Н. А. Залесская (Минск)

ТИПЫ ПОЛЬЗОВАТЕЛЕЙ TWITTER И КАТЕГОРИЗАЦИЯ ПУБЛИКУЕМЫХ ИМИ СООБЩЕНИЙ

Twitter стал одной из передовых платформ для обмена информацией. Следовательно, потребителям Twitter полезно знать происхождение твита, так как это влияет на характер просмотра информации и взаимодействия с этой информацией. В настоящей статье классифицируются твиты только текстового содержания на основе их происхождения. Разрабатывается контент-ориентированная категоризация типов сообщений, публикуемых пользо-вателями Twitter, а также классификация типов пользователей данной платформы на основе специфики производства и потребления информации. Предполагается, что анализ дея-тельности пользователей Twitter (производство и/или комментирование информации) оптимизирует потребление контента и управление последующим освещением события.

Twitter – одна из наиболее популярных социальных сетей: пользователи ежедневно публикуют в ней около 400 миллионов твитов. Эти общедоступные данные служат потенциальным источником информации, хотя такое количество твитов затрудняет нахождение релевантного контента. Поэтому тип пользователя учетной записи Twitter является важной переменной при анализе данных. Информация о происхождении твита влияет на восприятие контента (достоверность, местоположение и т.д.) конечным пользователем.

Одной из целей настоящего исследования является классификация твитов на основе их текстового содержания. Существует ряд причин, почему это может быть полезно. Одна из них заключается в том, что пользователи часто обмениваются контентом, взятом из сторонних источников, публикуя информацию на своей стене в Twitter. Следовательно, источник твита, опубликованного частным лицом, может изначально принадлежать организации. Вторая причина состоит в том, что, в связи с ограничением скорости, получить метаинформацию бывает достаточно сложно. И третья – в том, что информация профиля учетной записи пользователя бывает недоступна или неоднозначна (например, пользователи часто оставляют поля своего профиля пустыми или указывают некорректные данные).

Второй целью данного исследования является попытка показать, что классификация типов пользователей способна улучшить распознавание событий в Twitter. Типы пользователей Twitter могут быть проанализированы с точки зрения разных позиций. Проведенный анализ текстового содержания твитов позволил классифицировать их на два типа в зависимости от того, кем они написаны. Анализ твитов *организаций*, написанных вручную, выявил, что в твитах такого рода события упоминаются гораздо чаще, чем в *личных* твитах.

Для создания базы данных последующего анализа архивировались твиты на английском языке при помощи Twitter Streaming API. При отборе твитов внимание фокусировалось на полях с именами и описании профиля в каждой учетной записи пользователя, написавшего выбранный твит. Твит помечался как *личный*, если в поле имени или в описании профиля отсутствует лексика делового общения, и информация в поле имени является частью имени человека или поле описания профиля начинается либо с *I'm*,

либо с *I am*. Полученная информация проверялась вручную. Частный пользователь был определен как индивид, использующий Twitter в повседневной жизни для публикации информации о своей деятельности и личном статусе, комментирования социальных проблем и/или взаимодействия с близкими людьми. Был проведен анализ 100 пользовательских аккаунтов, с которых были отправлены твиты на английском языке. Учитывались имя, описание профиля, местоположение и URL-адрес учетной записи пользователя Twitter. Каждая учетная запись классифицировалась как *личная*, *неличная* или *неопределенная*. Среди 300 учетных записей, определенных вручную, 91,5 % были классифицированы как принадлежащие отдельным лицам, 5 % были определены как неличные и 3,5 % были отмечены как неопределенные.

Тем не менее классификация твитов в качестве *личных* только потому, что они были написаны пользователями, которым принадлежат учетные записи, может привести к неточностям (так как могут учитываться только те пользователи, которые предоставили в профиле более полную информацию о себе). Для решения этой проблемы был проведен анализ дополнительных учетных записей Twitter с учетом того факта, что приложения, разработанные специально для портативных устройств (например, Twitter для iphone), часто используются отдельными лицами для создания твитов, тогда как организации для создания контента в основном используют веб-версию Twitter и приложения для ПК.

Для дальнейшего анализа была собрана информация об источнике (т.е. о программных приложениях, использованных для создания твитов), отобраны англоязычные твиты, опубликованные в случайно выбранный день, и идентифицированы те, которые были явно написаны вручную и опубликованы из приложения. Данные твиты составили как минимум 1 % всех твитов. Табл. 1 показывает, какой процент твитов на английском языке пришелся на каждое приложение Twitter для различных платформ, на которые в совокупности приходилось примерно 66 % всех твитов.

Таблица 1

Твиты, опубликованные из приложений для мобильных устройств

Приложение	Процент
Twitter для iPhone	37,11
Twitter для Android	16,50
Twitter для BlackBerry	5,50
Twitter для iPad	2,55
Веб-версия (m2)	1,46
iOS	1,36
Echofon	1,29
Всего	65,77

Для проверки гипотезы о том, что высокий процент данных твитов – это *личные* твиты, был проведен еще один анализ. Было выбрано 300 учетных записей на английском языке, твиты из которых были отправлены с помощью одного из перечисленных выше приложений для мобильных устройств, и классифицированы в соответствии с прежними условиями. Из этих 300 учетных записей 87,1 % принадлежат индивидам. Только 1 % был оценен как явно неличный, тогда как 11,9 % были помечены как неопределенные.

Провести подобный анализ для идентификации твитов, написанных *организациями*, оказалось сложнее. Организации описывают себя различными способами, что затрудняет определение их названий в профилях пользователей. Кроме того, названия организаций часто упоминаются в учетных записях отдельных лиц как места работы (например, *I'm a software engineer at Microsoft Corporation/Я инженер-программист в корпорации Microsoft*). Поэтому для получения твитов организации использовались веб-каталоги организаций (например, www.twellow.com), и их твиты были собраны с помощью API Twitter. Было использовано 150 учетных записей организаций, с которых были отправлены твиты на английском языке. Чтобы не допустить преобладания сообщений одной организации, от каждой было выбрано не более 50 твитов. Чтобы автоматически отличать *личные* твиты от твитов *организации*, применялись два простых правила. Первое состоит в том, что «ответы» (@ ссылка на пользователя упоминается в начале твита) редко встречаются в твитах *организации*. Следовательно, если твит содержит ответ, скорее всего, это *личный* твит. Второе наблюдение заключается в том, что твиты *организации* часто содержат веб-ссылку на внешний контент. Если твит не является ответом и содержит веб-ссылку, он помечался как твит *организации*. В противном случае он классифицировался как *личный*.

Примеры твитов организаций

• *Diet Coke may be the new #2 popular, but U.S. soda market is shrinking/ Диетическая Кола может стать #2 по популярности, но в США рынок газированных напитков сокращается http://ow.ly/1bSNnh*;

• *Apple likely to introduce smaller, cheaper iPad mini today http://t.co/TuKBHZ3z/Вероятно, Apple представит сегодня более дешевый iPad mini http://t.co/TuKBHZ3z*.

Примеры личных твитов

• *Watching Black Mirror. It has totally sucked me in:D #notsomuch lol/Смотрю «Черное зеркало». Я подсел:D #несовсем лол*;

• *@john It's a stress fracture. Nah, no dancing was involved!/Это реально жесть. Нет, танцев не было.*

Twitter – это платформа, на которой пользователи выражают эмоции и свое отношение к происходящему. Для определения типа тональности твита (положительный, отрицательный или нейтральный) использовался инструмент Sentiment API. Предполагалось, что *личные* твиты передают позитивное или негативное настроение, а твиты *организации* – более нейтральные.

В ходе исследования выявлено, что твиты *организации*, как правило, содержат больше существительных. Авторы *личных* твитов чаще используют личные, возвратные и притяжательные местоимения (*me*, *us*, *you*, *myself*, *ourselves*, *my*, *mine*, *our*), так как обычно описывают себя (от первого лица) или обращаются к аудитории (второе лицо). Организации также часто обращаются к своей аудитории, используя в твитах местоимения второго лица (*Will you High Five the Penguins or Bears? #pittsburhgpenguins #hersheybears Sign up for a chance to win a trip to the Cup Final: http://t.co/XQP8ZDOINV/ Вы/Ты за кого? Пингвинз или Беарс? #pittsburhgpenguins #hersheybears. Подпишись, чтобы получить шанс выиграть поездку на финал Кубка: http://t.co/XQP8ZDOINV*). Глаголы в твитах *организаций* чаще встречаются в третьем лице. В *личных* твитах преобладает простое прошедшее время (Past Simple) вместо перфектных форм (*mattfellows: already arrived in LAX/уже прибыл в аэропорт Лос-Анджелеса*). Нередки в *личных* твитах глаголы в инфинитиве в сочетании с модальными глаголами (*funkeybrewster: @redeyechicago What should my next video be about?/О чем должно быть мое следующее видео?*).

Выявлено, что прилагательные в превосходной степени употребляются для выражения эмоций и мнений (*личные* твиты), тогда как в сравнительной степени – для констатации фактов и предоставления информации (твиты *организаций*). Наречия, в основном, используются в *личных* твитах для придания глаголу эмоциональной окраски. На удивление часто в *личных* твитах встречается вопросительное местоимение *whose/wh$* (*чей?*). Однако исследование выявило, что авторы *личных* твитов, как правило, используют «чей» в качестве сленговой версии выражения *who is* (*кто есть*). Например, *dinner & performance spectacular tonight!:) whose ready for some wine??/ сегодня вечером ужин и спектакль впечатляют! Кто готов выпить вина?*. Другим индикатором *личных* твитов являются наречия в превосходной степени, такие как *most* (*большинство/в большей степени*) и *best* (*лучше всего*). Например, *brrooooklyn: I love best the sound my iPod makes when I shuffle its apps. Boo bee boo))/brrooooklyn: Обожаю звук, который издает мой iPod, когда я переставляю иконки приложений. Буу бии буу))*. В *личных* твитах также нередко присутствует притяжательный падеж (Possessive Case).

Если тип тональности *личного* твита отрицательный, в нем чаще встречаются глаголы в прошедшем времени (Past Simple), потому что ряд авторов выражает негативные эмоции по поводу неких потерь или разочарованы в произошедших событиях. Вот пример наиболее часто применяемых глаголов: *missed* (*пропущено*), *bored* (*наскучило*), *gone* (*пропало*), *lost* (*потеряно*), *stuck* (*застряло/застопорилось*), *taken for* (*принято за что-л.*). *Личные* твиты часто демонстрируют эксцентричное и небрежное использование языка, тогда как твиты *организаций*, как правило, носят более формальный характер. Твит *организации* начинается с заглавной буквы, в нем правильно расставлены знаки препинания, все буквы в твите строчные. Новостные *организации* часто добавляют дескриптор темы к началу твита

(*Kanye West Scandal: Kanye West's signature reportedly forged on the contract at the upcoming NYFW event http://t.co/hpyLQYeL/Скандал с участием Канье Уэста: по сообщениям в прессе – на контракте, касающемся участия Уэста в предстоящей Неделе Моды в Нью-Йорке, подпись Уэста подделана*). С учетом этого представляется целесообразным анализировать употребление точки с запятой или дефиса в пределах первых трех слов твита. Использование удлиненных слов (например, *coooooooool; gooooal*) для усиления высказывания – еще одна особенность *личных* твитов. С учетом этого проводился поиск слов, в которых один и тот же символ повторяется три раза и больше (*i'm backkkk! guuudmorning twits/я веррнууулся! доообroее утро!*). Пользователи прибегают к произвольному употреблению сокращенных слов, в которых пропускаются или заменяются некоторые символы, чтобы соответствовать ограничению в 280 символов для одного твита (например, можно наблюдать переключение кодов – *2mrw – завтра*, *goodn8 – доброй ночи*). Кроме того, в *личных* твитах интенсивно используются хэштеги, сами твиты длиннее и упоминания пользователя в твите чаще (*Check out my blog I updated 2day 2 learn abt tuna #tunafish/Посмотрите мой блог, который я обновил сегодня, чтобы побольше узнать о тунце! #tunafish*).

Twitter предоставляет собой инструмент, с помощью которого пользователи ищут твиты по ключевым словам. Тем не менее запросы по ключевым словам для событий часто могут приводить к не относящимся к делу результатам из-за многозначности ключевых слов и переносного или метафорического их значения. Например, в Twitter поиск событий, посвященных гражданским волнениям (уличным беспорядкам) по нескольким репрезентативным ключевым словам (например, *strike* (*забастовка*), *rally* (*митинг*), *riot* (*бунт*) и т.д.) часто приводит к результатам, которые относятся к спортивным событиям, таким как *страйк* в боулинге или теннисное *ралли.* Например, *Robbie Meade: My new emote when I know I got a strike #bowling/ То самое чувство, когда понял, что у меня страйк #bowling.* Или к ситуациям, в которых ключевые слова используются в переносном значении. Например, *She is a riot!/Она бунтарка!*. Возникает вопрос о том, поможет ли определение типа пользователя, отправившего твит, преодолеть неоднозначность.

Чтобы исследовать гипотезу о том, что тип пользователя может влиять на поиск релевантных событий в Twitter, была проведена серия экспериментов на основе информации о двух типах событий – уличных беспорядков и вспышек заболеваний. В табл. 2 отражено процентное соотношение твитов, связанных с указанными событиями.

Таблица 2

Твиты, связанные с событиями, %

	Беспорядки	Вспышки заболеваний
Личные твиты	5,27	9,52
Твиты организаций	36,54	39,34
Все твиты	12,50	20,07

В табл. 3 представлены ключевые слова, которые используются для поиска в Twitter информации о рассматриваемых событиях.

Таблица 3

Ключевые слова, используемые для поиска в Twitter
для двух типов событий на английском языке

Уличные беспорядки	protest, protested, protesting, riot, rioted, rioting, rally, rallied, rallying, marched, marching, strike, struck, striking
Перевод	протестовать, протест, бунт, бунтовать, митинговать, митинг, маршировать, марш, забастовка, бастовать
Вспышки заболеваний	outbreak, epidemic, influenza, pandemic, quarantine, cholera, Ebola, flu, malaria, hepatitis, measles
Перевод	вспышка, эпидемия, грипп, пандемия, карантин, холера, лихорадка Эбола, грипп, малярия, гепатит, корь

С использованием API-поиск в Twitter был проведен отбор твитов, которые содержали хотя бы одно из ключевых слов, перечисленных в табл. 3. Анализ показал, что большинство твитов (> 80 %), в которых упоминаются ключевые слова, не относятся к событию, что подтверждает ненадежность использования только ключевых слов.

Таблица 4

На 200 твитах для каждого типа событий на английском языке

	Английский
Уличные беспорядки	89
Вспышки болезней	82

Между двумя типами пользователей существует заметная разница в количестве твитов, относящихся к реальным событиям. При рассмотрении фактов уличных беспорядков и вспышек заболеваний в твитах выявлен гораздо больший процент твитов *организации* с ключевыми словами, в которых упоминается событие, в отличие от *личных* твитов. Табл. 2 показывает, что в категории «уличные беспорядки» твиты *организации* на английском языке в 7 раз чаще (36,54 % против 5,27 %) сообщают о реальном событии, чем *личные* твиты с теми же ключевыми словами. В категории «вспышки заболеваний» твиты *организаций* имеют в 4 раза больше шансов сообщить о событии (39,34 % против 9,52 %). В твитах *организации* ключевые слова для поиска событий используются более согласованно и с большей конкретикой, чем в *личных*.

Итак, на основе текстового содержания твиты можно разделить на твиты *организаций* и *личные* твиты. Реальные события гораздо чаще упоминаются в твитах *организаций* с ключевыми словами событий, чем в *личных* твитах с теми же ключевыми словами, и поэтому включение информации о типах пользователей в модели распознавания событий весьма желательно. Создание отдельных классификаторов распознавания событий для твитов разных типов пользователей значительно более эффективно, чем использование

единой модели распознавания событий во всех твитах. Организации достаточно часто указывают на внешние источники информации через URL-адреса своих постов, тогда как частные пользователи больше отражают свой личный опыт и настроения в отношении событий и более активно взаимодействуют с другими участниками.