ОЦЕНКА КАЧЕСТВА СООБЩЕНИЙ TWITTER ДЛЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

А. И. Трубчик
Факультет компьютерных систем и сетей,
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: apri@tut.by

Интеллектуальный анализ данных на базе ленты Twitter требует учитывать не только объем отправленных сообщений, но оценивать их качество. Ретвиты - это ключевой механизм распространения информации в социальной сети Twitter. Любое сообщение в этой социальной сети можно опубликовать в собственной ленте, сделав «ретвит». Ретвит появляется под псевдонимом оригинального автора и ссылкой на него. Также это отличный индикатор популярности сообщения и его социального одобрения. В данной работе предлагается способ оценки качества сообщений на основе количества их ретвитов, даты публикации и числа подписчиков авторов.

Введение

В работе [1] исследовалась возможность использования социальной сети Twitter для повышения результативности прогнозирования на рынках электронной торговли. Анализ был направлен на нахождение корреляций (т.е. что данные связаны с событиями происходящими на фондовом рынке) и на те случаи, когда высокая активность в Twitter сигнализирует о будущих движениях рынка. Основываясь на эмоциональной оценке текста были выделены следующие категории твитов: позитивный, негативный, смешанный (негативный и позитивный) и нейтральный (не негативный и не позитивный). На этапе подготовки данных отфильтровывалось большое количество сообщений, включая ретвиты, и при нахождении корреляции использовался только количественный объем твитов за период времени. Можно сказать, что определение эмоционального содержания должно было обеспечить оценку качества сообщений, но небольшие объемы, сленг и специальные символы в текстах не позволяют достаточно точно это сделать.

Альтернативой или дополнением к определению эмоций в тексте может быть оценка качества сообщений на основе публичных данных, предоставляемых Twitter: количество ретвитов сообщения, дата его публикации и число подписчиков автора.

I. Моделирование

Грубой оценкой качества твита может являться простое деление количества подписчиков автора сообщения на число ретвитов, но в то же время пользователи с огромным числом подписчиков будут оказывать чрезмерное влияние на всю выборку.

Другим способом оценки качества твита может выступить кластерный анализ, который разобьет совокупности объектов на однородные группы и это даст возможность выделить плохие/хорошие сообщения. В целом методы машинного обучения позволяют решить данную проблему, но требуют качественной подготовки тренировочных данных.

Можно разработать модель, приняв согласно [2], что посещение пользователями сайта Twitter представляет собой процесс Пуассона и параметр распределения Пуассона λ - это среднее количество посещений в единичном интервале.

Пусть s_1 - число подписчиков автора, s_2 это число подписчиков второго порядка и одновременно количество пользователей, которое может просмотреть твит, если кто-то из подписчиков первого порядка сделает ретвит. Соответственно s_k - число подписчиков k-го порядка графа подписчиков Twitter. Далее $v_k(t)$ будет представлять процент пользователей, которые увидели оригинальное сообщение в течение времени t (количество единиц времени до даты публикации твита). Можно выразить количество ретвитов $Y(t)$, которое получит сообщение за время t:

$$Y(t) = \sum_{k=1}^{\infty} v_k(t) s_k p. \qquad (1)$$

Здесь p вероятность того, что будет сделан ретвит и соответственно наша оценка качества твита.

Среднее значение количества подписчиков в сети Twitter (все регионы) [3] около 91, поэтому s_2 можно приблизительно принять равным $91 s_1$. Оценить количество подписчиков третьего и далее порядков не представляется возможным вследствие отсутствии информации об этом, поэтому ограничимся k в интервале от 1 до 2.

Поскольку посещение Twitter пользователями это процесс Пуассона, то количество пользователей $v_k(t)$, просмотревших сообщение в течение времени t, описывается экспоненциальным распределением. Для определения количества просмотров за время t от подписчиков пер-

вого порядка s_1 получим:

$$v_1(t) = \int_0^t \lambda e^{-\lambda\tau} d\tau = 1 - e^{-\lambda t}.$$

Для подписчиков второго порядка нужно принять во внимание тот факт, что они увидят оригинальное сообщение только если его «ретвитнут» читатели из первого порядка. Поэтому учтем вероятности ретвита p и просмотра $v_1(t)$, а также временные рамки $t-\tau$, где τ - это время начала просмотра ретвита:

$$v_2(t) = \int_0^t p v_1(t-\tau)\lambda e^{-\lambda\tau} d\tau =$$

$$= p(1 - e^{-\lambda t}(\lambda t + 1)).$$

Для порядка $k = 2$ уравнение (1) принимает следующий вид:

$$Y(t) = v_1(t)s_1 p + v_2(t)91 s_1 p =$$

$$= (1 - e^{-\lambda t})s_1 p + 91(1 - e^{-\lambda t}(\lambda t + 1))s_1 p^2.$$

Чтобы найти p решим это квадратное уравнение и получим формулу (2). Таким образом вероятность p, которая является искомой оценкой твита, выражена в виде зависимости от переменных, которые Twitter предоставляет публично: количество ретвитов, промежуток времени и число подписчиков автора.

II. Пример использования

Для примера допустим $\lambda = 1$, а время $t = 0,5$ будет соответствовать 30 минутам. Результаты вычислений для различных значений количества подписчиков и ретвитов представлены в таблице 1.

Таблица 1 – Результаты при $\lambda = 1$

№	Подписчиков s_1	Ретвитов $Y(t)$	Время публикации t	Вероятность p
1	500	10	0,5	0,03090
2	500	20	0,5	0,04983
3	500	30	0,5	0,06482
4	500	10	1	0,01854
5	500	20	1	0,02970
6	500	30	1	0,03850
7	1000	10	0,5	0,01837
8	1000	20	0,5	0,03090
9	1000	30	0,5	0,04106
10	1000	10	1	0,01111
11	1000	20	1	0,01854
12	1000	30	1	0,02454
13	10000	10	0,5	0,00242
14	20000	100	0,5	0,01044
15	30000	20	0,5	0,00164
16	10	100	0,5	1,08003

Закономерно, что твит №1 от автора с числом подписчиков 500, набравший за 30 минут 10 ретвитов можно оценить, как более лучший по качеству, чем твит №7 от автора с 1000 подписчиков.

Твит №15 был отправлен популярным автором, но имеет низкую оценку, так как не набрал должное количество ретвитов.

Твит №16 набрал ретвитов больше, чем подписчиков у автора, и p оказалась чрезмерно высокой. Это возможно только в одном случае: если на сообщение была дана ссылка с внешнего по отношению к Twitter источника информации (крупные новостные веб-сайты). Подобную оценку при интеллектуальном анализе данных можно считать выбросом и никак ее не учитывать.

Заключение

Ретвиты являются индикаторами популярности сообщений и их социального одобрения. Социальная сеть Twitter предоставляет небольшое количество информации и предложенная модель предполагает множество допущений. Невозможно оценить количество подписчиков третьего и далее порядков, а также их вклад в полученную формулу.

Также твиты могут быть добавлены в «избранные» сообщения пользователей. Число добавленных в «избранные» отображается публично, но в этой работе они не учитываются, так как «избранные» твиты не публикуются в ленте и подписчики второго и более порядков их не увидят.

Оценка качества с помощью представленной формулы может быть использована напрямую вместо простого подсчета объема твитов, например при исследовании корреляции с использованием данных социальной сети Twitter. Такая оценка качества твита сгладит всплески большого объема несложных сообщений от непопулярных авторов, которые могут почти не оказывать влияния на исследуемые объекты.

1. Трубчик, А. И. Twitter как индикатор в задачах электронной торговли / А. И. Трубчик // 51–я научная конференция аспирантов, магистрантов и студентов по направлению 4: Компьютерные системы и сети – Минск : БГУИР, 2015. – С. 22.
2. Lee, K. Who Will Retweet This? Automatically Identifying and Engaging Strangers on Twitter to Spread Information / K. Lee, J. Mahmud, J. Chen, M. Zhou, J. Nichols – 2014.
3. Myers, S. A. Information Network or Social Network?: The Structure of the Twitter Follow Graph / S. A. Myers, A. Sharma, P. Gupta, J. Lin // Proceedings of the 23rd International Conference on World Wide Web – 2014. – P. 493–498.

$$p = \frac{-s_1(1 - e^{-\lambda t}) + \sqrt{(s_1(1 - e^{-\lambda t}))^2 + 364 s_1 Y(t)(1 - e^{-\lambda t}(\lambda t + 1)}}{182 s_1 (1 - e^{-\lambda t}(\lambda t + 1))} \quad (2)$$