Анализ отклика юзеров Twitter на новости из средств массовой информации

Аннотация. В работе проводится подсчет количества участников социальной сети Twitter, которые разместили сообщения к основной новости дня. Для получения статистических данных использовалась универсальная программа, которая была разработана для оценки уровня заинтересованности, проявляемой пользователями Twitter к главной новостной теме. База новостных текстов получена из архива новостей сайта РИА Новости, тексты сообщений пользователей сети Twitter – из Twitter Api (интерфейс, позволяющий пользоваться данными социальной сети Twitter).

Ключевые слова: статистические данные, Twitter, новостная аналитика, СМИ, новости, события дня.

1. Введение

Одной из основополагающих потребностей человека считается коммуникация. Для виртуального общения сегодня создано множество Интернет-порталов, к которым относится и Twitter, где зарегистрированные пользователи общаются, обмениваются эмоциями и новостями, выражают свою точку зрения относительно любых вопросов.

Таким образом, Twitter – это надежный источник новостных тем, к которым проявляют заинтересованность участники социальной сети. Благодаря данному ресурсу можно узнать, как повлияла новость из СМИ на общественность, какие события взволновали пользователей, а какие нет. В этой связи изучение реакции пользователей к сообщениям средств массовой информации и выступает главной целью данной научной работы.

Исследования в этом направлении проводятся регулярно (к примеру, сайт Лента.Ру, компания BrandAnalytics). Технология данных исследований заключается в определении популярности СМИ посредством подсчета в сообщениях пользователей числа размещенных ссылок. Задача статьи заключается в установлении взаимосвязи между сообщениями участников социальной сети и актуальными темами, которые размещаются в СМИ.

2. Поиск новостных тем в сообщениях пользователей

Для достижения главной цели, нужно решить следующие задачи:

1) отыскать способ выделения ключевых слов в актуальных темах для текущего дня;

2) разработать порядок генерации запросов в социальной сети Twitter, в которых упоминаются интересуемые ключевые слова.

Источником новостей послужил сайт РИА Новости, который обладает существенной базой новостных тем. Расчет важности упомянутых слов проводился при помощи статистической меры tf*idf [TF-IDF], которая позволяет оценить важность слов в исходном документе, являющемся, в свою очередь, составляющей коллекции документов.

TF – это показатель соотношения частоты встречаемости конкретного слова с общим количеством слов в документе. Его значение рассчитывалось путем сопоставления частоты встречаемости определенного слова с общим количеством всех слов общего объема документов. Такой подход позволил более точно определить важность ключевых слов. IDF представляет собой инверсию частоты, которое встречается с некоторым словом в документах базы. Мера TF-IDF – это произведение рассматриваемых сомножителей.

В ходе эксперимента были выгружены все новостные темы за день и отсеяны стоп-слова. Далее были выбраны три ключевых слова, значение tf*idf которых было наибольшим. К каждому из этих слов составлялся список тем, в которых они упоминались.

Данный список также был проанализирован, в результате чего вновь были отобраны три слова с наибольшим значением tf*idf.

Если вторая группа ключевых слов совпадала с первой, то считалось, что за этот день случилось одно существенное происшествие, которое и было описано данными словами. Иначе создавалась новая тема. Данная совокупность слов использовалась в роли поискового запроса в Twitter через Twitter Api. При этом запросы осуществлялись через равные временные промежутки. Из списка сообщений отбирались лишь те, которые были получены в первый раз. Из уникальных сообщений формировалась база данных, которая предназначена для обработки собранной информации.

Мера заинтересованности новостями из СМИ определялась посредством количества уникальных сообщений, которые были получены из запросов по этой теме. Полученной в результате эксперимента картины достаточно, чтобы сделать качественные выводы. Следует также отметить, что те сообщения, которые были отправлены с учетных записей СМИ, не брались во внимание. Данная мера была предпринята по той причине, что подобные ресурсы не отражают мнение пользователей.

3. Экспериментальная часть

Тестирование проводилось в два этапа. Первый период длился 38 дней (начался с 04.11.14 и закончился 20.12.14). За это время программа выдала 124268 сообщений, из которых 56122 – сообщения новостных сайтов, а 68146 – сообщения участников социальной сети. Полученные данные представлены в виде диаграммы (рис. 1):

1) столбчатая диаграмма красного цвета отображает количество сообщений, авторство которых принадлежит новостным агентствам;

2) столбчатая диаграмма синего цвета – количество сообщений, которые были написаны Интернет-пользователями;

3) линейная диаграмма – число пользователей, сообщения которых в этот день попали в выборку.

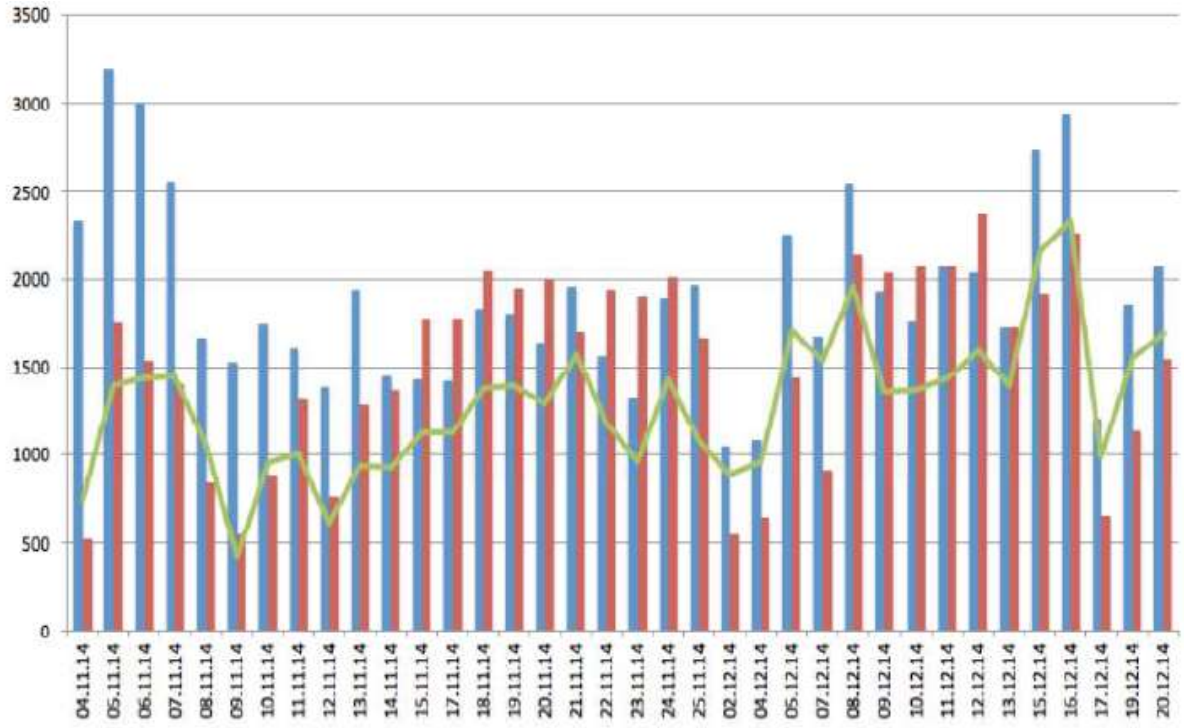

Рисунок 1 – Статистические данные за период 04.11.14-20.12.14гг.

На рисунке 1 можно увидеть резкие колебания. Так, к примеру, 5 ноября 2014 года в новостной ленте главной темой была Украина и Россия. В этот день в Twitter число сообщений пользователей увеличилось до 3200, что практически в 2 раза превышает упоминания об этом информационных агентств. Однако период 06.11.14-15.11.14 отражает сокращение заинтересованности участников социальной сети, несмотря на неизменность главной темы, что подтверждает общее количество сообщений.

С 08.12.14 пользователи обсуждали новости, связанные с возрастанием курса валюты. К 16 декабря 2014 года количество заинтересованных пользователей увеличилось до 2300 человек.

Рассмотрим второй период, который продолжался с 21.12.14 по 18.02.15. За это время программа выдала 148175 сообщений, из которых 48890 – это сообщения новостных агентов, а 99285 – сообщения самих участников социальной сети Twitter. На этом этапе число сообщений от пользователей превысило количество сообщений информационных агентств почти в 2 раза. Полученные данные можно наглядно увидеть на диаграмме, представленной на рисунке 2.

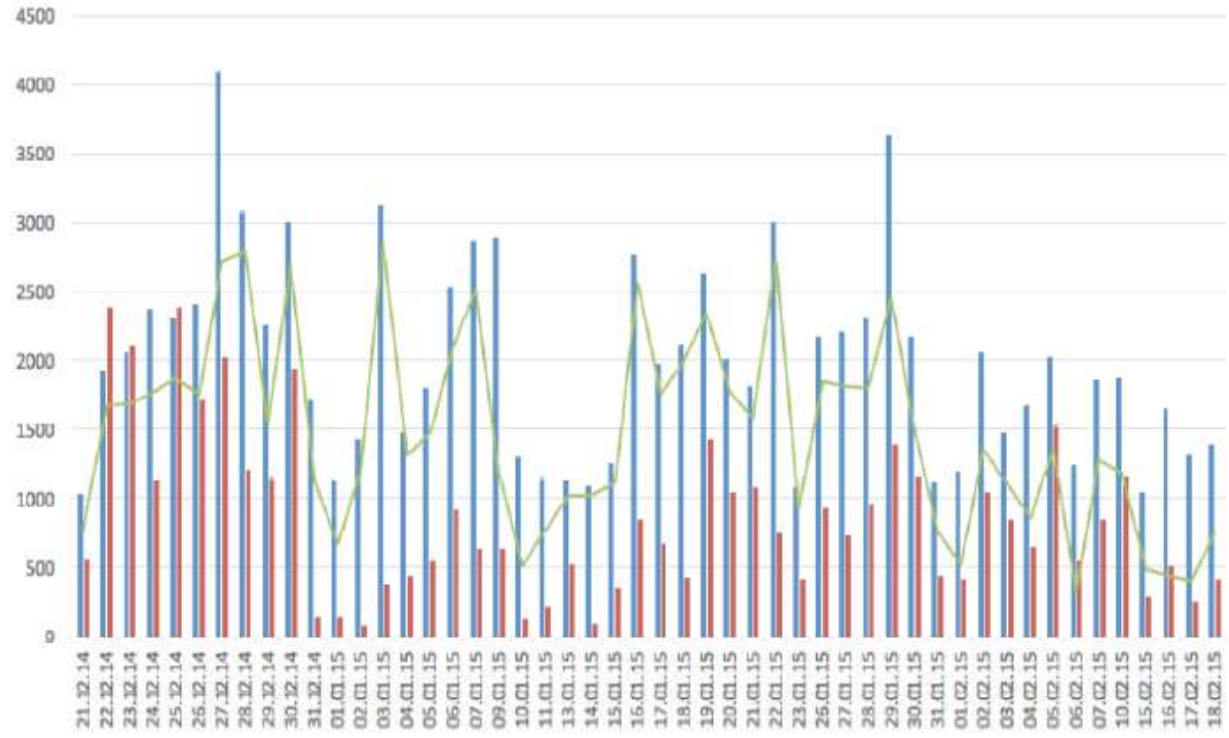

Рисунок 2 – Статистические данные за период 21.12.14-18.02.15гг.

Диаграмма отображает явные скачки числа сообщений. Период 21.12.14-31.12.14 отражает обсуждение в сети новости, связанной с ростом валюты. По динамике сообщений можно сделать вывод, что данная тема стремительно набирала популярность – 27 декабря число твитов увеличилось до 4000, что в 2 раза больше количества сообщений новостных агентов. Следовательно, можно сделать вывод, что общественность действительно была взволнована данным вопросом. Позже наблюдается снижение тенденции, скорее всего причиной этому стали новогодние праздники.

Период 03.01.2015-29.01.2015гг. отражает обсуждение в социальной сети сразу нескольких тем: события на Украине и снижение курса рубля, по которым число сообщений пользователей лидировало. Однако с 10.01.15 наблюдалось падение активности. 29 января произошел еще один скачек – количество сообщений достигло отметки 3600, после чего популярность тем начала постепенно снижаться.

Следует заметить, что представленную статистику сложно назвать большой. В сравнении с общей картиной публикуемых сообщений в Twitter, полученные программой

цифры незначительны. Кроме того, Twitter Api имеет ограничения для осуществления запросов. Следовательно, для эксперимента было получено меньше твитов, что может отразиться на результатах.

4. Заключение

В результате проведенной работы, можно сделать вывод, что при помощи программы удалось собрать некоторые статистические данные и воссоздать картину активности участников социальной сети Twitter. Из диаграмм можно предположить, какие новости больше волнуют общественность. Однако эксперимент усложняется тем фактом, что новостные агентства зачастую дублируют новости в сообщениях, а пользователи просто делятся ссылками информационных сайтов. Такие обстоятельства не позволяют отследить реакцию заинтересованных людей, поэтому сложно узнать их мнение.

Список использованных источников

1. TF-IDF – https://ru.wikipedia.org/wiki/TF-IDF
2. Twitter Api – https://developer.twitter.com/en/docs.html