Statistical Analysis of Hockey-Tweeting Twitter Users' Habits and Interactions

10.17975/sfj-2015-007

AUTHORS: Nikki Sigurdson and Mahsa Naserifar**SCHOOL:** Earl Haig Secondary School, Grade 12

Abstract

Big data analysis techniques can make significant impacts on social trend information. Hockey is a popular sport internationally, and online communities have formed on social media websites such as Twitter. This paper aims to investigate information available on Twitter about users connected to hockey. It also aims to explain Twitter data collection processes and the significance of social media information collection. Using a set of routines developed by the authors in python 3.3 and with the Twitter 1.16 API, 25,189 messages ("tweets") matching hockey keywords were collected. From it, further information about users' tweeting habits and the overall communities' habits were found. The highest percentages of frequent tweeting about hockey was during large sports events not directly related to hockey.

Key Words

sports, social networks, twitter

Introduction

BIG DATA is one of the fastest growing fields in computer science and data analysis. Social media networks contain extremely valuable information about society and consumers in the 21st century by providing users with quick access to updates and other information.

TWITTER is a social media network that allows users to "tweet" with 140 character phrases to update their followers about their lives and ideas. A popular method of labelling trends is through "hashtags". Hashtags prove to be very valuable tools for Twitter user analysis. The fast data communication method that twitter provides to its users has attracted over 302 million monthly active users. (1) Twitter claims to observe 500 million tweets sent per day, and says 80% of its active users are on mobile. This trend towards mobile communication makes Twitter data a valuable resource of information about users' lives and habits and is, therefore, a source 8 insight into many aspects of the users themselves.

HOCKEY is a popular sport in North America and around the world, and the hockey online community has a presence on Twitter. Trends show a tendency towards online hockey information retrieval and away from traditional television broadcasting methods,

which makes social media hubs critical to hockey industry leaders who want a grasp of their consumer community. Rogers owns the NHL broadcasting rights of Canada, and has noticed a severe shift in the preferred medium of it's consumers. (2) As fans choose online methods over other forms of information consumption, social media becomes an integral aspect of understanding hockey consumers.

This paper investigates the habits and community interaction of some Twitter users who tweeted about hockey between April 30 2015 and May 7 2015. Several questions were asked regarding the topic, including whether or not the hockey community is made up of a small number of users who tweet about hockey frequently, or if it is a large number of users tweeting about hockey infrequently. As well, the presence of commercial hockey Twitter accounts is observed.

Background

Several research studies have been conducted using twitter public stream data. Erik Tjong Kim Sang from Netherlands eScience Center, and Antal Van Den Bosch from Radboud University used Twitter's public stream data to analyse large amounts Dutch language tweets. (3) Their work shows how tweet data can be used to track trends in a population. Analysis of societal trends through social network tracking shows huge promise in the future for epidemic prediction.

Bakhshandeh et. al analyzed Twitter users and their degrees of separation with a variety of different search algorithms in an attempt to find the best method. (4) The best method was the algorithm that both found the smallest degree of separation between users, and made the least amount of requests to Twitter for information. The algorithms tested were originally designed for memory based search. The application of these algorithms to online social networks for degree of separation calculation exemplifies how numerous the possibilities of social network research are. This area of research will continue to grow in the future.

Miguel Rios and Jimmy Lin of Twitter Inc. studied Twitter user interactions during the 2010 FIFA world cup and the 2011 Tohoku Earthquake to understand how users are communicating on Twitter. (5) Their use of visualizations highlighted how users are communicating internationally and their results indicate significance in social network analysis.

Methods

Twitter Application Program Interface (API)

Twitter REST application program interface (API) v1.1 allows information that is already publically available on the Twitter website to be accessed and collected. (6) The API library used in this research was twitter 1.16. (7) Twitter requires REST API users to authenticate requests, and it utilizes OAuth services to enable 3rd party users to access its public stream data. (8) Data that is restricted by users cannot be accessed through the public stream. Other data however that is not easily accessible online can be collected through the public stream, such as a user's profile stylistic choices.

Collecting Data

A set of routines developed in Python 3.3.5 (9) was used to obtain 2.74 Mb of text data about users tweeting with certain keywords through Twitter's API. The keyword focused on in this investigation was "hockey". With Twitter's API, the letter case of the string was irrelevant. The text of a tweet is considered as one single string by Twitter's API; thus, both "hockey" and the hashtag form "#hockey" were accepted when a request was sent by the keyword "hockey".

Using twitter's REST API, requests were sent for real time most recent 100 tweets, and this was looped to collect around 3000 tweets per collection period. User IDs and the tweet text data were kept and stored in text files. The tweet collection period run time average was about 3 hours.

To analyse the user's trends, the data was stripped of the tweet text component, and new cleaned text file data were created using SAS University Edition software. (10) SAS was chosen for its sim-

pler string cleaning functions compared to python. These cleaned data were returned to python for numerical analysis; the datasets were used to calculate user tweeting frequency per collection period.

Requests were sent as integer values for total collection, and the collection program ran until the requested number of tweets met the termination criteria. It was observed that the Twitter API simply returned an increasing number of duplicate tweets when request numbers were too large. Specifically, after about 400 requests per run, the number of duplicate tweets being returned became higher than the number of unique tweets being returned. Data collection was found to be optimal with more frequent, smaller runs. The data collecting algorithm had to incorporate 2 methods of duplicate tweet identification. One of these methods was checking for entirely identical tweet text strings; the other was to check only portions of the string data for duplicates.

Results

Results A: Frequency of Users Tweeting

A total of 25,189 unique tweets was observed over the 8 day data collection period. Data regarding unique tweets and unique users was computed during the initial data collection process. The percentage of frequent users is defined by the percentage of users who, on average, tweeted more than once during the collection run period. The formula used to compute this percentage was:

$$\text{Percentage} = ((1 - (\text{Unique Tweeters} / \text{Unique Tweets})) \times 100)$$

The quotient of unique users divided by the number of unique tweets would be the percentage of users tweeting only once during

Obs	Date	Time (+0000)	Unique Tweets	Unique Users	Percentage of Frequent Tweeters Avg.
1	April 30 2015	3:23:37	1635	1375	15.90%
2	April 30 2015	10:19:20	1387	1190	14.20%
3	April 30 2015	21:49:54	1525	1320	13.44%
4	May 1 2015	1:20:25	1655	1494	9.73%
5	May 1 2015	10:01:24	1044	890	14.75%
6	May 2 2015	0:52:47	3608	3103	14.00%
7	May 2 2015	9:02:32	1165	879	24.55%
8	May 3 2015	3:01:01	1165	879	24.55%
9	May 3 2015	3:59:55	3497	2800	19.93%
10	May 4 2015	2:13:13	2077	1817	12.52%
11	May 5 2015	4:17:47	1618	1277	21.08%
12	May 6 2015	2:33:36	2572	2313	10.07%
13	May 7 2015	3:57:15	2241	1796	19.86%

TABLE 1:

Twitter data results, collected in real time over a period of 8 days. Each observation is the total results of one instance of data collection at the time specified. Unique user means multiple instances of one user tweeting more than one time in the data collection period may only be counted once. Time is in GMT (Greenwich Mean Time).

the collection period: in order to collect statistical data on users tweeting more than once during this period, the difference between 1 and the quotient aforementioned was calculated.

Observations 9 and 13 were collected at almost the same time of day, 4:00:00 GMT, and although the data was collected 5 days apart, their frequency percentage is nearly identical, 19.93% and 19.86% respectively. This suggests 24 hour trends related to user tweet frequency could exist.

Observations 2,3,5, and 7 in Table 1 were collected at night, and this data is relatively consistent with the exception of observation 7 from May 2 2015. This was the night of the professional boxing match between Manny Pacquiao and Floyd Mayweather, (11) an event which impacted several collections of data during that period, including observation 8, which also possessed irregularly high user frequencies.

One possible factor affecting these results would be major commercial accounts which may be proportionally contributing to the appearance of frequent tweeters. An example of such an account would be @nhl with over 3.5 million followers. With a higher percentage of frequent tweeters, it could suggest one of the following things: Either the majority of individual users is tweeting more frequently, or individual users are tweeting less frequently, and the appearance of an increase in community activity results from the only activity being from large commercial accounts of that collection period.

Results B:

#MayPac and the Observed Connection Between Hockey and Boxing

The interconnectedness of sports communities in online settings was seen. In string analyses of hockey related tweet text, terms like “basketball”, “baseball”, and ‘football’ appeared relatively frequently throughout all the datasets. During the period of May 2 – 3 2015, a significant increase in the frequency of other non-hockey sports related content appeared in the data. This is likely the result of the boxing championship match played between Floyd Mayweather Jr. and Manny Pacquiao on May 2 2015. (11) The hashtags “#may-pac” and “#betterfightsthanmaypac” were also popular, as many users conversed online about past famous sports matches. It can be clearly seen that a major increase in user tweeting frequency occurred during this period in Table 1, specifically in observations 7 and 8, where the frequency percentage sharply increases. These two observations also have identical unique tweets and users’ data. While the other data collection runs had varying request numbers, these two queries sent identical numbers and received identical val-

ues in these two columns. This strange data provides an insight into how Twitter’s public stream operates. It is possible that Twitter enforces certain limits and regulations as to how public API users can obtain data during periods of high traffic.

Discussion

Sources of Error

The data worked with for statistical analysis is reflective of the overall group it is sampled from; in this case, users tweeting about hockey. Therefore, we can make the assumption the data used was homogeneous with all data related to the keyword “hockey” on Twitter.

Another issue that was found was that many tweets contained ASCII characters unrecognisable by python, due to the prevalence of “emojis”. Emojis are pictographs with particular assigned Unicode values. (12) The presence of emojis in the text of a tweet renders that tweet’s data unreadable by python. Any tweets that contained these characters had to be ignored - meaning that a particular portion of user data was not collected. Users who use many emojis tend to be frequent users of emojis - this could have impacted our percentage values for users who tweeted frequently about hockey by completely excluding certain individuals from the analyses.

Conclusions

Twitter’s provided data is not complete, but it is still meaningful. There is significant information that can be collected by observing social media communities. Between 9.73% and 24.55% of users tweeting about hockey were discovered to be doing so very frequently. A correlation between different online sports communities was also found. The hockey community is moving away from traditional methods of hockey news media consumption and towards newer methods: online communities. This means individuals and organizations involved with hockey and other sports media should consider social media to be a valuable resource for information about sports communities in the future.

Future Work

The geographical location of users tweeting is significant when considering real time tweet data. Geographical location information about users was attempted to be collected for this paper’s analysis; however, it was found that a very small percent (< 1%) of users chose to include locational data with their tweets. This made an attempt to analyse this data for user trend information insignificant. In the future, the use of user profile location data rather than tweet

metadata location data will be collected and used for a deeper analysis of the Twitter hockey community.

Another step for the future is to utilize parallel processing methods. Processing data in parallel would allow for more extensive analysis as it would provide more processing power and decrease time requirements. It would also allow for bigger quantities of data

to be analyzed, making the results more accurate and meaningful. Incorporating parallel processing and other HPC related programming tools into the research would mean that data like 24 hour user-tweet-frequency trends could be discovered; this was data beyond the capabilities of the tools used for the research presented in this paper.

References

1. Twitter. About Twitter [Internet]. 2015 [cited 1 June 2015]. Available from: <https://about.twitter.com/company>
2. Huffington Post. Digital Media Is Driving Down NHL Ratings, But Rogers Blames The Messenger. [Internet]. 2015 [cited 1 June 2015];:1. Available from: http://www.huffingtonpost.ca/2015/02/03/nhl-ratings-rogers_n_6607718.html
3. van den Bosch A, Tjong Kim Sang E. Dealing with big data: The case of twitter. Computational Linguistics in the Netherlands Journal [Internet]. 2015 [cited 1 June 2015];. Available from: <http://hdl.handle.net/2066/121271>
4. Bakhshandeh R, Samadi M, Azimifar Z, Schaeffer J. Degrees of Separation in Social Network. AAAI [Internet]. 2015 [cited 1 June 2015];. Available from: <http://www.aaai.org/ocs/index.php/SOCS/SOCS11/paper/viewFile/4031%20degrees%20of%20separation%20in%20social%20networks/4352>
5. Rios M, Lin J. [Internet]. 2015 [cited 1 June 2015];. Available from: http://www.umiacs.umd.edu/~jimmylin/publications/Rios_Lin_2012.pdf
6. 12. Twitter. Twitter REST API [Internet]. 2015 [cited 1 June 2015]. Available from: <https://dev.twitter.com/rest/public>
7. pypi. twitter 1.16 [Internet]. 2015 [cited 1 June 2015]. Available from: <https://pypi.python.org/pypi/twitter/1.16.0>
8. OAuth. About OAuth [Internet]. 2015 [cited 1 June 2015]. Available from: <http://oauth.net/>
9. Python. Python [Internet]. 2015 [cited 1 June 2015]. Available from: <https://www.python.org/>
10. SAS. SAS® University Edition [Internet]. 2015 [cited 1 June 2015]. Available from: <http://support.sas.com/software/products/university-edition/>
11. Brown M. Mayweather-Pacquiao And The Death Of Boxing. Forbes [Internet]. 2015 [cited 1 June 2015];:1. Available from: <http://www.forbes.com/sites/maurybrown/2015/05/03/mayweather-pacquiao-and-the-death-of-boxing>
12. Unicode. Emoji [Internet]. 2015 [cited 1 June 2015]. Available from: http://unicode.org/faq/emoji_dingbats.html

REVIEW

by Dr. Alison Burnham,
Chief Information Officer at ScoreBig Inc.

The authors have selected a very interesting and challenging subject area to investigate. It is very impressive that they managed to collect their own data for the project, clean it and analyze it. It is more of an exploratory paper than a standard research paper that presents a hypothesis and the subsequent proof through research and experimentation. As an exploratory paper I do feel that a bit more structure would be helpful for the reader. Only the last paragraph of the introduction is really an introduction to the paper. The paragraph on big data is not terribly useful as it does not define the subject area or add to understanding. The section on Hockey would make a better following paragraph as it does provide some context as to why the authors thought this would be an interesting topic to explore. The section on Twitter should be in the body of the paper as it is important context to the data collected and examined. An explanation of the role of hashtags on Twitter and why they are useful for the analysis would be helpful.

The authors did a literature review for the paper which did give some idea of other work done in the field. It would be helpful if the review and rest of the background section were tied more to the paper contents themselves. The section labelled Twitter Application Program Interface (API) would be better labelled 'Data Collection Methodology'. The details given around the data being stripped of the tweet text component were a bit unclear – an illustration with a specific example tweet to show the work done and actual data used would help to clear this up.

The definition of frequent users is confusing. The paper states that it is those who, on average, tweeted more than once during the collection period. However, looking at the analysis done it is the users who tweeted more than once – but this could actually be made up exclusively of people who tweeted twice. It seems that the work to determine the actual frequency distribution of users was not done. This means that the conclusion that between 9.73 % and 24.55 % of users were discovered to be doing so very frequently is misleading: the percentages given are actually of tweets, not users.

It was very helpful to future users to mention the issues that arose during the data collection. The future work section was interesting and definitely gave room for further exploration on the subject area.