# Сбор данных в социальной сети Twitter

События реального мира оказывают влияние на онлайн-системы, и следы, оставленные пользователями в таких системах, использовались для автоматического раннего обнаружения событий. Действительно, ряд известных исследований показал, что грипп [70], уровень безработицы [71] или продажи автомобилей [72] могут быть спрогнозированы на основе поисковых запросов в Интернете.

Ранее, но аналогичным образом, сообщения в блогах коррелировали с продажами книг [73] или валовыми доходами от фильмов [74]. Следует отметить, что, хотя запросы не поддаются сентимент-анализу, сообщения в блогах поддаются, и, следовательно, применяя сентимент-анализ к блогам, можно обнаружить более сильные корреляции – например в исследованиях Мишне [74], [75].

По сравнению с блогами микроблоги - явление недавнее; тем не менее, они привлекли большое внимание на том основании, что могут помочь следить за настроениями общества. Действительно, краткость микропостов в сочетании с тем фактом, что одна услуга (Twitter), являющийся стандартом де-факто для такого рода публикаций, превратил эти данные в новый фаворит для исследователей.

Поэтому, основываясь на изобилии данных, предлагаемых Twitter, и в свете превосходных результатов, полученных в прошлом с помощью интеллектуального анализа журналов запросов и блогов, неудивительно, что был проведен ряд исследований по прогнозированию как настоящих, так и будущих событий на основе твитов.

Например, Асур и Губерман [76] использовали данные Twitter для прогнозирования кассовых сборов фильмов, О`Коннор и Баласубраманьян [77] довольно успешно сопоставили твиты с несколькими временными рядами общественного мнения, Тумасджан, Спрингер и Сенднер [78] утверждали, что предсказали исход выборов в Германии, а Боллен, Мао и Дзенг [79] успешно прогнозировали фондовый рынок.

Поверхностный взгляд на этот растущий объем литературы может свидетельствовать о том, что данные Twitter обладают впечатляющей – и универсальной – прогностической силой. Однако это было поставлено под сомнение. Например, исследование Вонга, Сена и Чьянга [80] вызвало некоторые сомнения в предсказуемости кассовых сборов на основе данных Twitter; и аналогичным образом Джунгер и Шоен [81] опровергли основной тезис Тумасджана, Спрингера и Сенднера [78], а именно, что твиты и голоса сильно коррелируют.

Прежде чем выяснить, кто прав, следует признать сильную предвзятость, пронизывающую исследования: тенденцию исследователей сообщать о положительных результатах, не сообщая об отрицательных. Эта предвзятость, так называемый эффект «картотечного ящика», который описывал Фанелли [82], чрезвычайно вредна.

Во-первых, это заставляет предположить, что опубликованные (положительные) результаты являются нормой и, следовательно, их следует ожидать в будущем.

Во-вторых, это затрудняет публикацию отрицательных результатов, поскольку бремя доказывания лежит на тех, кто критикует преобладающую (позитивную) точку зрения.

Это далеко нетривиальное доказательство или опровержение большинства вышеупомянутых методов прогнозирования. По одной из причин большинство из них, а именно те, которые связаны с кассовыми сборами, фондовым рынком, эпидемиями или опросами общественного мнения, не предсказывают событие, а показывают корреляцию между двумя различными временными рядами: один создан на основе данных Twitter, а другой получен из «оффлайн» мира. В этом отношении такие методы являются моделями и, следовательно, они могут быть более или менее точными и, в свою очередь, более или менее полезными.

Но одним из главных подготовительных этапов прогнозирования – это сбор данных, т.е. тех самых твитов, которые будут рассматриваться и, исходя из которых будет выводиться окончательный результат прогнозирования.

Ниже будет проведен метаанализ существующих инструментов сбора данных и их сравнительный анализ.

## Способы сбора данных в Twitter

Помимо того, что Twitter является универсальной коммуникационной платформой для пользователей по всему миру, он также является отличным источником актуальной информации. Данные, извлеченные из Twitter, используются исследователями с разным опытом работы для ответов на множество вопросов, начиная от простой информации о конкретных пользователях или событиях на сложные запросы. В некоторых исследованиях рассматриваются отдельные лица или небольшие сообщества, в то время как другие требуют больших объемов информации, собранной в течение длительных периодов времени. В зависимости от целей для сбора данных могут использоваться различные инструменты — от веб-сервисов аналитики, которые объединяют сбор, анализ и визуализацию, до непосредственного интеллектуального анализа API Twitter и интерпретации данных с использованием специального пакета статистики. Сбор данных в рамках проекта, будь то непосредственно через API или с помощью специального программного обеспечения, остается одним из самых сложных аспектов исследований на основе Twitter. Хотя технические и методологические требования на первый взгляд могут показаться сложными, глубокое знание инструментов и видов данных, доступных с их помощью, может решить многие общие проблемы. Потому стоит рассмотреть различные методы и их соответствующих преимущества и ограничения. Сначала будет рассмотрен сбор данных через API Twitter, как напрямую, так и с использованием набора программных пакетов, а затем рассмотрен вопрос о том, как интегрировать данные Twitter в общие проекты социальных научных исследований.

## The Twitter API`s

Вместо того, чтобы предлагать единый API, исследователям, желающим запросить сервис, доступны три различных интерфейса данных Twitter: Streaming API, REST API и Search API. За немногими исключениями, совокупность исследований, подготовленных к настоящему моменту, опиралась на сбор данных из одного из этих трех источников.

### The Streaming API

The Streaming API, вероятно, является наиболее широко используемым источником данных для исследований Twitter. Как правило, крупномасштабный количественный анализ данных Twitter основан на необработанных данных, собранных с помощью этого источника [83], [84]. Стоит отметить, что Streaming API - крайне неортодоксальный вид ресурса по сравнению с тем, как функционирует большинство других API. В более традиционных конфигурациях API основан на «извлечении»: исследователь запрашивает страницу данных с сервера, запрашивая URL, после чего сервер возвращает запрошенную страницу. The Streaming API, однако, основан на «push», то есть данные постоянно поступают с запрошенного URL (конечной точки), и исследователь должен разработать или использовать инструменты, которые поддерживают постоянное соединение с этим потоком данных при одновременной их обработке. Этот поток данных предоставляется исключительно в виде живого опроса, что означает, что в тот момент, когда твит публикуется в Twitter, он становится доступным. Поскольку потоковые данные предоставляются в виде системы опроса в режиме реального времени, не предназначенной для исторического анализа, исследование, проводимое с диахронической точки зрения, намного сложнее, чем с помощью традиционной системы сбора данных, так как исследователь, по сути, должен работать по расписанию Twitter. При изучении форм относительно спонтанной организации, таких как движения «Arab Spring», «Occupy» и «Indignados», сбор данных особенно затруднен, поскольку событие может быть признано таковым только задним числом [85], [86], [87]. Исследования запланированных событий, таких как выборы, требуют от исследователя добросовестности в создании потока для сбора данных, в идеале задолго до события, чтобы составить аналитически полезный корпус. Главной задачей академического сообщества в будущем будет проведение сложных и многогранных анализов, несмотря на такие ограничения вместо того, чтобы адаптировать исследовательские вопросы к доступности данных.

#### The Streaming API: представительная выборка

К счастью, необходимость сбора данных в реальном времени не применима одинаково ко всем исследовательским контекстам, и большая часть исследований в Twitter на сегодняшний день задает более общие вопросы, например, фокусируясь на макроскопических структурных свойствах платформы [88] или описание того, как пользователи концептуализируют свои коммуникативные практики [89]. Хотя Twitter используется во многих странах и на многих языках, сообщества пользователей существенно различаются по своему размеру, составу и привычкам использования. Исследователи должны осознавать кажущиеся незначительными детали, которые могут быть отражены в данных, например, всплески использования в течение дня или колебания активности в выходные дни по сравнению с рабочими днями. Сбор данных в течение длительных периодов времени всегда предпочтительнее, когда это возможно, даже если не все собранные данные используются в анализе.

Многие активные пользователи не пишут в твиттере ежедневно или, возможно, даже еженедельно, в то время как другие очень активны и соответственно искажают репрезентативность выборки. Наконец, не все количественные исследования Twitter основаны на содержании твитов: в таких работах, как «Измерение влияния пользователей в Twitter» [90] используются доступные данные социального графика и полностью фокусируются на отношениях между подписчиками, а не на содержании сообщений.

#### The Streaming API: ограничения пропускной способности

The Streaming API поставляется в трех диапазонах: «spritzer», «gardenhose» и «firehose», которые доставляют до 1%, 10% и 100% всех твитов, размещенных в системе, соответственно. По умолчанию любой обычной учетной записи пользователя в Twitter предоставляется доступ «spritzer» к системе, чего часто бывает достаточно для исследовательских целей. «Gardenhose» иногда предоставляется пользователям с обоснованными и убедительными причинами для расширения доступа, а «firehose» доступен только как компонент «деловых отношений» с Twitter напрямую или через авторизованных перепродавцов [91]. Для «spritzer» и «gardenhose» ограничение в процентах вступает в силу только тогда, когда более соответствующего процента всех твитов соответствуют условиям, размещенным в потоке. Если, например, исследователь собирает данные для небольшой конференции, доступа «spritzer» будет достаточно, чтобы зафиксировать каждый твит, опубликованный под хэштегом конференции, поскольку только в самых крайних случаях твиты о таком событии будут превышать 1% всего трафика на платформе.

Следует отметить, что согласно документации Twitter, соответствующие размеры выборки основаны на совокупности всей информации, размещенной в Twitter в потоковом режиме, а не на подмножестве твитов, к которым применяется определенный критерий фильтрации (например, хэштег или ключевое слово). Если есть сомнения, был ли захвачен весь желаемый контент, исследователи должны проверить, возвращает ли их запрос количество результатов, близкое к 1% или 10% от текущей общей пропускной способности API. Служба потоковой передачи также возвращает сообщения о статусе, указывающие, сколько твитов было пропущено, если ограничение вступило в силу, уведомляя исследователя об общем количестве твитов, пропущенных с начала опроса.

#### The Streaming API: параметры и заключение

Помимо трех вариантов пропускной способности, Streaming API предлагает два различных метода, «sample» и «filter», в качестве точек доступа к данным. Выборка просто предоставляет до 1% или 10% всех твитов, выбранных случайным образом. Хотя данные никогда не были независимо проверены как случайные, обычно предполагается, что они имеют достаточную степень случайности. Важно отметить, что две пробы, взятые в один и тот же момент времени, идентичны, что делает возможной воспроизводимость результатов [92],[93]. Внутри метода «filter» параметры «track», «follow» и «locations» можно использовать для выбора конкретных результатов из потока.

«Track» позволяет исследователям искать несколько терминов, разделенных запятыми, для отправки в потоковом запросе Twitter в качестве опции. Когда Twitter получает запрос, он возвращает исследователю только твиты, содержащие эти слова, разделенные символами, не являющимися словами. Во всех случаях, за исключением очень немногих, результат не превысит 1% от общего трафика на платформе, что делает этот метод в сочетании с параметром «track» удобным способом составления корпуса на основе ключевых слов или хэштегов.

«Follow» возвращает только твиты от набора пользователей, представленных их коллективными ID, разделенными запятыми. В настоящее время этот параметр позволяет собирать данные до 5000 учетных записей. Исследователи, намеревающиеся изучать конкретные сообщества пользователей, могут счесть этот метод особенно полезным для исследования известных групп людей в течение длительных периодов времени.

«Locations» обеспечивает идеальную точку доступа для исследователей, заинтересованных в географически ограниченных исследованиях. Большое количество пользователей отмечают свои твиты «геотегами» — то есть к твиту добавляется дополнительный объект метаданных, указывающий на его географическое происхождение. Эти твиты с «геотегами» представлены в виде точек или пар координат, которые указывают точное местоположение, и многоугольников или прямоугольников, нарисованных четырьмя парами точек, которые могут быть размером с городской парк или с провинцию. Вполне вероятно, что в будущем доля аналогичных твитов увеличится, что делает их весьма привлекательным инструментом для различных видов географически ограниченных исследований.

### The REST API

API REST (REpresentational State Transfer) предоставляет набор методов для взаимодействия с данными, который принципиально отличается от The Streaming API, использующего более традиционную модель извлечения. В общей сложности в REST API доступно более сотни активных методов, лишь немногие, из которых были изучены в исследовательских целях. Используя комбинацию методов, с помощью этой системы можно собрать данные социального графика группы пользователей, то есть информацию о том, кто за кем следит, и другие данные, помимо непосредственного содержания отдельных твитов. В частности, для интересующего пользователя два метода (followers/ids и friends/ids) могут возвращать списки других пользователей, которые подписаны на интересующего пользователя или за которыми следует пользователь, с количеством ID`s до 5000 на запрос. Другие полезные методы в REST API предоставляют доступ к актуальным темам, позволяют выполнять пакетный поиск пользователей с группами идентификаторов пользователей и, как правило, выполняют функции, которые полезны в сочетании с информацией, которая может быть собрана через Streaming API.

#### REST API: ограничение скорости обработки запросов

REST API имеет одно серьезное ограничение: это ресурс с ограниченной скоростью. Подобно тому, как уровни доступа «spritzer» и «gardenhose» к Streaming API искусственно ограничены только частью трафика, ограничение скорости установлено в основном для обеспечения разумных ожиданий трафика для инфраструктуры Twitter. Это очень затрудняет для исследователей своевременный сбор желаемых данных, особенно в случае запросов на REST. В прошлом компьютер, не прошедший проверку подлинности, мог выполнять 150 запросов в час. Когда любая учетная запись входила в Twitter с помощью открытой аутентификации (OAuth), это ограничение скорости увеличивалось до 350 запросов в час. Для использования методов followers/ids или friends/ids это означало, что в лучшем случае только 150 (или 350, при входе в систему) пользователей могли обрабатываться для каждого метода в час. Однако позже эти ограничения будут дополнительно снижены примерно до 60 запросов в час, и будут выполняться только запросы OAuth к API [94]. Если, например, исследователь собирает информацию о пользователе, у которого 563 друга и 178 подписчиков, можно было бы собрать всех друзей и подписчиков с помощью одного запроса к каждому из этих методов, общей стоимостью двух запросов в одночасовом окне.

Если у этого пользователя 5630 друзей и 1780 подписчиков, количество запросов увеличится до трех (поскольку friends/ids могут возвращать не более 5000 учетных записей на страницу данных). Ранее исследователи могли подавать заявки на белые списки на основе IP-адресов и учетных записей, которые, в случае их предоставления, увеличивали это ограничение до 20 000 запросов в час. С тех пор Twitter прекратил эту практику и не публикует новые белые списки. Бойд и Кроуфорд [95] выдвинули убедительный аргумент в пользу искусственного разделения на классы, которое создается с помощью этого различия между учетными записями с высокой пропускной способностью и всеми остальными. Хотя существуют способы обойти ограничения — например, путем создания больших сетей компьютеров, которые собирают данные в тандеме, — такая практика активно отслеживается Twitter, и нарушители наказываются внесением их учетных записей в черный список.

### The Search API

Ранние анализы Twitter, такие как работа Гаффни [96] о выборах в Иране 2009 года, были в значительной степени основаны на The Search API. Первоначально это была единственная точка доступа для поиска твитов, в которых упоминались хэштеги, и поэтому широко использовалась для исследований, основанных на событиях. Подобно REST API, это ресурс, основанный на извлечении, повторяет функциональность функции поиска Twitter.

Хотя теоретически некоторый исторический сбор данных все еще возможен с помощью The Search API, на практике его полезность сильно ограничена. Данные свободно удаляются из поисковой системы в течение недели после публикации, и нет достоверной информации о их полноте. Twitter активно не поощряет использование The Search API и планирует прекратить его в ближайшем будущем, поскольку он дорогостоящий в обслуживании и никогда не предназначался для распространения данных с высокой пропускной способностью в режиме реального времени.

## Инструменты поиска

Хотя практически весь доступ к данным Twitter осуществляется через один из API (и часто с помощью комбинации нескольких методов API), существует ряд инструментов для упрощения этого процесса. Вместо того, чтобы напрямую вызывать API, исследователи могут использовать их, чтобы указать, какие данные они хотят собрать. Клиентские программы, такие The Archivist или TAGS, поставляются с ограничением, заключающимся в том, что они должны регулярно запускаться с компьютера пользователя для сбора данных. Напротив, серверные методы сбора данных, такие как yourTwapperKeeper и Twitter Database Server, работают круглосуточно, собирая данные всякий раз, когда они становятся доступными. Этот процесс еще больше упрощается с помощью веб-сервисов, таких как 140kit, которые предоставляют более глубокие аналитические возможности, чем сервисы, не предназначенные для исследований, но в то же время ограничивают доступ к твитам для загрузки в соответствии с Условиями предоставления услуг Twitter, которые запрещают повторную публикацию полных твитов без согласия компании. Наконец, реселлеры данных, такие как Gnip и DataSift, предоставляют обширные исторические данные без проблем со сбором, но они платные.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Инструмент | Требуется хостинг? | Требуется программировать? | Предоставляет необработанные данные? | Предоставляет аналитику? | Платный сервис? |
| 140kit | Нет | Нет | Нет | Да | Нет |
| 140kit Source Code | Да | Да | Да | Да | Нет |
| yourTwapper  Keeper | Да | Да | Да | Нет | Нет |
| The Archivist | Нет | Нет | Нет | Да | Нет |
| TAGS | Нет | Нет | Да | Да | Нет |
| Twitter Database Server | Да | Да | Да | Нет | Нет |
| Gnip | Нет | Нет | Да | Да | Да |
| DataSift | Нет | Нет | Да | Да | Да |

Таблица 1 – ПО для сбора данных в Twitter

### The Archivist

The Archivist - бесплатное настольное приложение с открытым исходным кодом, которое работает в Windows на базе Microsoft .NET framework и является одним из самых простых инструментов для сохранения и анализа твитов. В отличие от большинства других доступных инструментов, The Archivist не требует веб-сервера. Каждый экземпляр может собирать твиты, содержащие определенное ключевое слово или хэштег, полученные с помощью The Search API. Использование The Search API делает The Archivist подверженным его ограничениям, и проблема, вероятно, станет более серьезной в будущем.

Извлечение больших объемов информации или исторических данных, как правило, невозможно, и для непрерывного извлечения собирающая машина должна работать постоянно. The Archivist рекомендуется только для небольших коллекций твитов, которые можно вручную проверить на согласованность, таких как небольшие архивы хэштегов и отдельные потоки пользователей. Для любого исследования, требующего надежной выборки, которую нельзя проверить вручную, не рекомендуется использовать The Archivist или любое другое настольное программное обеспечение, поскольку проблемы с задержкой, пропускной способностью и стабильностью, вероятно, повлияют на качество выборки.

### TAGS

The Twitter Archiving Google Spreadsheet (TAGS) - веб-скрипт, который можно использовать для сбора твитов в облаке. Запущенный под управлением Google Spreadsheets, TAGS способен извлекать данные Twitter через REST API и, следовательно, подвержен ограничению скорости, особенно при использовании без аутентификации в Twitter. Он обходит некоторые ограничения The Archivist, будучи размещенным на хостинге, и в то же время не требует от пользователей запуска собственного сервера. Подобно The Archivist, TAGS выполняет ряд статистических операций с извлеченными данными, облегчая анализ. Хотя нет необходимости устанавливать какое-либо ПО, TAGS имеет интерфейс, который менее интуитивно понятен, чем у The Archivist, и требует наличия учетной записи Google.

### YOURTWAPPERKEEPER

YourTwapperKeeper (YTK) - один из самых популярных инструментов, доступных исследователям, желающим упростить процесс извлечения данных. Он использует как The Streaming API Twitter, так и The Search API для сбора твитов, соответствующих заданному термину. Чтобы запустить код, исследователь должен использовать активное PHP-соединение и иметь возможность запускать пару скриптов - скрипты «stream» (запросы The Streaming API) и «crawl» (запросы The Search API). Кроме того, для хранения собранной информации требуется база данных MySQL.

YTK (как и его предшественник TwapperKeeper) использовался в ряде исследовательских проектов [97],[98] и часто упоминается как подходящий способ сбора данных [92]. Однако у использования YTK есть некоторые недостатки. Главны: он захватывает лишь небольшую часть диапазона метаданных, доступных в настоящее время через The Streaming API с каждым твитом. Из данных, которые предоставляет Twitter, YTK собирает только текст твита и несколько основных атрибутов метаданных, которые затем сохраняются в одной таблице, вместо того чтобы сохранять всю доступную информацию и выполнять предварительную обработку для облегчения анализа. В результате вопросы, связанные со ссылками, географическими местами, и менее ортодоксальные вопросы, сосредоточенные на конкретных атрибутах метаданных, не могут быть легко исследованы, если исследователь не обработает данные постобработкой или не изменит исходный код своей установки.

### TWITTER DATABASE SERVER

Twitter Database Server (TDS) - еще одно серверное решение для сбора данных о хэштегах. Как и YTK, он основан на PHP и сервере базы данных MySQL. Он предоставляет только абсолютно минимальный интерфейс для просмотра захваченных данных, вместо этого полагаясь на способность пользователя взаимодействовать с таблицами MySQL, которые содержат захваченные твиты, через командную строку или с помощью сторонней утилиты, такой как популярный phpMyAdmin. Хотя запуск и взаимодействие с TDS несколько сложнее, чем с YTK, он, потребляет меньше вычислительных ресурсов и захватывает дополнительные поля, такие как разрешенные URL-адреса, не предоставляемые YTK.

### 140KIT

140kit - веб-инструмент для анализа данных Twitter. В отличие от других упомянутых сервисов, 140kit гарантирует сбор всех полей метаданных. В его онлайн-версии загрузка необработанных данных запрещена из-за ограничений условий предоставления услуг, налагаемых Twitter на общедоступные наборы данных. Программное обеспечение доступно как на размещенной платформе, так и в виде автономного пакета, аналогичного yourTwapperKeeper, хотя язык Ruby, на котором написана программа, менее распространен, чем PHP, и может быть сложнее в реализации, в зависимости от типа доступа к веб-серверу. Как и у YTK и TDS, для потоковой передачи данных также требуется компьютер для постоянного запуска программного обеспечения. Кроме того, он поддерживает только The Streaming API, хотя исследователи могут расширить функциональность, если потребуется. В отличие от большинства других аналитических инструментов, 140kit создан специально для исследователей в качестве целевой аудитории и, следовательно, может быть более подходящим для ответов на вопросы вне коммерческих контекстов.

### GNIP и DATASIFT

Gnip - один из лучших инструментов, доступных исследователям с точки зрения качества данных, но он стоит дорого. Компания собирает данные через firehose access в Twitter и перепродает эти данные как исследовательскому сообществу, так и частному бизнесу. Доступность исторических данных делает сервис потенциально актуальным, но значительные затраты на сервис и отсутствие полных метаданных (по аналогии с YTK) могут удерживать исследователей от его использования.

Подобно Gnip, DataSift предоставляет большую часть той же функциональности. Также, как и в Gnip, некоторые поля метаданных, которые могут представлять интерес для исследователя, опущены, хотя он также добавляет другие примечательные точки данных (интеграция с оценками Klout). Недостатком обеих служб является то, что, ориентируясь на предприятия, они предоставляют не всю информацию, которая может иметь отношение к исследованиям.

## Ограничения

Ограничения социальных научных исследований, основанных на данных Twitter, происходят от ограничений, которые влияют на исследовательские проекты на разных уровнях. Как и в случае с любой другой методологией, не все типы данных и формы анализа одинаково хорошо согласуются со всеми видами исследовательских вопросов. Из-за его тенденции основываться на данных, а не на вопросах, большая часть текущих количественных исследований в Twitter сосредоточена на измерении и сравнении конкретных структурных параметров в очень больших выборках данных, иногда без учета их же теоретической значимости. Это понятно на фоне исследований больших данных, как принципиально нового подхода к поиску закономерностей, взаимосвязей и связующих звеньев между элементами, а не парадигматического теоретизирования значения заранее указанных элементов [99]. Идеальное исследование должно быть хорошо обосновано определенным набором исследовательских вопросов и запрашивать данные в соответствии с ними. В отличие от традиционных инструментов, таких как опросы и обычный контент-анализ, важно отметить, что даже ознакомительная фаза исследования носит явно количественный характер при изучении социальных сетей. Поскольку поиск, фильтрация и ранжирование - единственно возможный способ сделать массы контента читаемыми для человека-исследователя, они формируют логический первый шаг в любом анализе, даже в качественных исследованиях.

В то же время количественные исследования должны представлять данные в том виде, в каком они относятся к заданным вопросам, а не просто потому, что это возможно и были собраны большие объемы данных. Кроме того, возникает вопрос о том, насколько актуальны пользователи Twitter для общей популяции — как в самой социальной сети, так и за его пределами. Показатели внедрения и стратегии использования сильно различаются, что ставит под сомнение заявления о репрезентативности. При вынесении суждений о группах пользователей Twitter на основе твитов могут быть упущены из виду те пользователи, которые в основном читают, но почти не публикуют сообщения, в то время как значению пользователей с высокой активностью голоса может быть придано слишком большое значение. Выводы о населении в целом на основе Twitter затруднены из-за этого врожденного перекоса, однако без обобщения потенциал социологических исследований ограничен, несмотря на большой энтузиазм по поводу Twitter как источника данных [100].

Третий и наиболее важный набор ограничений основан на технологиях. Как уже указывалось, нет никакого способа проверить, насколько полно данный набор данных отражает то, что проходило через Twitter в момент его компиляции. Не имея доступа к «firehose», исследователи полностью полагаются на Twitter, чтобы предоставить репрезентативную выборку того, что там есть. В то же время стоит отметить, что это происходит не только из-за необходимости Twitter монетизировать свои данные, но и в результате уникальной задачи создания инфраструктуры, достаточно мощной для хранения таких огромных объемов информации в режиме реального времени. Неполные наборы данных могут затруднить анализ, однако ретроспективно запросить полный архив твитов, связанных с прошлым событием, невозможно. Это не только затрудняет новые исследования, но и чрезвычайно затрудняет достижение абсолютной воспроизводимости, что имеет очевидные последствия для достоверности результатов исследований.

Качественные социологические исследования в Twitter обладают своими уникальными возможностями, но также и собственным набором ограничений, например, в отношении конфиденциальности. Можно утверждать, что сила Twitter заключается в способности извлекать интересные идеи из коротких и часто сильно привязанных к контексту сообщений, однако они также трудны для интерпретации и несут в себе широкий спектр значений для различных заинтересованных сторон. В то время как «глубокий», качественный подход имеет больше нюансов, чем вычислительные процедуры, он также сильно ограничен по своему масштабу. Выбирая объект, который может быть детально изучен, исследователь делает конкретный выбор в отношении своего объекта исследования. С другой стороны, поверхностная агрегация данных всегда сопряжена с риском получения суждений, которые плохо подкреплены, поскольку они основаны на неправильных или чрезмерно неявных допущениях. Полагаясь на свой информированный опыт из других контекстов при рассмотрении Twitter в качестве источника данных и прагматично решая, как цели исследования могут быть согласованы с существующими техническими и методологическими проблемами, как правило, можно получить наилучший результат.