A Practical Approach for Content Mining of Tweets

Sunmoo Yoon, RN, PhD, Noémie Elhadad, PhD, Suzanne Bakken, RN, PhD

Abstract: Use of data generated through social media for health studies is gradually increasing. Twitter is a short-text message system developed 6 years ago, now with more than 100 million users generating over 300 million Tweets every day. Twitter may be used to gain real-world insights to promote healthy behaviors. The purposes of this paper are to describe a practical approach to analyzing Tweet contents and to illustrate an application of the approach to the topic of physical activity. The approach includes five steps: (1) selecting keywords to gather an initial set of Tweets to analyze; (2) importing data; (3) preparing data; (4) analyzing data (topic, sentiment, and ecologic context); and (5) interpreting data. The steps are implemented using tools that are publically available and free of charge and designed for use by researchers with limited programming skills. Content mining of Tweets can contribute to addressing challenges in health behavior research.
(Am J Prev Med 2013;45(1):122–129)

Introduction

Use of data generated through social media for health studies is gradually increasing. Because of its growing pervasiveness, social media have the potential to support the collection and analysis of health-related data in real time in the real world.[1,2] One social medium that has shown exponential growth is Twitter, a short-message micro-blogging service system with a "what's happening" prompt and an allowance of 140 characters per "Tweet." People use Twitter to share their momentary feelings, observations, activities, and daily lives with others.

Twitter usage statistics report 140 million active users generating 340 million Tweets on average per day as of March 2012.[3] Tweet content has been used not only as a rapid and inexpensive way to glimpse public opinion in general,[4] but also within the health domain for purposes such as monitoring diseases[5,6] and delivering health care.[7,8] Despite the growing attention to analyzing user-generated content from social media, most health researchers have little knowledge about how to apply content-mining methods.

Applying content-mining methods to social media in order to study health behaviors is important because gaining a full understanding of such behaviors has been difficult because of their complexity. Tweets are a source of real-time, real-world data about health behaviors, and they share characteristics with traditional methods of ecologic momentary assessment that simultaneously capture a behavior and allow individuals to report their current activity, location, and social surroundings at any particular moment.[9] However, unlike most ecologic momentary assessment methods, Tweet contents are not dependent on a specific intermittent stimulus to the intended respondent. Thus, Tweets may represent more-naturalistic content and have the additional advantage of being available in large volume. This paper describes a practical approach to analyzing Tweet contents and illustrates application of the approach to physical activity, a substantial and challenging public health issue and a health behavior of interest to many researchers.[10]

Content Mining

Web mining focuses on the discovery of meaningful knowledge from data such as online mailing lists, blogs, and social media and includes analysis of structure, usage, and content.[11] Web content mining aims to extract and analyze useful information (e.g., opinions, sentiment, main topics) from web content by applying techniques from multidisciplinary fields including data mining, machine learning, natural-language processing, information retrieval, and statistics. Following the traditional framework of general data mining, a typical content-mining process[11] includes preparing data so they can be imported and read in data-mining software, reducing the dimensionality of data, applying classic data-mining techniques, and terminating or iterating the process according to interpretation. Dodds and Danforth[12] and Kleinberg[13] have reported how to mine

From the School of Nursing (Yoon, Bakken), the Department of Biomedical Informatics (Yoon, Elhadad, Bakken), Columbia University, New York, New York

Address correspondence to: Sunmoo Yoon, RN, PhD, School of Nursing, Department of Biomedical Informatics, 630 West 168th Street, Mail Code 6, New York NY 10032. E-mail: sy2102@columbia.edu.

0749-3797/$36.00

http://dx.doi.org/10.1016/j.amepre.2013.02.025

SIDEBAR

Obesity Example

Step 1: Select keywords: *obesity*, *overweight*, *body mass index*, *body fat*, *anti-obesity drug*, *appetite*
Step 2: Import data: specify period of time (week of March 3, 2013) for search of Twitter database and retrieval of Tweets using open-source NodeXL.
Step 3: Prepare data: (1) clean data to remove extraneous symbols and words (e.g., " , ; ' symbols, urls); and (2) use Weka to transform text into N-grams (e.g., obesity=unigram, obesity prevention=bigram, childhood obesity prevention=trigram).
Step 4: Analyze data: (1) calculate frequency vectors to create term-Tweet frequency table resulting in identification of frequently occurring terms (e.g., marijuana smoking, sugar, milk, and genetics); (2) compare terms across Tweet corpuses using chi-square tests (e.g., BMI corpus distinct terms include *measure*, *circumference*, *supplements*, *height*, and *calculator*, and obesity corpus distinct terms include *family*, *marijuana*, *smoking*, and *milk*; (3) apply sentiment analysis to identify topics associated with positive (e.g., metabolism, lifestyle) or negative (e.g., body fat, appetite) sentiment.
Step 5: Interpret data: (1) terminate; or (2) return to Step 2 for importing Tweets in a different month (e.g., national childhood obesity awareness month).

social media content using natural-language processing. However, unlike the practical approach presented here, those methods require understanding of sophisticated large-scale computing methods.

Steps of Tweet Content Mining Applied to Physical Activity

The practical steps of Tweet content mining are illustrated in Figure 1 (additionally, the Sidebar provides a condensed example of these steps to illustrate how the process could be used for the content area of obesity): (1) selecting keywords; (2) importing data; (3) preparing data; (4) analyzing data; and (5) interpreting data (Appendix A, available online at www.ajpmonline.org).[14] Preliminary steps are to obtain review for human subjects research and to identify specific research questions. For the physical activity example, the analysis met federal criteria for Human Subjects Exemption. Research questions included (1) What is the content of Tweets that mention specific physical activities? (2) Does Tweet content vary by specific physical activity? (3) Does Tweet content change over time? (4) What proportion of Tweets that mention specific physical activities express positive as compared to negative sentiments? and (5) How is context expressed in Tweets that mention specific physical activities?

Step 1. Selecting Representative Keywords

Initially, domain experts analyze a concept to be studied and identify appropriate key terms and phrases (e.g., synonyms/morphologic variants) for extraction of Tweets to create the analytic corpus. For physical activity, 17 diverse activities (e.g., aerobics, jogging, swimming), defined by "mypyramid.gov," were selected as key phrases.

Step 2. Importing Tweets

In the second step, Tweets are imported by searching the selected terms and phrases via Tweet import tools. For the current study, NodeXL was used, a publicly available open-source Microsoft Excel template for creating a Tweet data corpus. Unlike other tools, NodeXL offers convenient data manipulation for users without programming skills by facilitating searching Twitter for public Tweets and importing the Tweets as an Excel file. Up to 1000 Tweets for each activity were randomly imported via NodeXL for each of 12 weeks from Week 1 of March to Week 4 of May 2010 (total 174,394 Tweets) to create the initial analytic corpus.

Step 3. Preparing Tweets

Text cleaning. The preparation step includes text cleaning, text transformation to generate attributes, and reduction of dimensionality through attribute selection. Compared to other genres of documents such as news

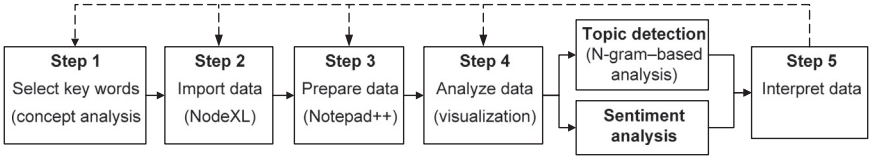

Figure 1. Steps of content mining

stories or traditional webpages, the linguistic characteristics of Tweets are noisy, due to use of a variety of languages, format/signs, unstructured grammar, and unofficial abbreviations. Thus, it is important to remove all nonstandard characters or special characters that would hinder the use of content-mining tools. For instance, Weka 3.7.1, a popular data-mining tool (www.cs.waikato.ac.nz/ml/weka/), has a special use for the typographic symbol for quotation ("). Consequently, quotation marks must be removed from Tweets prior to analysis.

Other examples of typologic symbols that must be removed to ensure readability as CSV (comma-separated value) data for import into Weka are apostrophes ('); single quotes (‘); commas (,); and semicolons (;). Other symbols (e.g., ^, @, space, line feed) and unnecessary letters (e.g., www, http://*) also can be removed in this step. Symbols can be removed via an open-source code editor (e.g., Notepad++) or Microsoft® Word, with the replace function.

Text transformation (generate attribute). In the text-transformation step, Tweet contents are represented as a vector of features. The simplest features are the individual words composing the Tweet, and the associated feature values are the frequencies of word occurrences in the Tweet.[15] Other examples of features are N-grams and numeric features, such as the length of a Tweet.

N-grams were used in the current physical activity example. An N-gram is a subsequence of N items in a given sequence, where the sequence items or grams can be anything from characters to words. Given the phrase "physical activity burns," there are three unigrams ("physical," "activity," and "burns"); two bigrams ("physical activity" and "activity burns"); and one trigram ("physical activity burns").

Because the use of a combination of unigrams, bigrams, and trigrams is reported as effective methods,[16] the authors used the combination of unigrams, bigrams, and trigrams in the physical activity content analysis. The Tweet term-frequency dictionary was computed by the N-gram method from the corpus of 174,394 publically available Tweets that were imported from twitter.com/ via NodeXL. Each unigram, bigram, and trigram generated is considered one attribute.

Attribute selection (dimensionality reduction). Given that many attributes can be generated from a single sentence (e.g., *I am swimming with my sister* generates 15 attributes), and that it is more difficult to algorithmically process large data sets with high dimensionality,[11] it is typically necessary to reduce the dimensionality of a data set by decreasing the number of attributes. The authors applied two methods for reducing dimensionality: removal of stop words and stemming. Stop words (i.e., words that are very common in any document such as "the" and "to") have little informational content and are unlikely to help with text mining and can be removed.[11,17] Individual stop words (e.g., of, a, an, for) were removed from the dictionary to reduce dimensionality, resulting in a 15%–20% data reduction among various physical activities, but stop words within phrases (e.g., "for" in "physical activity for," or "of" in "physical activity of") were retained. Stop words were removed using the stop-word removal function in Weka.[18] The number of features ranged from 500 to 1000 per 1000 Tweets per physical activity. Increasing the Java Heap size allows Weka to handle larger number of features than are required for the physical activity analysis.

Stemming reduces dimensionality by identifying a word root and removing suffices and prefixes from different word forms. For example, the two words "exercising" and "exercised" can be stemmed to "exercise" and, thus, the two features can be merged into one attribute representing all the features with "exercise" as a stem. Stemming has the disadvantage of discarding linguistic information.[11,19] However, for the physical activity example, stemming was necessary so that the dictionary was not too large to be processed by Weka. There are three different types of stemming algorithms: affix removing, statistical, and mixed. Porter's algorithm, an affix-removal approach, is the most popular and standard approach because it is concise and efficient[20] and was applied through "snowball stemmers" within Weka.[21]

Step 4. Analyzing Data

In this step, preprocessed data are analyzed in order to discover patterns such as "hot topics" or sentiments. Three approaches were used to discover patterns in the physical activity Tweet corpus: topic detection, sentiment analysis, and categorization of ecologic momentary context.

Topic detection. To detect and summarize topics, classic data-mining techniques are used on the structured data matrix that resulted from the previous stages. These include descriptive statistics (frequency counts); visualization; classification; and clustering.[14] The frequency of terms can be compared across diverse physical activities with vector values. Frequently occurring terms can be visualized with two-dimensional (2D) graphs and with three-dimensional (3D) motion charts to visualize data trends over time (www.excelcharts.com/blog/google-motion-chart-api-visualization-population-trends/). A keyword-only approach will create noise, and the level of noise may vary according to the concept studied. For example, the Tweet corpus of "run" contains more noise than "swim," "weight lifting," or "chopping wood."

Table 1. A sample of physical activity Tweets term-frequency dictionary from the corpus of 174,394 Tweets containing 31,489 terms

Terms	Weight	Day	Good	Time	Obesity	Study
aerobics	0.0229	0.0349	0.0331	0.0335	0	0.0026
bicycling	0.0062	0.0283	0.0182	0.0215	0	0.0024
chop wood	0	0.0705	0.0529	0.0353	0	0
dancing	0.0023	0.0255	0.0284	0.0242	0	0.0009
do yard work	0	0.1352	0.0526	0.072	0	0.0014
hiking	0.0039	0.0572	0.0343	0.0305	0	0.0013
jogging	0.0118	0.0251	0.0329	0.0308	0	0
just got back from gym	0.0052	0.0441	0.1159	0.0635	0	0.0038
light workout	0.0339	0.0928	0.1105	0.0715	0	0.0028
physical activity	0.0662	0.0548	0.0404	0.0381	0.037	0.0353
play golf	0	0.0077	0.0072	0.0135	0	0
play basketball	0.0011	0.001	0.0013	0.0033	0	0
ran miles	0.004	0.0573	0.2476	0.057	0	0
swim	0.0016	0.0448	0.0365	0.0402	0	0.0013
walk fast	0.0214	0.0271	0.0256	0.0384	0	0
weight lifting	0.9614	0.0333	0.0424	0.0363	0.0014	0.0034
weight training	0.9577	0.0496	0.0351	0.0256	0	0.0096

Classification is used to predict which category a new observation belongs to among a set of predefined categories. In contrast, clustering is an unsupervised learning approach, used to aggregate data into groups that are meaningful or useful.[22] Because a priori categorization of Tweet contents existed, the clustering data-mining technique was tested in the current study to summarize physical activity Tweet contents. However, because the technique was overly reliant on investigator interpretation for the corpus, only descriptive statistics (frequency counts) and visualization using 2D and 3D motion charts were used to summarize the Tweet corpus. Chi-square statistics also were applied to find the terms that were discriminative of a particular activity using the ChiSquaredAttributeEval mining algorithm in Weka.

Sentiment analysis and categorization of content. Recently, there has been more research in computational linguistics and machine learning about sentiment analysis, the automated detection of opinions or attitudes in text.[23] Sentiment analysis extracts subjective information about a topic or a document by applying computational analytic techniques. The authors relied on a sentiment analysis tool (twittersentiment.appspot.com/) that categorizes Tweets as positive or negative. Accuracy >80% has been reported for other corpora.[24] Tweets were qualitatively examined for aspects of context typically assessed in ecologic momentary assessment: time, purpose, environment, social context, and feeling.

Step 5. Interpretation

In the interpretation step, the decision is made to terminate the content-mining process or iterate, and domain experts play a critical role. When domain experts decide results are interpretable according to their study's scope and aims, the processes are terminated.[25] Conversely, when the results are not satisfactory, one must return to the previous iteration.[11] In the physical activity example, domain experts terminated the process because the study scope was to investigate the spring season. An alternative decision could have been to return to Step 2 in order to import Tweets in a different season (e.g., winter) to increase generalizability.

Results

Results are organized according to the five research questions: (1) What is the content of Tweets that mention specific physical activities? (2) Does Tweet content vary by specific physical activity? (3) Does Tweet content change over time? (4) What proportion of Tweets that mention specific physical activities express positive as compared to negative sentiments? and (5) How is context expressed in Tweets that mention specific physical activities?

Tweet Contents

The computed Tweet term-frequency dictionary contained 31,489 terms (Table 1). The most frequently occurring unigram, bigram, and trigram for each of the physical activity terms are described in Table 2. For example, "class," "aerobic class," and "water aerobics class" were the frequent unigram, bigram, and trigram in Tweets that mention aerobics. N-gram–based text computing produced the Tweet term-frequency dictionary containing 31,489 terms from the corpus of 174,394 Tweets. Table 1 shows the frequency of six sample terms.

Distinct Content Across Physical Activities

The term "good" appeared across all activities, whereas "obesity" occurred only for a few activities; "but should" (light workout); "mountain" (hiking); and "the basics" (bicycling) were the most distinct terms calculated by chi-square test. Interactive trends graphs (Appendix B, available online at www.ajpmonline.org) show the selected distinct terms that frequently appeared in Tweets that mention physical activity; those distinct terms include *student*, *women*, *unintentional physical activity*, *breast cancer prognosis*, *arthritis walk season*, *cell phone*, *improve symptoms gerd*, *everyone benefits*, and *CDC* over the course of 12 weeks.

Content Change Over Time

The snapshots of 3D motion charts (Figure 2) display 12-week trends of Tweets that mention bicycling or physical activity; on April 2010, "bike friendly" appeared as the most frequent distinct term on Tweets that mention bicycling. The interactive trends graph (Appendix B, available online at www.ajpmonline.org) shows how the contents of Tweets that mention physical activity changed over time. *Students* frequently appeared from Week 2 to Week 3 in March. *Normal bm (bowel movement)* and *women* occurred from late March to early April. *Unintentional physical activity* was shortly posted for a few days in early April. *Breast cancer prognosis* frequently appeared in mid-April for 10 days followed next most often by *cell phone*. In early May, *improve symptoms gerd* appeared frequently followed by *everyone benefits*. *CDC* was frequently discussed around the last week in May.

Table 2. Top three unigram, bigram, and trigram for each of the physical activity terms

Term	Unigram	Bigram	Trigram
aerobics	Class	aerobic class	water aerobics class
	Water	water aerobics	minutes doing aerobics
	Step	calories burned	my aerobics class
bicycling	Google	Google maps	bike-friendly cities
	Maps	bicycling magazine	Google maps adds
	Cities	bike friendly	50 bike friendly
chop wood	Fire	i chopped	I chopped wood
	Free	of wood	chopped some wood
	Today	wood for	chopped wood for
dancing	Stars	dancing with	dancing with the
	Video	dancing in	dancing in the
	love	dancing to	I'm dancing
do yard work	today	to do	some yard work
	time	going to	work to do
	good	have to	yard work today
hiking	day	go hiking	to go hiking
	today	I'm	a hiking accident
	trail	hiking with	British men die
jogging	morning	go jogging	to go jogging
	today	jogging with	jogging with a
	good	went jogging	went jogging with
just got back from gym	good	I feel	the gym now
	time	time to	to the gym
	shower	home from	the gym with
light workout	today	for a	light workout today
	good	I'm	do a light
	gym	workout today	light workout in
physical activity	health	activity plan	any physical activity
	weight	national physical	physical activity can
	exercise	weight loss	regular physical activity
play golf	Tiger	Tiger Woods	play gold with
	Woods	golf with	learn to play
	masters	will play	going to play
play basketball	hours	I played	I played basketball
	years	just played	played basketball for
	soccer	with my	played basketball in

(continued on next page)

Table 2. *(continued)*

Term	Unigram	Bigram	Trigram
ran miles	felt	felt good	mins and felt
	mins	5 miles	and felt good
	good	just ran	and felt great
swim	pool	swim in	for a swim
	lol	I am	can't swim
	today	swim with	how to swim
walk fast	lol	I am	making my way
	home	so fast	walking fast faces
	don't	to walk	downtown walking fast
weight lifting	muscle	build muscle	weight lifting for
	gym	weight loss	weight lifting gloves
	workout	to get	completed weight lifting
weight training	loss	weight loss	completed weight training
	completed	training for	weight training 400lbs
	fat	completed weight	weight training for

Positive Versus Negative Sentiments

For the physical activity data set during the period of July 21 to August 16, 2010, most Tweets reflected positive attitudes, with bicycling (77%) as the top-ranked category. Tweet categories that reflected more than 40% negative sentiments were as follows: walking fast, running, weight-lifting, physical activity, basketball, yard work, weight training, and jogging. Tweets mentioning hiking, golf, dancing, and swimming showed consistent positive sentiments, whereas others varied over time. Dancing-related Tweets were overwhelmingly positive (76%; n=84,217), and the associated trend graph shows that the relationship of positive and negative attitudes did not fluctuate (Appendix C, available online at www.ajpmonline.org).

Context of Various Physical Activities

Tweets that mention outdoor activities, such as hiking and yard work, contained information about physical contexts such as weather condition, trails, seasonal condition, and time of day (Table 3). In addition, Tweets that mention running provided detailed information about the activity (e.g., miles run, duration—hours/minutes/seconds, intensity—fast/slow). Emotional context (e.g., emotional obligation, feelings) was also prevalent in some Tweets. Tweets revealed social context (e.g., with my dad, friend, sister).

Discussion

The intent of the current paper was to introduce a simple and easy-to-apply method for mining Tweet contents, and to illustrate application of the content-mining pipeline to gain insights for health behavior research. All tools used for data collection and analysis are available publically and free of charge. Further, this study introduced a 3D motion chart as a visualization strategy of vast amount of mining results.[26] As opposed to 2D traditional visualization methods, a 3D motion chart was able to succinctly merge and present 12 weeks of Tweet topics within one chart.

Key challenges in the health behavior research field include the following: (1) assessment of complex health behaviors; (2) health-promoting behavior intervention design; and (3) motivation for health-promoting behaviors. The authors believe that the methods described may provide insights and address challenges for health behavior research. First, the study findings support the applicability of Twitter as an ecologic momentary assessment tool by demonstrating the relevance of naturally generated Tweet contents as a source of health behavior data. Such an approach may overcome the recognized limitations of self-reports of health behavior[27] and methods that include stimulation to generate responses.

For example, Tweets revealed situational momentary context prior to and right after physical activity. Analysis of frequently occurring terms provided situational context such as purpose (e.g., to build muscle); time (e.g., now, today); social context (e.g., gym with); environment situation (e.g., water, trail); feeling (e.g., felt great, love, hungry); and post-activity plans (sleeping, eating). Surprisingly, Tweets also captured fairly detailed measurement information such as the amount of calories burned ("157 calories burned") and distance covered ("ran 2.02 miles").

Second, the simple methods of topic detection may help health behavior researchers that have minimum programming skills utilize the detected topics for health behavior intervention design. In the physical activity content-mining case, distinct term lists indicate that some phrases appeared uniquely for a specific physical activity. Given that designing effective physical activity interventions is a considerable challenge,[28] frequently

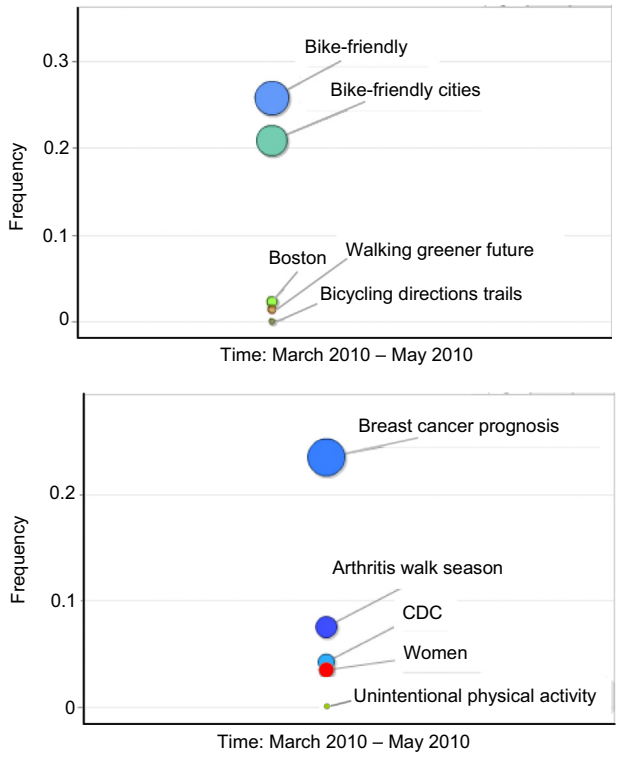

Figure 2. Weekly trends of Tweets that mention bicycling (top) and physical activity (bottom) with 3D motion chart visualization

occurring distinct terms may suggest intervention strategies. For example, *the basics*, *traffic*, and *safety* were frequently occurring distinct terms in bicycle-related Tweets. Researchers, government officials and health providers can harness such distinct terms for designing interventions that promote bicycling. In a similar way, terms related to time or distance measurement were distinctly common among Tweets that mention running, thus suggesting that these measures may be particularly important for interventions that incorporate running.

Finally, one of the biggest challenges in the health behavior research field is motivation for changing health behaviors. The authors observed that there is transformation of communication level reflected in the Tweet corpus. Although Tweet contents appeared at individual, community, and organization levels, the content of the three different levels of conversations is not independent. For example, when Google formally announced the new release of a bicycling map (i.e., organization level), Twitter users freely expressed their personal feelings with diary-like posts (individual level) and small group discussions (community level) occurred. In other words, the observations suggest that the Twitter medium was able to transform formal information into informal conversation, thus providing preliminary evidence that Twitter has the potential to support transformation of formal informational materials into conversations that may motivate behavior change.

Limitations

There are several limitations to the authors' application of content-mining methods to study the health behavior of physical activity that are applicable to other uses. The primary limitation of this study is its limited generalizability as a result of the characteristics of Twitter users (e.g., mainly young adults) and other factors (e.g., various languages having diverse linguistic structures, different Tweeting culture). The unit of analysis was the Tweet in the virtual Twitter community. However, each Tweet has a different probability of occurring in the data set, and its value is unknown. Further, some Tweets may be missed in a search because of the informal language used in Tweets. To minimize this issue, hash-tag lists (twitter.com/toptweets) were checked to avoid missing large volumes of Tweets referring to physical activity terms (informal terms or abbreviations) other than the search terms.

The findings are also limited because of the accuracy of the sentiment analysis tool used in this study. Although the accuracy of the tool is reported as being more than 80% according to its developers,[24] some experts in this field have expressed their concerns about the accuracy of sentiment analysis.[29] Further, Twitter users might have had a tendency to provide information that they believed

Table 3. Summary of Tweet content mentioning various physical activities

Criteria	Phrases in Tweets indicating context
Time	Now, just ran, just played, today, morning, time, will play, completed weight training, hours, years, free, went jogging, going to
Purpose	Calories burned, to build muscle, weight loss, to get, training for, for a swim, weight lifting for, played basket for, weight training for, have to, how to swim
Environment	Water, Google map, trail, bicycle-friendly cities, activity plan, pool, weight training 400lbs
Social	Gym with, play wit, swim with, with my, play in, class
Feeling	I felt great, felt good, love

to be consistent with social norms and expectations, and to over-report engagement in health-enhancing physical activity.[30–32] Last, it should be emphasized that the presented methods are simply based on frequency of the words to describe a phenomenon of interest. Other sophisticated natural-language tools and weighting techniques are necessary to provide a deeper understanding of the semantics of the Tweets and a representative sample.

Conclusion

This application of text-mining and sentiment analysis methods to analyze physical activity–related Tweets enhanced understanding of physical activity behaviors and their associated situational contexts. Such approaches offer an alternative to traditional self-reports and ecologic momentary assessments for capturing health-related behaviors.

The authors thank Drs. Mary W. Byrne, Elizabeth Cohn, PoYin Yen and Jacqueline Merrill for their contributions to the dissertation research on which this article is based.

The study was supported by grant T32NR007969 from the National Institute of Nursing Research. Article preparation was also supported by grant R01 HS019853 from the Agency for Healthcare Research and Quality.

No financial disclosures were reported by the authors of this paper.

References

1. Atienza AA, Patrick K. Mobile health: the killer app for cyberinfrastructure and consumer health. Am J Prev Med 2011;40(5S2):S151–S153.
2. Nathan KC, Amanda LG. Health behavior interventions in the age of Facebook. Am J Prev Med 2012;43(5):571–2.
3. Twitter Team. Twitter turns six. Twitter Blog 2012.
4. O'Connor B, Balasubramanyan R, Routledge B, Smith N. From Tweets to polls: linking text sentiment to public opinion time series. Proc Int AAAI (ICWSM 2010) 2010:122–9.
5. Eysenbach G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. Am J Prev Med 2011;40(5S2):S154–S158.
6. Paul MJ, Dredze M. You are what you Tweet: analyzing Twitter for public health. Proc ICWSM 2011.
7. Fisher J, Clayton M. Who gives a Tweet: assessing patients' interest in the use of social media for health care. Worldviews Evid Based Nurs 2012;9(2):100–8.
8. Militello LK, Kelly SA, Melnyk BM. Systematic review of text-messaging interventions to promote healthy behaviors in pediatric and adolescent populations: implications for clinical practice and research. Worldviews Evid Based Nurs 2012;9(2):66–77.
9. Schwartz JE, Stone AA. Strategies for analyzing ecological momentary assessment data. Health Psychol 1998;17(1):6–16.
10. DHHS. Office of disease prevention and health promotion. Physical activity guidelines advisory committee report. Washington DC: DHHS, 2008.
11. Liu B. Carey MJ, Ceri S, eds. Web data mining: exploring hyperlinks, contents and usage data. In carey MJ, Ceri S, eds. Berlin: Springer, 2006.
12. Dodds PS, Danforth CM. Measuring the happiness of large-scale written expression: songs, blogs, and presidents. J Happiness Stud 2010;11(4):441–56.
13. Kleinberg J. Bursty and hierarchical structure in streams. Data Min Knowl Disc 2003;7(4):373–97.
14. Grobelnik M, Mladenic D. Text-mining tutorial. In the Proceedings of Learning Methods for Text Understanding and Mining, Grenoble, France, 2004.
15. Salton G, Wong A, Yang CS. Vector-space model for automatic indexing. Commun ACM 1975;18(11):613–20.
16. Conway M, Doan S, Kawazoe A, Collier N. Classifying disease outbreak reports using N-grams and semantic features. Int J Med Inform 2009;78(12):e47–e58.
17. Wilbur WJ, Sirotkin K. The automatic identification of stop words. J Inform Sci 1992;18(1):45–55.
18. Salton G, Buckley C. Improving retrieval performance by relevance feedback. J Am Soc Information Sci 1990;41(4):288–97.
19. Jackson P, Moulinier I. Natural language processing for online applications: text retrieval, extraction and categorization. 2 rev. ed. Amsterdam: John Benjamins Pub Co, 2007.
20. Tomlinson S. Comparative evaluation of multilingual informtion access systems. Lecture Notes Comput Sci 2004;3237:169–82.
21. Porter M.F. Snowball: a language for stemming algorithms. 2001. snowball.tartarus.org/texts/introduction.html.
22. Tan P, Steinbach M, Kumar V. Introduction to data mining. Reading MA: Addison Wesley, 2006.
23. Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends Inform Retr 2008;2(1-2):1–135.
24. Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision, 2010. cs.wmich.edu/~tllake/fileshare/TwitterDistantSupervision09.pdf.
25. Gershon N, Eick SG. Information visualization applications in the real world. IEEE Comput Graph Appl 1997;17(4):66.
26. Mabry PL. Making sense of the data explosion: the promise of systems science Am J Prev Med 2011;40(5S2):S159–S161.
27. Oenema A, Brug J, Dijkstra A, de Weerdt I, de Vries H. Efficacy and use of an internet-delivered computer-tailored lifestyle intervention, targeting saturated fat intake, physical activity and smoking cessation: a randomized controlled trial. Ann Behav Med 2008;35(2):125–35.
28. Cavallo DN, Tate DF, Ries AV, Brown JD, DeVellis RF, Ammerman AS. A social media-based physical activity intervention: a randomized controlled trial original research article. Am J Prev Med 2012;43(5):527–32.
29. Pak A, Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining. Proc Int LREC'10, European Language Resources Association (ELRA). Valletta, Malta, 2010.
30. Adams SA, Matthews CE, Ebbeling CB, et al. The effect of social desirability and social approval on self-reports of physical activity. Am J Epidemiol 2005;161(4):389–98.
31. Klesges LM, Baranowski T, Beech B, et al. Social desirability bias in self-reported dietary, physical activity and weight concerns measures in 8- to 10-year-old African-American girls: results from the Girls Health Enrichment Multisite Studies (GEMS). Prev Med 2004;38(S):S78–S87.
32. Warnecke RB, Johnson TP, Chavez N, et al. Improving question wording in surveys of culturally diverse populations. Ann Epidemiol 1997;7(5):334–42.

Appendix

Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.amepre.2013.02.025.