

<https://doi.org/10.5281/zenodo.3894532>

УДК 316.37

Соина А.С., Ченгарь О.В., Шевченко В.И.

Соина Анастасия Сергеевна, ассистент кафедры «Теория и практика перевода», Севастопольский государственный университет, 299011, Россия, Республика Крым, Севастополь, ул. Университетская, 33. E-mail: anastasiasoina@gmail.com.

Ченгарь Ольга Васильевна, кандидат технических наук, доцент кафедры «Информационные технологии и компьютерные системы», Севастопольский государственный университет, 299011, Россия, Республика Крым, Севастополь, ул. Университетская, 33. E-mail: OVChengar@sevsu.ru.

Шевченко Виктория Игоревна, кандидат технических наук, доцент кафедры «Информационные технологии и компьютерные системы», Севастопольский государственный университет, 299011, Россия, Республика Крым, Севастополь, ул. Университетская, 33. E-mail: VIShevchenko@sevsu.ru.

Общество через призму больших данных социальных сетей: современный зарубежный опыт

Аннотация. В данной статье рассматриваются некоторые современные публикации зарубежных исследователей из Британии, Индии, Греции, Перу, Италии, Нигерии, направленные на изучение социальных сетей с помощью технологии больших данных. Сделан вывод о том, что технологию больших данных для анализа социальных сетей стали использовать относительно недавно. В настоящее время большие данные позволяют предотвращать теракты, отслеживать умонастроения людей, делать сервис более удобным для клиента. Основным источником информации является соцсеть Twitter. В качестве наиболее значимых вызовов современности исследователи указывают на ограничения в интерпретации изображений.

Ключевые слова: большие данные, социальные сети, Twitter, терроризм, уровень жизни, туризм, психология.

Soina A.S., Chengar O.V., Shevchenko V.I.

Soina Anastasiia Sergeevna, teaching assistant of «Translation Theory and Practice» Department, Sevastopol State University, 299011, Russia, Republic of Crimea, Sevastopol, Universitetskaya st., 33. E-mail: anastasiasoina@gmail.com.

Chengar Olga Vasilevna, Candidate of Technical Sciences, Associate Professor of the «Information Technologies and Computer Systems» Department, Sevastopol State University, 299011, Russia, Republic of Crimea, Sevastopol, Universitetskaya st., 33. E-mail: OVChengar@sevsu.ru.

Shevchenko Viktoriya Igorevna, Candidate of Technical Sciences, Associate Professor of the «Information Technologies and Computer Systems» Department, Sevastopol State University, 299011, Russia, Republic of Crimea, Sevastopol, Universitetskaya st., 33. E-mail: VIShevchenko@sevsu.ru.

Society through the perspective of big data and social networks: modern foreign experience

Abstract. This paper discusses some recent publications of foreign researchers from Britain, India, Greece, Peru, Italy, and Nigeria aimed at studying social networks using big data technology. It is con-

cluded that big data has recently been used for social network analysis. Currently, big data allows humanity to prevent terrorist attacks, follow people's way of thinking, and make service more appealing to clients. The main source of information is the social network Twitter. Also, the researchers consider limitations in the interpretation of images to be the most significant challenges of our time.

Key words: big data, social networks, Twitter, terrorism, standard of living, tourism, psychology.

Актуальность темы. В настоящее время технологии стремительно развиваются, прогресс идет все быстрее. Одной из новых технологий последних двух десятилетий стали большие данные, big data. Большие данные – это социально-экономический феномен, связанный с новой возможностью исследовать весь мировой объем данных для выполнения целого спектра задач. За рубежом в 2010 году начали появляться исследования, направленные исключительно на различные способы обработки больших данных, а с 2013 года технологию больших данных изучают в рамках ряда вузовских дисциплин. Другой, не менее важной новой технологией, а точнее, площадкой для взаимодействия людей, стали социальные сети, число пользователей которыми только растет. Активные пользователи ежедневно генерируют огромное количество данных, а современные ученые находят все новые способы использования этих данных. Таким образом, данная работа, направленная на изучение зарубежного опыта использования технологии больших данных социальных сетей видится нам актуальной.

Постановка проблемы. Технологию больших данных можно использовать в автомобилестроении, медицине, торговле, обеспечении порядка и безопасности, маркетинге и так далее. Неотъемлемой составляющей всех этих процессов является человек, мы с вами. Именно с помощью социальных сетей мы можем узнать больше о нас самих, о том, какие услуги мы хотели бы получить, какая опасность может нам угрожать и какой товар мы, возможно, хотели бы приобрести. Цель данной статьи: проанализировать современные зарубежные научные разработки

с точки зрения применения больших данных социальных сетей.

Анализ проблемы. В ходе исследования было просмотрено около 150 источников и отобрано 60, на английском, французском и немецком языках, преимущественно на таких сайтах, как <https://www.academia.edu/> и <https://www.researchgate.net/>. Отобранные публикации можно условно разделить на три категории: публикации о развитии общества в широком смысле, проблемах общества («гуманитарный» аспект), публикации о способах анализа больших данных («технический» аспект), публикации об исследовании какого-либо общественного явления с помощью технологии больших данных и социальных сетей, либо приложений (синтез «гуманитарного» и «технического» аспектов). Примечательно, что последних исследований – немного. Однако именно на них мы остановимся в рамках данной статьи.

Итак, широко известные зарубежные соцсети – это Facebook, Twitter и Instagram. В данной работе речь в основном пойдет о них, хотя будут встречаться и альтернативные площадки. Facebook имеет максимально широкий функционал: здесь можно создавать сообщества, делиться фотографиями, видеозаписями, ссылками, писать довольно объемные посты. Twitter обычно используется для коротких сообщений, изображений, видео, преимущественно формат общения – текстовый. Основной контент Instagram, напротив, – визуальный, площадка предлагает делиться фотографиями и видео, используя минимум текста.

Начнем с наиболее острой, на наш взгляд, проблемы современности – безопасности. Значительную угрозу современному обществу представляет терроризм. Не секрет, что для вербовки терро-

ристы используют различные соцсети и мессенджеры [1]. Среди зарубежных исследователей есть те, кто видит решение с помощью технологии больших данных. Авторы из, Федерального университета в Кашере, Нигерия, в работе «Аналитика социальных сетей: незаменимая антитеррористическая система для улучшения сбора разведанных в целях борьбы с терроризмом в Нигерии», 2018 (здесь и далее транслитерация личных имен, а также перевод заголовков исследования мои. – А.С.) настаивают на изменении принципа работы спецслужб [8]. По их мнению, анализ информации из YouTube, Twitter, Instagram, Facebook позволит получить данные о деятельности террористов или террористических групп, о членстве, мотивации и оперативных методах таких групп, а также блокировать их источники финансирования и препятствовать вербовке. Исследователи рекомендуют использовать как аналитику социальных сетей с открытым исходным кодом, такую как Gephi, Igraph, NetworkX, так и коммерческую аналитику, такую как I2Analyst, Sentinel Visualizer. В своей работе они предлагают издать новые законы, позволяющие службе безопасности Нигерии предсказывать теракты на основе анализа поведения местных граждан в социальных сетях.

Коллеги из Саудовской Аравии, А. Алзахрани, К.Б. Бажва, Т. Алгхамди, Х. Алдахави в исследовании «Борьба с терроризмом в социальных сетях при использовании больших данных», 2018, показывают конкретные способы вычленения необходимых данных из общего массива [4]. Как только объем данных будет уменьшен до приемлемого размера, методы интеллектуального анализа данных, включающие классификацию, кластеризацию, ассоциацию, последовательное обнаружение закономерностей и регрессию, могут быть более эффективно использованы для выявления признаков терроризма и извлечения связанной с вербовкой информации и стратегий различных групп террористов. Дальнейшие

действия подразумевают обработку естественного языка для построения и расширения набора данных. Анализ социальных сетей позволяет визуализировать членов группы в виде узлов (акторов) и их взаимосвязей. В отличие от других количественных стратегий, которые сосредоточены вокруг образа и полного исследования качеств акторов, составляющих исследуемую группу, анализ социальных сетей предполагает, что для понимания социального феномена необходимо определить и нарушить связь между акторами. Авторы считают, что этот подход поможет получить информацию о принятии решений, групповой динамике и результатах коллективных действий.

Еще одной проблемой современности является изучение поведенческой психологии. В качестве примера приведем работу исследователей Кембриджского университета, М. Косински, Д. Стиллвелл и Т. Грэпп «Предсказание личностных особенностей посредством цифрового следа», 2013 [9]. Исследователи продемонстрировали, что легкодоступные цифровые записи поведения, такие как лайки в Facebook, могут быть использованы для автоматического и точного предсказания целого ряда личностных характеристик, включая сексуальную ориентацию, этническую принадлежность, религиозные и политические взгляды, личностные черты, интеллект, счастье, употребление аддиктивных веществ, разлуку родителей, возраст и пол. Представленный анализ основан на сборе данных более чем 58 000 добровольцев, которые предоставили свои лайки в Facebook, подробные демографические профили, а для проверки результатов, – прошли несколько психометрических тестов. Предложенная исследователями модель правильно различает гомосексуалистов и гетеросексуалов в 88% случаев, афроамериканцев и кавказцев – в 95% случаев, а также демократов и республиканцев в 85% случаев. Точность предсказания близка к точности стандартного личностного теста, который проходили впоследствии испытуемые. Исследовате-

ли полагают, что прогнозирование предпочтений пользователей может быть использовано для улучшения многочисленных продуктов и услуг. Например, цифровые системы и устройства (интернет-магазины или автомобили) могут быть сконструированы таким образом, чтобы их поведение наилучшим образом соответствовало предполагаемому профилю каждого пользователя. Кроме того, можно повысить релевантность маркетинговых рекомендаций за счет добавления психологических аспектов в существующие модели пользователей. Например, реклама онлайн-страхования может подчеркивать безопасность при взаимодействии с эмоционально нестабильными (невротическими) пользователями, но подчеркивать потенциальные угрозы при работе с эмоционально стабильными личностями.

Итальянским исследователям удалось измерить уровень счастья итальянских пользователей Twitter. В своей работе «Как использовать большие данные из социальных сетей: субъективный индикатор благополучия через Twitter» (2017) они утверждают, что индекс качества жизни, применяемый официальной статистикой, имеет не так много общего с действительностью, так как указанные там показатели исключительно материальны: «Этот индекс охватывает только материальное качество жизни и является эталонным показателем объективного благополучия в Италии... Он не охватывает всех людей и содержит только данные за прошедший год, из-за чего показывает уже неактуальную информацию. Это серьезный недостаток, когда лица, принимающие решения, должны основывать свой выбор на такой информации. Различия между этими двумя измерениями (объективным и субъективным) ясно вытекают из сравнения», – считают исследователи [7]. В работе они приводят расчеты и карту Италии, где действительно заметна разница между объективными показателями удовлетворения и реальным положением вещей. Таким образом им удалось создать композитный

показатель благосостояния, который смешивает стабильную официальную статистику и изменчивые данные социальных сетей.

Разумеется, в медийной среде не обойтись без журналистов. Ряд исследователей университета Ньюкасл в 2019 году опубликовали работу под названием «Twitter, Трамп и изменение климата: вычисление общественного мнения в режиме онлайн с помощью больших данных» [6]. В ней ученые пытались понять, какие слова чаще всего ассоциируются с изменением климата в онлайн-дискурсе, а также как публичный дискурс в Twitter соотносится с дискурсом информированных ученых и журналистов, пишущих в блогах об изменении климата. «После того как мы собрали текстовый корпус из Twitter Stream API и четырех блогов, мы преобразовали весь текст в нижний регистр и удалили все артикли (например, a, the, and). Мы также тщательно следили за использованием терминов (например, преобразовали *climate change*, *clichange*, и *climatch* в одно-единственное *climatechange*). Наконец, мы загрузили весь предварительно обработанный текст в движок Facebook *fasttext engine* и сгенерировали наиболее похожие списки слов, относящиеся к сочетанию *climatechange*. Затем проанализировали результат и интерпретировали разницу между постами в Twitter и онлайн-блогах, которые были представлены пользователями и авторами, с точки зрения понимания явления *climatechange*». Исследователями было проанализировано около миллиона твитов и 4 блога за определенные промежутки времени и выполнено сопоставление.

К. Брантен и Ю. Пфеффер в работе «Контент-анализ Twitter», 2018, провели анализ настроения пользователей и классификацию текстов на основе машинного обучения [5]. Исследователи собрали необходимые твиты с помощью API Twitter – аббревиатура API расшифровывается как «интерфейс прикладного программирования», то есть прямое подключение к данным Twitter, к которым можно получить доступ с помощью программного

кода. Основная причина, по которой анализ твитов популярен во многих научных областях, заключается в том, что к данным можно легко и без затрат получить доступ через API Twitter. Авторам удалось создать облака слов с топ-100 словами из 5101 твитов, связанных с новостями, и 60168 твитов, не связанных с новостями. Более того, они сравнили настроение, показанное в твитах, связанных с новостями, с настроением, продемонстрированным твитами, не связанными с новостями. Основная концептуальная проблема, как подчеркивают исследователи, связанная с автоматизированным анализом контента, заключается в несоответствии между тем, что эти методы должны делать (анализировать настроения, находить темы) и тем, что они делают на самом деле (подсчитывают прилагательные, совпадающие слова). Автоматизированные методы анализа контента далеки от понимания текста. Вместо этого текст рассматривается как «гора слов». Написанный текст содержит в себе некий культурный пласт и обладает разной лингвистической «сложностью», и ничего из этого не может быть изучено с помощью методов, описанных в статье. Также существенной проблемой выступает то, что методы предварительной обработки текста и контент-анализа оптимизированы для текстов на английском языке. Текст на других языках с более сложными флексиями и словосочетаниями может потребовать вмешательства человека, например, для «очистки» текста и составления словаря, чтобы показать значимые результаты. Сведение сообщений пользователей к «горе слов» также игнорирует еще одну важную особенность общения в социальных сетях: изображения.

Исследователи из Колумбии и Перу в совместной работе «Метод сбора релевантных тем из Twitter с использованием больших данных», 2019, разработали способ вычленения контента путем отбора профилей пользователей, являющихся экспертами в области спорта и политики [12]. База эксперимента насчитывает 10 750 твитов, связанных со спортом, и

8758 твитов, связанных с политикой. В исследовании предлагается к рассмотрению метод, который может быть включен в любое приложение, требующее исследования по темам. После того как были выбраны начальные пользователи и темы, было замечено, что наиболее релевантные темы можно определить с помощью LDA. По словам ученых, большая часть полученных тем могла быть связана с газетными заголовками El País, выпущенными в те же дни, когда был применен этот метод. В перспективе ученые смогут применять индекс релевантности каждой темы и отслеживать его изменение относительно часового пояса, чтобы определить, когда тема достигла большей релевантности.

Ещё одной сферой применения больших данных является туризм. По словам молодого исследователя с острова Крит, И. Макридиса, это может принести выгоду предприятиям и местным органам власти. В своем исследовании «Аналитика больших данных и поиск знаний с помощью социальных сетей с геолокацией» автор разработал приложение для получения знаний с помощью фактического пользовательского контента в социальных сетях, основанных на данных о местоположении, например, Twitter, Foursquare, Instagram и Flickr [10]. Цель исследования заключалась в анализе данных о поведении посетителей, их впечатлениях и предпочтениях в отношении туристических направлений, а также в предложении решений для местных органов власти и бизнеса в отношении более эффективного продвижения туристических направлений, улучшения существующих объектов, а также привлечения интереса большего числа потенциальных посетителей. Исследование проводилось для городов Ираклион и Ханья на острове Крит. Результаты анализа оформлены с помощью трех круговых диаграмм, чтобы представить впечатления пользователей в двух городах (положительном, отрицательном или нейтральном), которые чаще всего упоминались в их сообщениях (например, места, люди, события, орга-

низации и т.д.). Происхождение пользователей определяли по используемому ими языку. Далее исследователь извлек хэштеги из всех сообщений (как текстовых, так и медийных), написав пользовательский алгоритм для обнаружения тем, которые чаще всего фигурируют в сообщениях пользователей. Полученную информацию он представил в виде облака слов. Чтобы определить наиболее влиятельных пользователей в двух городах, которые своими сообщениями могут влиять на мнение других людей, была создана линейная диаграмма, показывающая социальную активность в день для каждого интересующего региона. Также в работе представлен рейтинг туристических мест.

Выводы. Статья не претендует на исчерпывающее освещение направлений современных зарубежных исследований в отношении социальных сетей. Мы лишь делаем попытку поиска неких общих явлений, характерных для сферы в целом. Синописис ряда исследований позволил выявить следующее:

1. Специалисты начали изучать социальные сети с применением технологии больших данных относительно недавно. Зарубежные исследователи обращают свое внимание в большей степени на Twitter, чем другие на сети. Возможно, это связано с меньшими способами самовыражения для пользователей, по срав-

нению, например, с Facebook, и это существенно упрощает задачу по скачиванию и обработке данных.

2. Наиболее популярными направлениями современных зарубежных исследований являются поиск террористов, психологические аспекты поведения людей в связи с инфоповодами, удовлетворенность жизнью, туризм. Данные социальных сетей способны давать картину происходящего в конкретный момент времени.

3. Наибольшим вызовом для современных исследователей является обработка изображений.

Перспектива исследования. В дальнейшем возможно рассмотреть причины возникновения конфликтов в обществе и применение технологии больших данных для выявления пользователей, разжигающих конфликт, что, по нашим данным, на сегодняшний момент никем не проводилось.

Отметим, что в России в 2017 году на базе Томского государственного университета был организован Университетский консорциум исследователей больших данных, куда на данный момент входят 27 университетов, в том числе и Севастопольский государственный университет [2]. Данная работа выполнена при финансовой поддержке Севастопольского государственного университета в рамках внутреннего гранта № 31/06-31.

СПИСОК ЛИТЕРАТУРЫ

1. Байназаров Н. Как террористы используют соцсети для вербовки // Rusbase. URL: <https://rb.ru/story/isis-uses-social-media/>
2. О нас // Университетский консорциум исследователей больших данных: сайт. URL: <https://opendata.university/about>
3. Academia: platform for academics to share research papers // Academia. URL: <https://www.academia.edu/>
4. Alzahrani A., Bajwa K. B., Alghamdi T., Aldahawi H. Countering Terrorism On Social Media Using Big Data // Academia. URL: https://www.academia.edu/37819783/COUNTERING_TERRORISM_ON_SOCIAL_MEDIA_USING_BIG_DATA
5. Brantner, C., Pfeffer, J. Content analysis of Twitter: Big data, big studies. In S. A. Eldridge II & B. Franklin (Eds.), *The Routledge Handbook to Developments in Digital Journalism Studies*. Abingdon: Routledge, 2018. URL: https://www.researchgate.net/publication/322753687_Content_analysis_of_Twitter_Big_data_big_studies_In_S_A_Eldridge_II_B_Franklin_B_Eds_The_Routledge_Handbook_to_Developments_in_Digital_Journalism_Studies_Abingdon_Routledge

6. Bush D, Xing J, Sieber R. Twitter, Trump, and Climate Change: Examining the Big Data of Public Opinion Online // American Association of Geographers, 2019. URL: https://www.academia.edu/40054449/Twitter_Trump_and_Climate_Change_Examining_the_Big_Data_of_Public_Opinion_Online
7. Iacus S.M., Porro G., Salini S., Siletti E. How to Exploit Big Data from Social Networks: a Subjective Well-being Indicator via Twitter // SIS 2017. Statistics and Data Science: new challenges, new generations: scientific article. URL: https://www.academia.edu/33994733/How_to_Exploit_Big_Data_from_Social_Networks_a_Subjective_Well-being_Indicator_via_Twitter
8. Jibril L. M., Ali G.A. Social Media Analytics: Indispensable Counterterrorism Framework to improve Intelligence Gathering towards Combating Terrorism in Nigeria // 3rd Big Data Analytics & Innovation Conference, National Defence College, Abuja. Volume: Vol. 1. 2016. URL: https://www.researchgate.net/publication/309537533_Social_Media_Analytics_Indispensable_Counterterrorism_Framework_to_improve_Intelligence_Gathering_towards_Combating_Terrorism_in_Nigeria
9. Kosinski M., Stillwell D., Graepel Th. Private traits and attributes are predictable from digital records of human behavior // Proceedings of the National Academy of Sciences of the United States of America. 2013. Vol. 110, № 15. P. 5802–5805.
10. Makridis J. Big Data Analytics and Knowledge Discovery through Location Based Social Networks // Technological Educational Institute of Crete, 2018. URL: https://www.academia.edu/40223892/Big_Data_Analytics_and_Knowledge_Discovery_through_Location_Based_Social_Networks_LBSN_MSc_Thesis
11. Research Gate: platform for academics to share research papers // ResearchGate GmbH. URL: <https://www.researchgate.net/>
12. Silva J, Naveda A.S., Suarez R.G. Method for Collecting Relevant Topics from Twitter supported by Big Data // Journal of physics, 2020. URL: https://www.academia.edu/42689289/Method_for_Collecting_Relevant_Topics_from_Twitter_supported_by_Big_Data

REFERENCES (TRANSLITERATED)

1. Bajnazarov N. Kak terroristy ispol'zujut socseti dlja verbovki // Rusbase. URL: <https://rb.ru/story/isis-uses-social-media/>
2. nas // Universitetskij konsorcium issledovatelej bol'shih dannyh: sajt. URL: <https://opendata.university/about>
3. Academia: platform for academics to share research papers // Academia. URL: <https://www.academia.edu/>
4. Alzahrani A., Bajwa K. B., Alghamdi T., Aldahawi H. Countering Terrorism On Social Media Using Big Data // Academia. URL: https://www.academia.edu/37819783/COUNTERING_TERRORISM_ON_SOCIAL_MEDIA_USING_BIG_DATA
5. Brantner, C., Pfeffer, J. Content analysis of Twitter: Big data, big studies. In S. A. Eldridge II & B. Franklin (Eds.), The Routledge Handbook to Developments in Digital Journalism Studies. Abingdon: Routledge, 2018. URL: https://www.researchgate.net/publication/322753687_Content_analysis_of_Twitter_Big_data_big_studies_In_S_A_Eldridge_II_B_Franklin_B_Eds_The_Routledge_Handbook_to_Developments_in_Digital_Journalism_Studies_Abingdon_Routledge
6. Bush D, Xing J, Sieber R. Twitter, Trump, and Climate Change: Examining the Big Data of Public Opinion Online // American Association of Geographers, 2019. URL: https://www.academia.edu/40054449/Twitter_Trump_and_Climate_Change_Examining_the_Big_Data_of_Public_Opinion_Online
7. Iacus S.M., Porro G., Salini S., Siletti E. How to Exploit Big Data from Social Networks: a Subjective Well-being Indicator via Twitter // SIS 2017. Statistics and Data Science: new challenges, new generations: scientific article. URL: https://www.academia.edu/33994733/How_to_Exploit_Big_Data_from_Social_Networks_a_Subjective_Well-being_Indicator_via_Twitter

8. Jibril L. M., Ali G.A. Social Media Analytics: Indispensable Counterterrorism Framework to improve Intelligence Gathering towards Combating Terrorism in Nigeria // 3rd Big Data Analytics & Innovation Conference, National Defence College, Abuja. Volume: Vol. 1. 2016. URL: https://www.researchgate.net/publication/309537533_Social_Media_Analytics_Indispensable_Counterterrorism_Framework_to_improve_Intelligence_Gathering_towards_Combating_Terrorism_in_Nigeria
9. Kosinski M., Stillwell D., Graepel Th. Private traits and attributes are predictable from digital records of human behavior // Proceedings of the National Academy of Sciences of the United States of America. 2013. Vol. 110, № 15. P. 5802–5805.
10. Makridis J. Big Data Analytics and Knowledge Discovery through Location Based Social Networks // Technological Educational Institute of Crete, 2018. URL: https://www.academia.edu/40223892/Big_Data_Analytics_and_Knowledge_Discovery_through_Location_Based_Social_Networks_LBSN_MSc_Thesis
11. Research Gate: platform for academics to share research papers // ResearchGate GmbH. URL: <https://www.researchgate.net/>
12. Silva J, Naveda A.S., Suarez R.G. Method for Collecting Relevant Topics from Twitter supported by Big Data // Journal of physics, 2020. URL: https://www.academia.edu/42689289/Method_for_Collecting_Relevant_Topics_from_Twitter_supported_by_Big_Data

Поступила в редакцию 30.05.2020.
Принята к публикации 01.06.2020.

Для цитирования:

Соина А.С., Ченгарь О.В., Шевченко В.И. Общество через призму больших данных социальных сетей: современный зарубежный опыт // Гуманитарный научный вестник. 2020. №5. С. 220-227. URL: <http://naukavestnik.ru/doc/2020/05/Soina.pdf>