: представления, извлекаемые моделью, становятся пригодны для семантической кластеризации (silhouette score составил 0,65).

Дополнительный подход к оценке интересов, который мы обнаружили во время разработки, основан на контенте блогера, аудиторию которого необходимо изучить:

1) строим для каждого пользователя многомерный вектор, соответствующий взаимодействию с постами из профиля блогера;

2) выделяем группы пользователей, взаимодействующих с похожим контентом с помощью метода k-средних (k-means);

3) оцениваем тематики контента вручную, чтобы интерпретировать аудиторные кластеры.

Основное преимущество этого метода — относительная простота реализации и низкие требования к вычислительным мощностям. Основные недостатки — ресурсоемкость в плане человеческих ресурсов для интерпретации данных и неполнота информации об аудитории. Сочетание этих недостатков не позволяет эффективно использовать метод как основу для самостоятельного продукта.

Рыночные задачи показали, что жизнеспособное решение для анализа онлайн-аудитории блогеров — сочетание подходов, основанных на всей доступной информации. В этом случае каждый пользователь социальной сети характеризуется данными из своего профиля, социального окружения и контента, с которым он взаимодействует. Группы пользователей, в свою очередь, определяются через сочетание признаков из каждой категории данных.

Перспектива

Любая работа в области анализа социальных медиа сопряжена с рядом сложностей при сборе данных. Платформы склонны ограничивать возможности API, повышая стоимость разработки решений, а неопределенный правовой статус повышает риски коммерческого использования подобных технологий.

Если не учитывать эти риски, наиболее перспективные, на наш взгляд, направления для дальнейшей работы — объединение профилей между разными платформами и улучшение алгоритмов для обработки текстовой информации.

Технические решения для обработки текстов позволят извлечь больше полезной информации из всего массива данных. Существующие подходы для тематического моделирования и определения тональности недостаточно точны, чтобы отследить все нюансы в поведении пользователей.

Объединение профилей технически осуществимо через поиск идентичного контента в аккаунтах. Более ресурсоемкий способ — алгоритмы распознавания и верификации лиц. Решение этой задачи позволит работать с более полной информацией об аудитории каждого блогера.