# Tweet Analysis for User Health Monitoring

Ranjitha Kashyap* and Ani Nahapetian‡

* Rubicon Project, Los Angeles, CA, USA
‡California State University, Northridge, CA, USA
ranjithabv@gmail.com
ani@csun.edu

***Abstract*—** Data analysis of social media postings can provide a wealth of information about the health of individual users, health across groups, and even access to healthy food choices in neighborhoods. In this paper, we analyze Twitter postings of 140 characters or less, known as tweets, to infer user health status over time. Tweets and in turn their users' health are scored according to semantic analysis, sentiment analysis, emoticon classification, meta-data analysis, and profiling over time. The purpose of the analysis includes individually targeted healthcare personalization, determining health disparities, discovering health access limitations, advertising, and public health monitoring. The approach is analyzed on over 12,000 tweets spanning as far back as 2010 for 10 classes of users active on Twitter.

***Keywords—Big Data, Semantic Analysis, Sentiment Analysis, Twitter.***

## I. INTRODUCTION

The data made public on social media sites, such as Twitter, provide a plethora of information about individuals, groups, and neighborhoods. Twitter is an online social networking medium, popular since 2006, where registered users share or post messages under 140 characters known as tweets. Tweets have been composed from daily conversations, updates, and critiques on news, movies, politics, life, etc. In this work we leverage this data to monitor general user health.

We collect and query tweets to carry out health analysis of users. We analyze individual tweets to infer user health, and how it changes over time. Our approach is composed of semantic analysis, sentiment analysis, emoticon classification, meta-data monitoring, and profiling over time. We validate our approach on an over 12,000 tweet data set that we have collected. We show that health of users can be inferred from their tweets, and thus the analysis can be used for targeted healthcare, determining health disparities, advertising, and even public health analysis.

For each user, each tweet of the user is passed through a series of analysis to determine individual tweet content's implications of user health.

The tweet health score is composed using sentiment analysis, i.e. positive tweets are determined to be healthier than negative tweets. The polarity of the tweet is also used to determine the mood of the user, and thus inferring happiness or sadness/anger and incorporating that into the tweet health score. Similarly, emoticons are used in this mood classification.

Semantic analysis involves mapping certain words in the tweet to scores. For example, the word 'exercise' is given a positive increment, while the phrase 'fast food' is given a negative increment.

The meta-data of the tweet (e.g. time and location of posting) is used to further tailor the score. The tweet data is aggregated over time to assign a user a running health score, which leverages a moving window of the tweet health score.

To verify that the health score calculated is indeed a valid marker of user health, we collected and analyzed a large set of Twitter data for various classes of users.

## II. RELATED WORK

Social Media has become a treasure trove of data for the analysis of various facets of life. Twitter specifically with textual messages, twitter handles signified with @ symbols, and topic areas signified with # (hashtags) has become a popular and accessible medium for exploration of user and group information and trends. Articles have explored using Twitter to determine news trends [1], as well publication site relevance [2]. Public health information has been gleaned from tweet analysis, including spread of disease [3][4][5] and depression risks across groups [6]. Personal health characteristics have also been examined in Twitter data, as we do in this work. Specifically health concerns related to insomnia [7] and to child birth [8] have been examined.

## III. APPROACH

### A. Overview

We analyzed the effectiveness of using a variety of information that is available from an individual tweet, to assign a health score to the tweet. The approach includes analysis of the content of the tweets, in terms of word choice, emoticon usage, and polarity of language. Meta data about the tweet, such as the time of day is also used to determine the healthy habits of the user. For example, tweeting in the middle of night has a negative effect on the health score. The high-level overview of the approach is presented in Figure 1.

The health score algorithm takes in the polarity data, healthy and unhealthy semantics, any previous health score and processes the text for the appearance of the listed keywords and makes a decision to assign a health score The health score for each user is stored in the database and is used again for analyzing the history of scores.



978-1-63190-014-3 © 2014 ICST
DOI 10.4108/icst.mobihealth.2014.257537

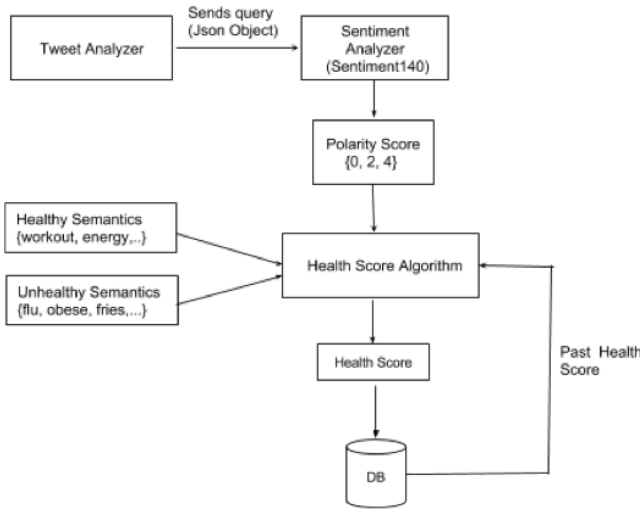


Fig. 1. Overview of the tweet data analysis process.

### B. Semantic Analysis

We developed a semantic analyzer for tweets. For example, if a user tweets more about unhealthy topics such as being sick, not well, lazy, McDonalds, French fries, the health score will be lower as compared to the user who tweets more about fitness, diet, fruits, vegetables, exercises or workout.

Semantics related to hundreds of healthy and unhealthy behaviors were collected and added to our semantic analyzer. Names of fast food restaurants, unhealthy habits such as tobacco or drug use, lists of diseases, and unhealthy foods were signified as 'unhealthy'; conversely lists of active pursuits, topics related to nutrition, and health foods were signified as 'healthy.'

### C. Sentiment Analysis

Sentiment analysis has been an important topic in natural language processing research, where the positive, neutral, or negative tone of text is determined. There are a number of sentiment analyzers available, and we used the open source Sentiment140 sentiment analysis tool designed specifically for Twitter data. For each tweet, it returns a polarity score 0 indicating negative, 4 indicating positive and 2 indicating neutral.

This semantic analysis was further augmented with the mapping of emoticon usage to certain polarity, as shown in Table 1.

TABLE I. MAPPING OF POLARITY SCORE TO INDIVIDUAL EMOTICONS

| **Emoticon-Polarity Mapping** | |
|---|---|
| ***Emoticon*** | ***Polarity*** |
| :-) :) :o) :] :3 :c) :D C: | Positive (4) |
| :-( :( :c :[ D8 D; D= DX v.v | Negative (0) |
| : | Neutral (2) |

### D. Meta-Data Usage

In addition to using the text of the tweet, we used meta-data relating to place and time of tweet to determine the healthiness of the user and the tweet. If the time of the tweet was between 10 PM to 6 AM, then the tweet health score was decremented. The location of the tweet was used to determine the appropriate time, given the time zone. If available, the geocode of a tweet can also be used to determine the healthiness of tweet. For example, a tweet from a fast food restaurant is decremented in terms of health score, while a tweet from a gym increments the score. The Google Places API was used to determine the type of establishment. Location types including gym, park, and restaurant can be obtained with a geocode (i.e. longitude and latitude).

### E. Tweet Health Trends

The general health of a user should not be swayed by one tweet. Moreover, there is some error, for example in the case of humor, in polarity and sentiment analysis. As a result, the history of a user's tweet health is used to calculate the user health. Three different calculations of user health score were analyzed. First approach assigns the user health score as the tweet health score of the most recent tweet, as shown in Equation 1. The health score of a tweet is calculated with the following formula, incorporating semantic, sentiment, and meta-analysis of the tweet.

Eq 1. $TweetHealthScore = TweetPolarity + HealthySentimentCount - UnhealthySentimentCount - UnhealthyMetaDataCount + HealthyMetaDataCount$

According to the first approach, the user health score at a given time, *t*, is then equivalent to the latest tweet health score, as given in Equation 2.

Eq 2. $UserHealth\ Score\ (t) = TweetHealthScore\ (t)$

The second and third options looked at a weighted average of the recent tweet and some or all previous tweets. In the second option, the health score of user is based on a weighted average of the current tweet and all past tweets health scores, as shown in equation 3.

Eq 3. $$UserHealthScore(t) = W \cdot TweetHealthScore(t) + (1 - W) \cdot \sum_{i=0}^{t-1} TweetHealthScore(i)$$

Equation 4 gives the third option where there is a sliding window, and so the last few tweets can temper a recent tweet. The variable $k$ is used to demonstrate the size of the sliding window. In both the second and third calculation, $W$ is the weighting given the most recent tweet. In our case we assigned W=0.8.

Eq 4. $$UserHealthScore = W \cdot TweetHealthScore(t) + (1 - W) \cdot (\sum_{i=t-k}^{t-1} TweetHealthScore(t-1))/k$$

## IV. EXPERIMENTAL ANALYSIS

### A. *Twitter API Interfacing*

The tweets from a user can be downloaded through REST API calls, with the data obtained in JSON format. The downloaded tweets are limited by count and ordered by time. The username is referred to as twitter handle. The collection of tweets is termed a timeline. The timeline seen with the API call is equivalent to the tweets seen in the user profile on twitter.com. It is required to paginate the timeline to calculate the time and text of every tweet. The REST API call GET statuses/user timeline returns a set of tweets identified by the username or twitter handle of the user.

Twitter4J, a Java library for accessing Twitter API, was used in our analysis. As shown in Figure 2, tweets were collected by REST API call using twitter handles. The data, returned in JSON format, was then parsed with our in-house Java parser. Of the JSON data returned for each Twitter handle, the tweet text, including punctuation, was extracted, along with the Tweet time, corrected for the time zone of the user, and geocode if available.

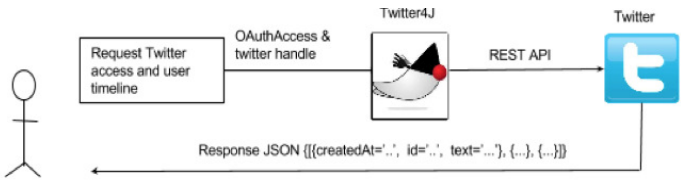


Fig. 2. Overview of the protocol for obtaining JSON data from a Twitter handle.

### B. *User Group Selection Approach*

To determine the validity of the approach, we took classes of users with strong association with either healthy or healthy characteristics and classified them according to their health score. We then examined how correctly each user, from each user group, was classified and if it matched their expected group. For example, a dietician was expected to be a 'healthy' user and so their health score over time was hypothesized to be healthy.

The classes of 'healthy' users include fitness gurus, dieticians, and physicians. Neutral users were selected from classes including IT industry leaders, popular celebrities, and general Twitter users. Additionally, individuals with commented or mentioned certain types of unhealthy behavior or diseases were also chosen. Table II provides the classes of users, along with their selection process. A total of 120 users were sample, with over 100 tweets per user.

TABLE II. MAPPING OF POLARITY SCORE TO INDIVIDUAL EMOTICONS

| **User Classes** | |
|---|---|
| ***Class*** | ***User selection process*** |
| Dieticians | List of top dieticians on social media. http://www.mamavation.com/2012/02/top-17-dietitians-you-can-count-on-in-social-media.html<br><br>List of influential dieticians. http://dietitians-online.blogspot.com/2010/12/top-100-influential-dietitians-on.html |
| Physicians | List of top physicians on Twitter. http://www.healthcareitnews.com/news/10-physicians-follow-twitter<br><br>List of top celebrity doctors on Twitter. http://msndegree.net/2010/top-25-celebrity-doctors-on-twitter-worth-following/ |
| Fitness Gurus | List of top fitness Gurus on Twitter. http://www.huffingtonpost.com/2012/11/01/twitter-fitness-experts_n_2038510.html |
| Twitter Celebrities | Top tweeters on Twitter. http://twittercounter.com/pages/100 |
| IT Professionals and Followers | Top tech people on twitter. http://www.businessinsider.com/the-best-tech-people-on-twitter-2013-7?op=1 |
| @BurgerKing Commentors | User search by @BurgerKing between 7:30am - 7:50am 03/16 |
| @KrispyKreme Commentors | User search by @KrispyKreme for the 03/16 |
| #postpartumdepression Mentioners | Included #postpartumdepression in any March 2014 post |
| #CrohnsDisease Mentioners | Included #CronhsDisease in any March 2014 post |
| General Users | Users selected according to March 2014 posting with a minimum of 90 tweets over the course of their Twitter usage. |

### C. *Experimental Results*

Table III provides the average user health score across different classes of users. Data for popular dieticians, doctors, fitness gurus was collected. People who had commented on the @KrispyKreme and @BurgerKing Twitter handles were chosen, as well as those who have mentioned #CrohnsDisease or #PostpartumDepression in their tweets. Tweets from celebrities, IT professionals, and the general Twitter population were also analyzed.

As shown in Table III, dieticians and fitness gurus had high health scores, while doctors perhaps with their frequent mention of disease names had lower health scores. All users sampled who had the @BurgerKing and/or @KrispyKreme had low health scores. Similarly, users who had #PostpartumDepression and/or #CrohnsDisease had low health scores. IT professionals had average health scores, while the general selection of users used in this experiment had pretty low health scores. Celebrities had fairly high health scores.

TABLE III. AVERAGE USER HEALTH SCORE ACROSS DIFFERENT CLASSES OF USERS

| Average User Health Score | | | | |
|---|---|---|---|---|
| **User Category** | **Estimation Approaches** | | | |
| | *Eq. 2* | *Eq. 3* | *Eq. 4 k= 1* | *Eq. 4 k=2* |
| Dietitians | 2.991429 | 2.94617 | 2.983143 | 2.98 |
| Doctors | 2.181538 | 2.167962 | 2.176923 | 2.174923 |
| Fitness Gurus | 2.658215 | 2.658366 | 2.652201 | 2.649531 |
| Tech&Enterpreneur | 2.3255 | 2.320208 | 2.3217 | 2.31885 |
| Celebreties | 2.623 | 2.621736 | 2.6184 | 2.6159 |
| General | 1.66534 | 1.671582 | 1.662933 | 1.661265 |
| @BurgerKing | 2.157224 | 2.173599 | 2.156 | 2.154784 |
| @KrispyKreme | 1.704082 | 1.699159 | 1.700408 | 1.698163 |
| #PostpartumDepression | 2.212245 | 2.211514 | 2.205306 | 2.203061 |
| #CronhsDisease | 1.838776 | 1.873195 | 1.836735 | 1.834694 |

TABLE IV. THE CLASSIFICATION OF USERS INTO 'HEALTH' 'UNHEALTHY' AND "NEUTRAL' CATEGORIES

| **User Class** | **Category** | | |
|---|---|---|---|
| | *Healthy* | *Neutral* | *Unhealthy* |
| Dietitians | 85.7% | 14.3% | 0% |
| Doctors | 46.1% | 23.1% | 30.8% |
| Fitness Gurus | 86.7% | 0% | 13.3% |
| Tech&Enterpreneur | 50% | 35% | 15% |
| Celebreties | 73.3% | 16.7% | 10% |
| General | 0% | 13.3% | 86.7% |
| @BurgerKing | 60% | 0% | 40% |
| @KrispyKreme | 0% | 20% | 80% |
| #PostpartumDepression | 40% | 0% | 60% |
| #CronhsDisease | 20% | 0% | 80% |

Table IV provides the classification of the average user health score for each of the 120 users examined in our analysis. The health score of the user assigned by their Eq. 2 user health score was used to classify them into one of three categories: health, neutral, unhealthy.

The health score assignment approach was further validated through human analysis. A total of 100 tweets, 50 tweets each from two users, was analyzed and compared with the calculated user health score calculated for the tweets. A famous dietician and a user who had referenced @KrispyKreme in a tweet were chosen. The difference between the health scores calculated with human inspection and with the algorithm is shown in Table V. As the results demonstrate, the human reading of the tweet and assignment of the health score is in line with the output of our tweet analysis algorithm.

TABLE V. VALIDATION RESULTS

| | **Average User Health Score** | |
|---|---|---|
| **User Category** | *Average Health Score {Eq.2}* | *Average Human Analysis Health Score {Eq.2}* |
| Dietitian | 3.06 | 2.84 |
| @KrispyKreme | 1.74 | 1.62 |

## V. CONCLUSION

Social media information is an important and large source of information about individuals. Twitter is one of the leading sources for social media information. In this work we analyze tweet data across a user's history to determine the user health. We demonstrate the effectiveness of the approach by using over 12000 tweets, across a time period of over two years, for 10 classes of users.

## REFERENCES


[1] Rong Lu, Zhiheng Xu, Yang Zhang, and Qing Yang. 2012. Life activity modeling of news event on twitter using energy function. In *Proceedings of the 16th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining - Volume Part II* (PAKDD'12).

[2] Roja Bandari, Sitaram Asury, Bernardo Huberman. 2012. The Pulse of News in Social Media: Forecasting Popularity In *Proceedings of the Sixth international conference on Web Blogs and Social Media*, (ICWSM'12).

[3] Sean Brennan, Adam Sadilek, and Henry Kautz. 2013. Towards understanding global spread of disease from everyday interpersonal interactions. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence* (IJCAI'13).

[4] Nattiya Kanhabua and Wolfgang Nejdl. 2013. Understanding the diversity of tweets in the time of outbreaks. In *Proceedings of the 22nd international conference on World Wide Web companion* (WWW '13 Companion).

[5] Xiang Ji, Soon Ae Chun, and James Geller. 2013. Monitoring Public Health Concerns Using Twitter Sentiment Classifications. In Proceedings of the 2013 IEEE International Conference on Healthcare Informatics (ICHI '13).

[6] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference* (WebSci '13).

[7] Sue Jamison-Powell, Conor Linehan, Laura Daley, Andrew Garbett, and Shaun Lawson. 2012. "I can't get no sleep": discussing #insomnia on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12).

[8] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Major life changes and behavioral markers in social media: case of childbirth. In *Proceedings of the 2013 conference on Computer supported cooperative work* (CSCW '13).