Big Data Analysis for User Health Tracking

A thesis submitted in partial fulfillment of the requirements
For the degree of Master of Science in
Software Engineering

By

Ranjitha Kashyap

May 2014
The thesis of Ranjitha Kashyap is approved:

_________________________________________  ____________
Dr. Taehyung Wang  Date

_________________________________________  ____________
Dr. Richard Covington  Date

_________________________________________  ____________
Dr. Ani Nahapetian, Chair  Date

California State University, Northridge
Dedication

I dedicate my dissertation work to my beloved parents

Ramā and Venkatesha Murthy and to

my husband Narayana Kashyap

for supporting me all the way!
I would like to express my deep gratitude to my advisor and committee chair Prof. Ani Nahapetian, for her continuous support without which the thesis wouldn’t have been possible to complete. Her suggestions, guidance, thorough knowledge and expertise helped me immensely in understanding and developing this thesis. I thank her immensely for her patience and generous time spent to guide me through the entire process.

I would also like to thank my committee members Prof. Richard Covington and Prof. Taehyung Wang for their support for reviewing and certifying the thesis. I also thank Prof. Robert Lingard for helping me for various reasons.

I am grateful to my manager at work Nader Salehi, for his support, encouragement and being considerate to pursue my studies along with a fulltime job as a Software Engineer at Rubicon Project.

Warmest thanks to my parents, family and friends for their moral inspirations. I thank my sister Rashmi Raghav who is very special to me and has never left my side.

Finally, I thank my dear husband Narayana Kashyap, for his continuous love and support throughout, and being by my side during difficult situations and encouraging me to pursue my graduate career. Many thanks to my lovely daughter Shriya for cheering me with her smiles and making me forget all the stress and difficulties during the journey.
List of Tables

TABLE 1: SAMPLE JSON RESPONSE RETURNED FROM A TIMELINE................................. 6
TABLE 2: TWEET PARSER RESULT ................................................................. 6
TABLE 3: SAMPLE SEMANTICS CONSIDERED DURING THE ANALYSIS......................... 7
TABLE 4: SENTIMENT140 API CALL AND OUTPUT .................................................. 11
TABLE 5: EMOJICONS AND POLARITY ...................................................................... 11
TABLE 6: USER SELECTION PROCESS FOR DIFFERENT CATEGORY OF CLASSES .......... 20
TABLE 7: AVERAGE USER HEALTH SCORE ACROSS DIFFERENT CLASSES OF USERS .... 22
TABLE 8: THE ACCURACY OF CLASSIFICATION OF USERS INTO HYPOTHESES ‘HEALTH’ ‘UNHEALTHY’ AND ‘NEUTRAL’ CATEGORIES ......................................................... 23
TABLE 9: VALIDATION RESULTS ........................................................................... 25
# Table of Contents

Signature Page .............................................................................................................. ii

Dedication .................................................................................................................. iii

Acknowledgement ....................................................................................................... iv

List of Tables ............................................................................................................... v

Abstract .................................................................................................................... viii

1. Introduction............................................................................................................. 1

2. Related Work ....................................................................................................... 4

3. Approach............................................................................................................... 5
   3.1 Overview ......................................................................................................... 5
   3.2 Collecting Data ............................................................................................... 5
   3.3 Parsing Data .................................................................................................. 7
   3.4 Semantic Analysis ......................................................................................... 8
   3.5 Sentiment Analysis ....................................................................................... 9
   3.6 Database Schema ......................................................................................... 12
   3.7 Data Flow .................................................................................................... 13
   3.8 Meta-Data Usage .......................................................................................... 15
   3.9 Tweet Health Trends .................................................................................... 15
   3.10 Graphical Representation ......................................................................... 17
4. Experimental Analysis ........................................................................................................ 18
   4.1 Twitter API Interfacing .................................................................................................. 18
   4.2 User Group Selection Approach .................................................................................. 19
   4.3 Experimental Results .................................................................................................. 20
   4.4 Validation Results ....................................................................................................... 20
5. Conclusion ....................................................................................................................... 26
Abstract

Big Data Analysis for User Health Tracking

By

Ranjitha Kashyap

Master of Science in Software Engineering

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 thesis, I analyze the messages posted in Twitter that are 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.

With this user health scores classification approach, we can say approximately 65%-70% of the users can be accurately classified as healthy or unhealthy.
1. Introduction

Not many people enjoy talking about health and fitness, especially when it concerns their own health problems. But with the advent of various social media avenues like Twitter, Facebook, etc., people have started sharing in the public domain expressing feelings, reporting and updating activities and whereabouts. 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 I leverage this data to monitor general user health.

I collect and query tweets to carry out health analysis of users and analyze individual tweets to infer user health, and how it changes over time. This approach is composed of semantic analysis, sentiment analysis, emoticon classification, meta-data monitoring, and profiling over time. This approach is validated on an over 12,000-tweet data set that has been collected and shows 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.

The health analysis is classified as Data Collection, Data Parser, Sentimental Analysis and Semantic Analysis to derive the health score as shown in Figure 1. I have explained the details in Chapter 3.
For a user, every tweet from the user is passed through a series of analysis to determine individual tweet content 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 word ‘insomnia’ 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; I carried out the following experiment with a large set of Twitter data. Classes of users with characteristically healthy or unhealthy behaviors were selected. (For example, I classified users as ‘unhealthy’ if they use “unhealthy” hashtags such as #BurgerKing. Famous dieticians were classified as ‘healthy.’) Their tweets were then analyzed and a user health score was generated. This score was then algorithmically used to classify each user as healthy, unhealthy, or neutral. The original user class was then compared to determine the validity of the classification.

The experimental analysis demonstrates the high accuracy with which users can be classified into their expected groups, thus supporting the inference that our tweet analysis approach can accurately determine user health in many situations.
2. 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 I do in this work. Specifically health concerns related to insomnia [7] and to child birth [8] have been examined.
3. Approach

3.1 Overview

In this thesis, the effectiveness of using a variety of information that is available from an individual tweet is analyzed and assigned a health score.

The approach includes querying Twitter for tweets, parsing the response for tweets and analyzing 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 as compared to the one tweeting during normal day.

3.2 Collecting Data

It is required to get authorization for accessing the user data. The AccessToken returned from Twitter is stored in a configuration file for future usage.

Authorized requests with user handle are sent through REST API and tweets are collected as a timeline. The timeline returned is the correspondent of the data seen on a user’s profile in www.twitter.com. The GET statuses/user_timeline API returns up to 3200 tweets per user.

```java
statuses = twitter.getUserTimeline(userHandle, page.count(100));
array = getJson(statuses);
page.setPage(page.getPage() + 1);
jsonObj.put("data", array);
```
The method getUserTimeline() returns the recent statuses posted by the given user. It also takes paging information, which controls pagination i.e., here it returns 100 tweets from a given page. The getters and setters page.setPage() and page.getPage() can be used to paginate and retrieve historical tweets. The timeline Statuses are transformed into a JSON as shown in Table 1.

| createdAt=M| id=440517009038528512, text=Good morning, god loves you http://t.co/Q4GaHOu5P0, source=Twitter for iPhone&lt;/a&gt;, isTruncated=false, inReplyToStatusId=1, inReplyToUserId=-1, isFavorited=false, isRetweeted=false, favoriteCount=45260, inReplyToScreenName='null', geoLocation=null, place=null, retweetCount=37111, isPossiblySensitive=false, isoLanguageCode=null, contributorsIDs=[J@3288df60, retweetedStatus=null, userMentionEntities=[], urlEntities=[], hashtagEntities=[], mediaURL=https://pbs.twimg.com/media/Bh0H4rjCMAA5maY.jpg, mediaURLHttp=htt|
3.3 Parsing Data

A Tweet Parser is designed to filter the datetime, text and id from every tweet. The JSON returned from the REST API is fed to a Tweet Parser as shown in Figure 2., and the corresponding pseudo code in Table 2. It removes the hashtags, re-tweets, stop words[13] for easy analysis. The parser generates JSON object with id, text and date fields.

```java
//date format
SimpleDateFormat sdf = new SimpleDateFormat("yyyy-MM-dd");
JSONArray jArray = new JSONArray();
JSONObject jObject = null;
for (Status status : <ResponseList>statuses) {
    jObject = new JSONObject();
    jObject.put("date", sdf.format(dateFormat.parse(status.getCreatedAt())));
    jObject.put("text", status.getText());
    jObject.put("id", status.getId());
    jArray.put(jObject);
}
```

**Figure 2: Tweet Parser**

**Table 2: Tweet Parser Result**

```
Tweet Object: {"data":[{"text":"I see all your comments. Keep strong. Be the best that you can be. Much love."},
3.4 Semantic Analysis

A Semantic analyzer was developed for the 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.” Sample list is shown in Table 3.

<table>
<thead>
<tr>
<th>Healthy Semantics</th>
<th>Unhealthy Semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>excercise</td>
<td>fries</td>
</tr>
<tr>
<td>workout</td>
<td>flu</td>
</tr>
<tr>
<td>happy</td>
<td>infections</td>
</tr>
<tr>
<td>meditation</td>
<td>depression</td>
</tr>
<tr>
<td>diet</td>
<td>smoking</td>
</tr>
<tr>
<td>yoga</td>
<td>obese</td>
</tr>
<tr>
<td>healthy</td>
<td>stress</td>
</tr>
<tr>
<td>energy</td>
<td>drugs</td>
</tr>
<tr>
<td>calories</td>
<td>diseases</td>
</tr>
</tbody>
</table>
3.5 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. It determines the demeanor of the user with respect to a subject of interest by evaluating the sentiment and emotional state of the tweet. There are a number of sentiment analyzers available, and we used the open source Sentiment140 sentiment analysis tool designed specifically for Twitter data. There are two-classification services API that takes JSON as input.

- Simple Classification Service: Individual tweet messages are passed via HTTP GET requests. The API call looks like below:

  http://www.sentiment140.com/api/classify

- Bulk Classification Service: The JSON containing several tweets is passed and receives the response in bulk. This API can service thousands of tweets and get a single response. I use this service in my analysis to get the polarity score of the collected tweets. The requests are sent via HTTP POST.

  http://www.sentiment140.com/api/bulkClassifyJson
The JSON returned from Tweet Parser is sent to Sentiment analyzer. For each tweet, it returns a polarity score 0 indicating negative, 4 indicating positive and 2 indicating neutral. The response from the Sentiment Analyzer for the Table 2 data is shown in Table 4.
This sentiment analysis was further augmented with the mapping of emoticon usage to certain polarity, as shown in Table 5.

**Table 5: Emoticons and Polarity**

<table>
<thead>
<tr>
<th>Emoticons</th>
<th>Polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>:-) :) :o) :] :3 :e):D C:</td>
<td>Positive (4)</td>
</tr>
<tr>
<td>:-( (; :e :[ D8 D; D= DX v.v</td>
<td>Negative (0)</td>
</tr>
<tr>
<td>:</td>
<td>Neutral (2)</td>
</tr>
</tbody>
</table>
3.6 Database Schema

The health score results with the calculated polarity are stored in the MYSQL database for further analysis.

- **Create Table**

```
CREATE TABLE Health_Score (tweet_id INT (20) NOT NULL, user_handle VARCHAR (25) NOT NULL, date DATETIME NOT NULL, sentiment_polarity DOUBLE (5,2) NOT NULL, health_score DOUBLE (5,2) NOT NULL, comments VARCHAR (100) DEFAULT NULL, PRIMARY KEY (ID))
```

- **Store Results**

The sentiment polarity and health scores calculated are stored in Health_Score table with the latest tweet date for further analysis of the history.

```
"INSERT INTO Health_Score VALUES (?, ?, ?, ?, ?)"
```

- **Previous Results**

The previous polarity, id and health score results are fetched from Health_Score table for the current analysis and a score is generated accordingly. If it returns multiple rows, the latest row is fetched. The tweet_id is used as a reference to fetch the most recent tweets i.e. the API returns results with an ID larger than the specified tweet_id.

```
"SELECT tweet_id, sentiment_polarity, health_score "
+ "FROM Health_Score " +
"WHERE user_handle = ? and date > ?"
```
3.7 Data Flow

As shown in Figure 3, Data Collector collects data from Twitter via REST API call and sends it to the Data Parse to separate the required fields in required format. It reviews the tweet to remove any stop words like (a, an, of, or, the, to, etc..), removes the url in the tweet by looking for the http regex, removes the hash tag symbol (#), user references with @ symbol and numerals.

Data Parser generates a JSON with date, text and id fields and sends it to Sentiment Analyzer (Sentiment140) via HTTP POST which then analyzes the data and generates a bulk response with polarity score of {0,2,4} which indicates negative, neutral and positive attitude of the user on the topic of interest respectively. The polarity score is assigned for every tweet message and the output generated is another JSON.

Several texts and words were collected, a few listed below for sentiment analysis and categorized into different groups like diseases, healthy habits, unhealthy habits and fast-food chains.

Diseases: arthritis, cancer, pneumonia, ulcer, cyst, infection

Healthy Habits: workout, exercise, meditation, yoga, organic, healthy

Unhealthy Habits: depression, stress, snoring, violent, refined, processed

Fast-Food Chains: McDonald, Burger King, taco bell, KFC
The health score analyzer/algorithm takes in the polarity data, healthy semantics and unhealthy semantics (discussed in section 3.4), 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.
3.8 Meta-Data Usage

In addition to using the text of the tweet, I used meta-data relating to time 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 “createdAt” field of the tweet was used to determine the appropriate time, given the time zone.

3.9 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 data analysis of the tweet.

\[
\text{TweetHealthScore} = \text{TweetPolarity} + \text{HealthySentimentCount} - \text{UnhealthySentimentCount} - \text{UnhealthyMetaDataCount} + \text{HealthyMetaDataCount}
\]

According to the first approach, the user health score is then equivalent to the tweet health score, as given in Equation 2.

\[
\text{UserHealth Score (t)} = \text{TweetHealthScore (t)}
\]
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. \( UserHealthScore(t) = W \times TweetHealthScore(t) + (1 - W) \times \sum_{i=0}^{t-1} TweetHealthScore(i) \)

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 this case I assigned \( W=0.8 \).

Eq 4. \( UserHealthScore = W \times TweetHealthScore(t) + (1 - W) \times \left( \sum_{i=t-k}^{t-1} TweetHealthScore(t - 1) \right) / k \)
3.10 Graphical Representation

The Graph 1 shows the health score for a span of time from Dec 2013 and Graph 2 show the average results during that period.

**Health Score Results with Trendline**

**Graph 1:** Sentiment polarity and Twitter health score for all the tweets during the period.

**Graph 2:** Sentiment polarity and Twitter health score for an average set of tweets.
4. Experimental Analysis

4.1 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. The page count is also passed to retrieve the specified number of tweets per page and max id is passed to get only the most recent tweets whose id is greater than the value passed.

Twitter4J, a Java library for accessing Twitter API, was used in the analysis for interfacing with the Twitter. The interface for Twitter Access is as shown below

```java
public interface ITwitterAccess {
    public void getConnection() throws FileNotFoundException, IOException;
    public JSONObject getUserTimeline(String userHandle) throws TwitterException, ParseException, JSONException, FileNotFoundException, IOException;
    public String parseCreatedDate(Date createdDate) throws ParseException;
}
```
As shown in Figure 4, tweets are collected via REST API call using twitter handles. The data, returned in JSON format, was then parsed with the in-house Java parser. Of the JSON data returned for each Twitter handle, the tweet text, unique tweet identification id, was extracted, along with the Tweet time.

![Diagram](image_url)

**Figure 4: Overview of the Protocol for Obtaining JSON Data from a Twitter Handle**

4.2 User Group Selection Approach

To determine the validity of the approach, I took classes of users with strong association with either healthy or healthy characteristics and classified them according to their health score. I then examined how correctly each user, from each user group, was classified and if it matched their original user group. For example, a dietician was determined 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 6
provides the classes of users, along with their selection process. A total of 120 users were sample, with over 100 tweets per user.

<table>
<thead>
<tr>
<th>User Classes</th>
<th>User selection process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dieticians</td>
<td>List of top dieticians on social media.</td>
</tr>
<tr>
<td></td>
<td>List of influential dieticians. <a href="http://dietitians-online.blogspot.com/2010/12/top-100-influential-dietitians-on.html">http://dietitians-online.blogspot.com/2010/12/top-100-influential-dietitians-on.html</a></td>
</tr>
<tr>
<td>Physicians</td>
<td>List of top physicians on Twitter.</td>
</tr>
<tr>
<td></td>
<td>List of top celebrity doctors on Twitter.</td>
</tr>
<tr>
<td>Fitness Gurus</td>
<td>List of top fitness Gurus on Twitter.</td>
</tr>
<tr>
<td>Twitter Celebrities</td>
<td>Top tweeters on Twitter.</td>
</tr>
<tr>
<td></td>
<td><a href="http://twittercounter.com/pages/100">http://twittercounter.com/pages/100</a></td>
</tr>
<tr>
<td>IT Professionals and Followers</td>
<td>Top tech people on twitter.</td>
</tr>
<tr>
<td>@BurgerKing Commentors</td>
<td>User search by @BurgerKing between 7:30am - 7:50am 03/16</td>
</tr>
<tr>
<td>@KrispyKreme Commentors</td>
<td>User search by @KrispyKreme for the 03/16</td>
</tr>
<tr>
<td>#postpartumdepression Mentioners</td>
<td>Included #postpartumdepression in any March 2014 post</td>
</tr>
<tr>
<td>#CrohnsDisease Mentioners</td>
<td>Included #CronhsDisease in any March 2014 post</td>
</tr>
<tr>
<td>General Users</td>
<td>Users selected according to March 2014 posting with a minimum of 90 tweets over the course of their Twitter usage.</td>
</tr>
</tbody>
</table>
4.3 Experimental Results

Table 7 provides the average user health score across different classes of users. Dieticians, doctors, fitness gurus were chosen to demonstrate the ‘healthy’ class of users. People who had commented on the @KrispyKream and @BurgerKing Twitter handles were chosen to represent the ‘unhealthy’ class of users, along with those who has mentioned #CrohnsDisease or #PostpartumDepression in their tweets.

Celebrities, IT professionals, and the general twitter population were selected to demonstrate the average health value, as a baseline.

As shown in Table 7, 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.

As expected 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 7. Average User Health Score across Different Classes of Users

<table>
<thead>
<tr>
<th>User Category</th>
<th>Estimation Approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Eq. 2</td>
</tr>
<tr>
<td>Dietitians</td>
<td>2.991429</td>
</tr>
<tr>
<td>Doctors</td>
<td>2.181538</td>
</tr>
<tr>
<td>Fitness Gurus</td>
<td>2.658215</td>
</tr>
<tr>
<td>Tech&amp;Enterpreneur</td>
<td>2.3255</td>
</tr>
<tr>
<td>Celebrities</td>
<td>2.623</td>
</tr>
<tr>
<td>General</td>
<td>1.66534</td>
</tr>
<tr>
<td>@BurgerKing</td>
<td>2.157224</td>
</tr>
<tr>
<td>@KrispyKreme</td>
<td>1.704082</td>
</tr>
<tr>
<td>#PostpartumDepression</td>
<td>2.212245</td>
</tr>
<tr>
<td>#CronhsDisease</td>
<td>1.838776</td>
</tr>
</tbody>
</table>

Table 8 provides the classification error for the average user health score for each of the 120 users examined in our analysis and depicted in the Graph 3. Each user was assigned an expected health category, according to how the user’s Twitter handle was added to the list. This categorization was compared with the determined user health scores classification. 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 expected classification was then compared with these results.

**TABLE 8. THE ACCURACY OF CLASSIFICATION OF USERS INTO HYPOTHESISED ‘HEALTH’ ‘UNHEALTHY’ AND ‘NEUTRAL’ CATEGORIES**

<table>
<thead>
<tr>
<th>User Class</th>
<th>Expected Category</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Healthy</td>
</tr>
<tr>
<td>Dietitians</td>
<td>Healthy</td>
<td>6/7</td>
</tr>
<tr>
<td>Doctors</td>
<td>Healthy</td>
<td>6/13</td>
</tr>
<tr>
<td>Fitness Gurus</td>
<td>Healthy</td>
<td>13/15</td>
</tr>
<tr>
<td>Tech&amp;Enterpreneur</td>
<td>Neutral</td>
<td>10/20</td>
</tr>
<tr>
<td>Celebrities</td>
<td>Neutral</td>
<td>22/30</td>
</tr>
<tr>
<td>General</td>
<td>Neutral</td>
<td>0/15</td>
</tr>
<tr>
<td>@BurgerKing</td>
<td>Unhealthy</td>
<td>3/5</td>
</tr>
<tr>
<td>@KrispyKreme</td>
<td>Unhealthy</td>
<td>0/5</td>
</tr>
<tr>
<td>#PostpartumDepressio</td>
<td>Unhealthy</td>
<td>2/5</td>
</tr>
<tr>
<td>#CronhsDisease</td>
<td>Unhealthy</td>
<td>1/5</td>
</tr>
</tbody>
</table>
In this analysis, it showed over 71% ‘healthy’ users actually had ‘healthy’ user health scores. 65% of ‘unhealthy’ users had ‘unhealthy’ health scores. The less correlated ‘unhealthy’ scores can be due to the fact that the selection of the ‘unhealthy’ users was weaker than that of the ‘healthy’ users, who make a career out of writing about healthy habits.
4.4 Validation Approach

As a second approach to validate the health score algorithm with human analysis, I gathered a user with hashtag krispykreme as an unhealthy user and a dietitian as a healthy user and collected 50 tweets for both the users. The average difference between the human inspection and algorithm result is as shown in Table 9.

\[ \Delta = \text{Average Health Score - Average Human Analysis Health Score} \]

<table>
<thead>
<tr>
<th>User Category</th>
<th>Validation Approach</th>
<th>Average Health Score</th>
<th>Average Human Analysis Health Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Δ Healthy</td>
<td>Δ Unhealthy</td>
</tr>
<tr>
<td>Dietitians</td>
<td></td>
<td>0.32%</td>
<td>0.08%</td>
</tr>
<tr>
<td>@KrispyKreme</td>
<td></td>
<td>0.08%</td>
<td>0.16%</td>
</tr>
</tbody>
</table>
5. 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 I analyze tweet data across a user’s history to determine the user health.

The effectiveness of the approach is demonstrated by using over 12000 tweets, across a time period of over two years, for 10 classes of users. And also, demonstrate that we can classify ‘healthy’ users in over 70% of cases and ‘unhealthy’ users in 65% of cases.

Bigger hope is to save the lives by predicting the user health progress by detecting a symptom of a deadly disease, motivating people in depression, targeting relevant ads and information for their well being.

This can be used to generate reports and dashboards, which would provide some awareness in the user and improve their health with decrease cost of care.
References


