Применение автоматизированного сбора информации из сообществ социальных сетей для выявления активных пользователей

Б.А. Низомутдинов, Л.А. Видясова

***Аннотация* — В статье представлен разрабатываемый авторами метод определения наиболее активных подписчиков сообществ в социальных сетях. Объектом изучения выступили 18 официальных групп районных администраций Санкт-Петербурга ВКонтакте. В работе описывается реализация алгоритма сбора и анализа данных. В ходе исследования было определено общее количество комментариев в каждом сообществе и выделены наиболее активные подписчики, оставляющие больше всего комментариев под постами. С применением автоматизированного инструментария парсинга данных социальных сетей проведен анализ социально-демографических характеристик активных пользователей. Все собранные данные были деперсонализированы. Результаты исследования показали существование в каждом сообществе ядра активных пользователей, которые оставляют больше всего реакций на посты, публикуемые в группах. Сделан вывод о большом потенциале, который можно извлечь из применения автоматизированного сбора данных из социальных сетей.**

***Ключевые слова* — социальные сети, общественные активисты, активные пользователи, лидеры мнений, парсинг, анализ социальных сетей, VK API, парсинг социальных сетей.**

I. Введение

Для успешного мониторинга и управления социально-экономическим пространством города необходимо объединение традиционного социально-психологического подхода для учета ценностно-эмоционального капитала жителей и достижений современных цифровых технологий. Большие объемы информации из социальных сетей в настоящее время представляет все больший интерес для исследователей, изучающих поведение людей в городах. Тем не менее, ценность данных социальных сетей для исследований в области устойчивого развития городов все еще недостаточно изучена.

В последние годы активное вовлечение горожан в развитие городской среды является одним из ключевых приоритетов городского управления на всех уровнях. Создаются различные онлайн площадки для взаимодействия представителей органов власти и населения: порталы решения городских проблем, сервисы электронных обращений и петиций, официальные группы в социальных сетях.

Современные виртуальные городские сообщества — это открытое диалогичное пространства, которое можно идентифицировать как коммуникативную площадку для обмена информацией и опытом, взаимодействия с различными объектами городской среды.

За последние годы социальные сети стали мощным инструментом распространения информации, влияющей на мнения и поведение людей. Популярные аккаунты в социальных сетях охватывают широкую аудиторию, что приводит к появлению в сети профессиональных влиятельных лиц, способных влиять на поведение и выбор потребителей.

В современной медиа-среде активные пользователи играют все более важную роль в регулировании доступа к онлайн-контенту [1]. При оценке достоверности информации люди склонны полагаться на свои социальные сети. Исследование поведения людей, читающих новости в Интернете, показывает, что при определении достоверности новостной статьи чрезвычайно важно, кто ею делится. Люди склонны полагаться на информацию, которую рекомендует человек, которому они доверяют, до такой степени, что одна и та же новость воспринимается как более заслуживающая доверия, если она рекомендована другом из социальной сети, чем при чтении ее на исходном новостном сайте [2].

Официальные площадки для обсуждения вопросов городского развития в сети публикуют новости и информацию по ключевым проблемам, а также собирают комментарии от жителей города. Для того, чтобы оценить реагирование на публикуемый контент, следует учитывать специфику таких реакций, а также возможности определения ключевых пользователей, активно выражающих гражданскую позицию и оказывать влияние на общественное мнение по поднятым вопросам.

Впервые, термин «лидеры мнений» был использован Полом Лазарсфельдом в 1955 г. для обозначения людей, имеющих авторитет в какой-либо социальной группе и способных оказать на нее влияние в процессе обсуждения сообщений [3]. В современной научной

Статья получена 08 ноября 2021.

Низомутдинов Борис Абдуллохонович, Университет ИТМО (email: boris@itmo.ru)

Видясова Людмила Александровна, Университет ИТМО (bershadskaya.lyudmila@gmail.com)

литературе лидеры мнений определяются как заинтересованные и компетентные люди, мнения которых сторонники считают честными и заслуживающими доверия [2]. Лидеры мнений уделяют пристальное внимание рассматриваемой теме, часто обсуждают ее с аудиторией, убеждая других принять то или иное мнение [4]. В то же время, лидеры общественного мнения не всегда занимают влиятельную позицию. Чаще всего, они занимают такое же социальное положение, что и те, на кого они стремятся повлиять, но воспринимаются как хорошо осведомленные по обсуждаемой теме [3,4]. Лидеры общественного мнения обладают несколькими атрибутами, связанными с личностью, опытом или сетями, включая доверие, знания, энтузиазм, взаимосвязь и центральность [5, 6].

II. Сбор данных

В качестве объекта исследования были использованы 18 официальных групп администраций районов Санкт-Петербурга в социальной сети Вконтакте. Данные сообщества относятся к информационным, в таких группах пользователи не могут самостоятельно публиковать посты, единственный вариант коммуникации - комментарии под записями, которые разместил администратор. В данных сообществах регулярно публикуют новости о жизни района, предстоящих событиях и важные анонсы администрации.

Для сбора информации был использован авторский набор инструментов, состоящий из парсера текстового контента, публичного сервиса для сбора информации, сервиса, с использованием методов API VK и специального программного обеспечения для обработки цифровых отпечатков пользователей социальных сетей.

API интерфейс ВКонтакте позволяют загружать посты и комментарии публикаций, всю информацию о пользователях из сообществ, если только эти сообщества не закрыты.

В результате, удалось собрать комментарии исследуемых сообществ и сгенерировать набор данных в закодированном виде об основных характеристиках активных пользователей, без хранения персональной информации.

Для сбора информации о сообществе был задействован метод VK groups.getById. Данный метод возвращает массив объектов, описывающих сообщества. В настройках есть возможность указать интересующие, например количество подписчиков и другие характеристики. Метод wall.get возвращает список записей со стены пользователя или сообщества, метод wall.getComments позволяет получить список комментариев к одной записи на стене. Используя два этих метода можно собирать все комментарии в сообществе. Метод account.getProfileInfo дает возможность собрать информацию о пользователях.

В соответствии с Правилами обработки персональных данных мы разработали инструмент для деперсонализации данных: был оставлен только идентификационный номер без привязки к нему имени/фамилии. Мы использовали следующее шифрование при хранении данных: оценивалась доступность определенной информации без ее значения, таким образом, сохраняя в нашей базе данных только значение 1 или 0; в случае, если значение было числовым, например, количество друзей/подписок, то сохранялась цифра. Такой подход предусматривает анонимный сбор и обработку данных пользователей социальных сетей.

В будущем, запланирован анализ и обработка фотографии пользователей с применением облачной технологии Microsoft Azure, для выделения основных объектов на изображении (аватаре). На данной платформе реализовано множество API-сервисов для компьютерного зрения, машинного обучения, параллельных вычислений и обработок и др. Облачный сервис Azure позволяет использовать свои вычислительные мощности для различных задач, в том числе и для машинного анализа изображений и текста.

В данном исследовании мы анализировали только основные публично-доступные параметры.

Общий алгоритм сбора информации состоял из 3 этапов:

— Сбор и выгрузка данных из исследуемых сообществ, с применением методов VK APIc
— анализ комментариев, поиск наиболее активных авторов;
— парсинг данных из профилей активных подписчиков.

III. Результаты обработки данных

Из официального источника [9] - Управления Федеральной службы государственной статистики по Санкт-Петербургу и Ленинградской области была определена численность постоянного населения в разрезе муниципальных образований Санкт-Петербурга по состоянию на 1 января 2021 года.

Далее, для каждого официального сообщества района во Вконтакте, автоматизированными методами были собраны следующие показатели:

— общее кол-во участников в сообществе;
— % участников от населения;
— общее кол-во комментариев в сообществе;
— всего постов;
— среднее количество комментариев на пост.

В результате автоматизированного сбора данных была получена информация о количестве участников, публикаций и комментариев в каждом сообществе. По данным исследования, численность участников районных сообществ не превышает 3-4% от общей численности населения, проживающего на данной территории, за исключением Курортного (16%), Калининского (10%) и Выборгского районов (6%).

Общая аудитория всех 18 сообщест составила 183440 подписчиков, в ходе анализа было выявлено, что все указанные группы имеют свою уникальную аудитори, для этого, была выгружена вся аудитория всех сообществ, и проведен поиск пересечений, в итоге, большинство подписчиков состоят только в одной из изучаемых групп. Только 7619 пользователей состоят более чем в 2 сообществах из списка, и далее по

убыванию, общее распределения вовлеченности в разные сообщества изображено на риснуке 1.

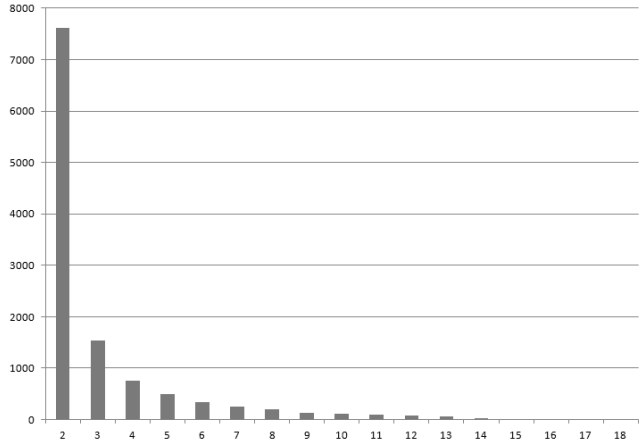

Рис 1. Пересечение аудитории сообществ во ВКонтакте

На рисунке 2 представлено распределение общего количества комментариев пользователей в каждом районном сообществе. Следует отметить, что среди сообществ с максимальной долей участников от всего населения Калининский и Выборгский район также лидируют по количеству публикаций.

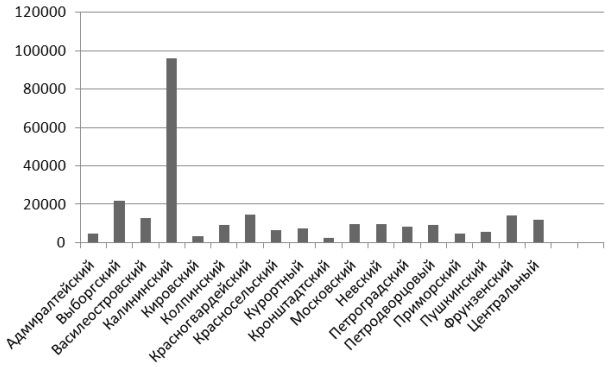

Рис 2. Общее количество комментариев в каждом районе

Дополнительно, с помощью сервиса, на основе VK API, была собрана усредненная информация о коэффициенте вовлеченности пользователей в публикуемый контент, по типу информации, для определения, на какой тип контента в подобных сообществах оставляют больше комментариев и лайков. Подобная информация помогает понять какое содержание у постов в целом привлекает больше внимания аудитории и насколько эффективны публикации.

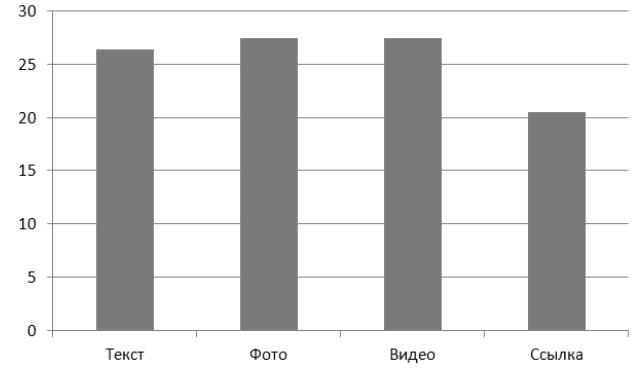

Рис 3. Коэффициент вовлеченности пользователей в публикуемый контент по типу информации

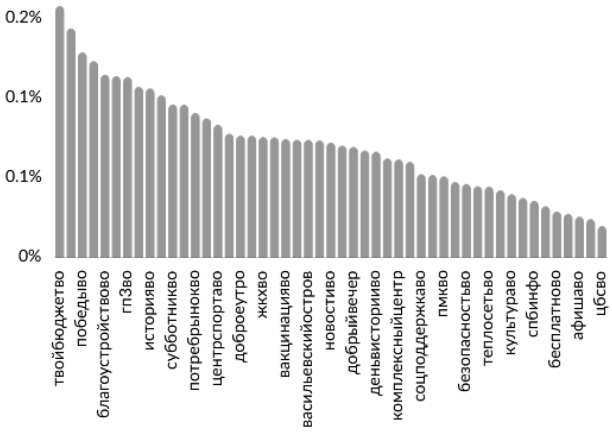

Рис 4. Распределение активности в зависимости от хештега

Часть постов в исследуемых сообществах имеет хештеги, для хештегов был проведен дополнительный расчет активности, построено распределение, на какие хештеги чаще происходит реакция - оставлен комментарий, сделан репост или поставлен лайк. На рисунке 4 представлено распределение активности в сообществе Василеостровского района для разных хештегов за период в 12 месяцев.

В ходе исследования был проведен анализ собранной информации, рассчитаны средние показатели активности пользователей в районных сообществах Санкт-Петербурга. С использованием разработанного скрипта были выделены самые активные подписчики в каждой из групп, подсчитано количество комментариев и лайков в расчете на каждый пост.

Для определения средних показателей активности пользователей в районные сообщества в Санкт-Петербурге были рассчитаны следующие значения:

— общее количество постов;
— средне количество просмотров на пост;
— средне количество лайков на пост;
— % участников, поставивших лайк;
— средне количество комментариев на 1 пост;
— среднее количество репостов на 1 пост.

Полученные данные позволяют говорить о достаточно низкой активности пользователей в сообществах районных администраций в Санкт-Петербурге. Об этом свидетельствуют показатели среднего количества оставленных реакций в виде лайков (одобрения) и комментариев под постами от администрации групп. Однако дальнейший анализ позволили сделать вывод о существовании в каждом сообществе определенных активных пользователей, которые оставляют в разы больше реакций, чем остальные члены группы. Как было отмечено ранее, все данные при сборе и обработке были деперсонализированы. При сборе комментариев была сделана надстройка для скрипта, позволяющая исключить комментарии модераторов сообщества, это позволило избежать искажения статистики.

В данном исследовании, за самых активных были

взяты те профили, которые оставили максимальное количество комментариев под записями в своей районной группе. На рисунке 5 результаты представлены в виде диаграммы, для каждого района отмечено количество комментариев в сообществе от выявленного активного пользователя.

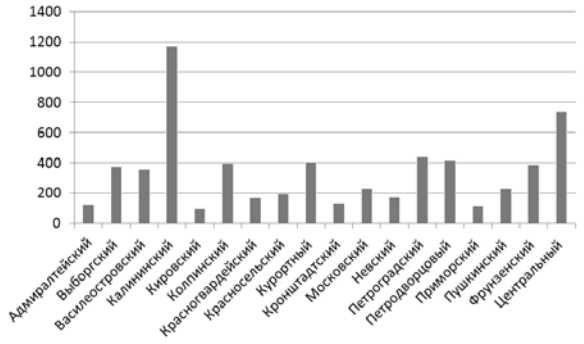

Рис 5. Общее количество комментариев активных пользователей в своем районе

В ходе анализа было выявлено 33 самых активных пользователей в районных сообществах.

На следующем этапе был проведен анализ профилей 33 самых активных участников сообществ. С помощью VK API была собрана общедоступная информация о профиле (в случае, если профиль открытого типа), а также произведен подсчет общего количества комментариев от пользователя, которые он оставлял в других сообществах и страницах социальной сети ВКонтакте. Дополнительно, оценивалась дата регистрации профиля и наличия реального человека на фотографии, данная информация использовалась для исключения рекламных ботов.

По данным исследования, среди активистов 39% составляют мужчины и 61% - женщины. Определить возраст удалось только у 59% активных пользователей. Как показывает статистика, в основном это люди старше 30, причем 12% составляют активисты старше 60 лет.

Сложнее обстояло дело с определением основных сфер занятости среди группы активных пользователей. Почти у 70% в изучаемой группе было невозможно определить сферу занятости. Однако среди пользователей, чья сфера занятости была указана в открытом доступе основную долю составляют работники сферы образования и науки.

Данные исследования свидетельствуют о проявлении активности выявленных лидеров мнений в виде комментариев к постам и на других площадках (сообществах). Эти факты позволили сделать вывод о существовании определенного социально-психологического типа лидеров мнений в районных сообществах. У собранных профилей активность начинается от 192 комментариев во всей социальной сети ВКонтакте, и достигает значения 1625 комментария, что можно считать высоким показателем активности. В расчете общего числа комментариев во всем ВКонтакте учитываются только записи из открытых сообществ. В закрытых сообществах, количество комментариев рассчитать невозможно.

IV. Выводы

В работе исследовались комментарии, которые оставляют подписчики под официальными записями, определено общее количество комментариев, выявлены наиболее активные комментаторы. На основе информации о самых активных участниках групп был проведен анализ социально-демографических и иных параметров, характеризующих активистов в городских сообществах в Петербурге.

В результате проведения исследования был отработан механизм выявления активных пользователей в районных сообщества Санкт-Петербурга в социальной сети ВКонтакте. Механизм был реализован на основе API ВКонтакте и позволяет собирать деперсонализированные наборы данных. На основе применения данного метода было выявлено существование в каждом сообществе ядра активных пользователей, которые оставляют больше всего реакций на посты, публикуемые в группах, а также развивают дискуссии.

Метод показал свою перспективность, и полученные данные могут быть использованы как исследователями, так и представителями государственных ведомств.

Дальнейшие направления исследования будут связаны с анализом текстов комментариев пользователей и использованием качественных психологических методов для более глубокого анализа профилей. Также планируется расширить масштаб исследования за счет включения в анализ данных из других социальных сетей, а затем их сравнительного анализа

В перспективе мы также планируем выполнить анализ и обработку фотографий пользователей с помощью технологии Microsoft Azure, текущее исследование посвящено только общедоступным основным характеристикам.

Библиография

[1] Mutz D., Young L. Communication and public opinion: Plus ça change? // Public Opinion Quarterlyю 2011. Vol. 75. No 5. P. 1,018–1,044. DOI: 10.1093/poq/nfr052

[2] Turcotte J., York C., Irving J., Scholl R.M., Pingree R.J. News recommendations from social media opinion leaders: Effects on media trust and information seeking // Journal of Computer-Mediated Communication. 2015. Vol. 20. No. 5. P. 520–535. DOI: 10.1111/jcc4.12127

[3] Katz E., Lazarsfeld P.F. Personal influence: The part played by people in the flow of mass communication. Glencoe, Ill.: Free Press. 1955.

[4] Nisbet M.C., Kotcher J.E. A two-step flow of influence? Opinion-leader campaigns on climate change // Science Communication. 2009. Vol. 30. No 3, P. 328–354. DOI: 10.1177/1075547008328797

[5] Bakshy E., Hofman J.M., Mason W.A., Watts D.J. Everyone's an influencer: Quantifying influence on Twitter // WSDM '11: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, 2011. P. 65–74. DOI: 10.1145/1935826.1935845

[6] Katz E. "The two-step flow of communication: An up-to-date report on an hypothesis // Public Opinion Quarterly. 1957. Vol. 21. No 1. P. 61–78. DOI: 10.1086/266687

[7] Cha M., Haddadi H., Benevenuto F., Gummadi K.P. Measuring user influence in Twitter: The million follower fallacy // ICWSM '10: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. 2010. P. 10–17.

[8] Diehl T., Weeks B.E., de Zúñiga H.G. Political persuasion on social media: Tracing direct and indirect effects of news use and social

interaction // New Media & Society. 2016. Vol. 18. No 9. P. 1,875–1,895. DOI:10.1177/1461444815616224

[9] Численность населения Санкт-Петербурга по состоянию на 1 января 2021 года. Федеральная служба государственной статистики https://petrostat.gks.ru/storage/mediabank/TCgYnq5o/Числ.СПб%20на%2001.01.2021.pdf (дата обращения: 07.11.2021).

Низомутдинов Борис Абдуллохонович, ведущий аналитик Центра технологий электронного правительства исследований Университета ИТМО, Санкт-Петербург (https://itmo.ru/), email: boris@itmo.ru, elibrary.ru: authorid=794641, ORCID: orcidID= 0000-0002-4090-9564

Видясова Людмила Александровна, кандидат социологических наук. Начальник отдела мониторинговых исследований Университета ИТМО, Санкт-Петербург (https://itmo.ru/), email: bershadskaya.lyudmila@gmail.com, elibrary.ru: authorid=642903, ORCID: orcidID= 0000-0002-8006-7066

Application of automated collection of information from social network communities to identify active users

B.A. Nizomutdinov, L.A. Vidiasova

***Abstract*— The article presents a method developed by the authors to determine the most active subscribers of communities in social networks. The object of study was 18 official groups of district administrations of St. Petersburg VKontakte. The paper describes the implementation of an algorithm for data collection and analysis. The study determined the total number of comments in each community and identified the most active subscribers who leave the most comments under posts. The analysis of socio-demographic characteristics of active users was carried out using automated tools for parsing social network data. All collected data has been depersonalized. The results of the study showed the existence in each community of a core of active users who leave the most reactions to posts published in groups. The conclusion is made about the great potential that can be extracted from the use of automated data collection from social networks.**

***Keywords*— social networks, social activists, active users, opinion leaders, parsing, social network analysis, VK API, social network parsing.**

REFERENCES

[1] Mutz D., Young L. Communication and public opinion: Plus ça change? // Public Opinion Quarterlyю 2011. Vol. 75. No 5. P. 1,018–1,044. DOI: 10.1093/poq/nfr052

[2] Turcotte J., York C., Irving J., Scholl R.M., Pingree R.J. News recommendations from social media opinion leaders: Effects on media trust and information seeking // Journal of Computer-Mediated Communication. 2015. Vol. 20. No. 5. P. 520–535. DOI: 10.1111/jcc4.12127

[3] Katz E., Lazarsfeld P.F. Personal influence: The part played by people in the flow of mass communication. Glencoe, Ill.: Free Press. 1955.

[4] Nisbet M.C., Kotcher J.E. A two-step flow of influence? Opinion-leader campaigns on climate change // Science Communication. 2009. Vol. 30. No 3, P. 328–354. DOI: 10.1177/1075547008328797

[5] Bakshy E., Hofman J.M., Mason W.A., Watts D.J. Everyone's an influencer: Quantifying influence on Twitter // WSDM '11: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, 2011. P. 65–74. DOI: 10.1145/1935826.1935845

[6] Katz E. "The two-step flow of communication: An up-to-date report on an hypothesis // Public Opinion Quarterly. 1957. Vol. 21. No 1. P. 61–78. DOI: 10.1086/266687

[7] Cha M., Haddadi H., Benevenuto F., Gummadi K.P. Measuring user influence in Twitter: The million follower fallacy // ICWSM '10: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. 2010. P. 10–17.

[8] Diehl T., Weeks B.E., de Zúñiga H.G. Political persuasion on social media: Tracing direct and indirect effects of news use and social interaction // New Media & Society. 2016. Vol. 18. No 9. P. 1,875–1,895. DOI:10.1177/1461444815616224

[9] The population of St. Petersburg as of January 1, 2021. Federal State Statistics Service https://petrostat.gks.ru/storage/mediabank/TCgYnq5o/Числ.СПб%20на%2001.01.2021.pdf (accessed: 07.11.2021).

Nizomutdinov Boris Abdullohonovich, leading analyst at the Center for E-Government Research Technologies at ITMO University, St. Petersburg (https://itmo.ru/), email: boris@itmo.ru, elibrary.ru: authorid=794641, ORCID: orcidID= 0000-0002-4090-9564

Vidiasova Lyudmila Alexandrovna, Candidate of Sociological Sciences. Head of the Monitoring Research Department at ITMO University, Санкт-Петербург (https://itmo.ru/), email: bershadskaya.lyudmila@gmail.com, elibrary.ru: authorid=642903, ORCID: orcidID= 0000-0002-8006-7066