A Tool for Monitoring and Analyzing HealthCare Tweets

Ahmed Ali, Walid Magdy, and Stephan Vogel
Qatar Computing Research Institute
Qatar Foundation
Doha, Qatar
{amali, wmagdy, svogel}@qf.org.qa

ABSTRACT

The amount of data exchanged over social media is witnessing a major growth in the last few years. Various studies in different domains investigated extracting useful information from this exchanged data. Less attention was directed toward studying healthcare in social media compared to other topics such as politics and marketing. In this paper, we present a platform for monitoring healthcare tweets on social media in different regions. The platform offers a solution to governments or healthcare providers to monitor public health and measure public satisfaction with healthcare services from what people post on Twitter. It helps in the early detection of disease outbreak and healthcare public view. Our platform uses an automatic classification method for detecting healthcare related tweets. It presents comprehensive reports that provide the most popular topics people share, sentiment analysis of tweets, multimedia content related to health. The platform is tested and demonstrated for three different locations: London, Boston, and Dublin. In addition, its effectiveness is tested for Arabic and English tweets with no specific location.

1. INTRODUCTION

Social media is currently playing a fundamental role in the life of many internet users. Public posts on a social website such as Twitter include personal status, opinion sharing, discussions, marketing, campaigning … etc. Among the material users share on Twitter are tweets related to health and healthcare. Some users share information and updates on their health or the health of loved ones. This is a common behavior by many people who seek support during difficult times. These social posts can be of tremendous value if detected and monitored by healthcare providers: they provide indicators about the general public health, they can help in early detecting of an epidemic, or can alert about ongoing concerns with healthcare [1-3]. In addition, people often share their experience with healthcare facilities like hospitals, clinics, and health centers. People can recommend specific doctors or clinics, or they can complain about particular aspects of health facility, such as queuing in emergency department, food service in hospital, competence and attitude of caregivers … etc. In general, healthcare related tweets can be utilized as an indicator about the quality of services provided by these health facilities, and can help to improve the provided services to patients. Unfortunately, studying healthcare material posted on social media has received less attention compared to other topics [2, 4].

In this paper, we present a tool that monitors and analyzes health related tweets and presents a comprehensive report for what the public shares about health in specific locations. The presented work in this paper addresses the following questions:

1. How to collect health related posts from social media, such as Twitter?
2. What analysis is required to extract indicative information from the collected tweets?
3. How extracted information could be presented?

Our proposed approach for collecting healthcare tweets uses a semi-supervised approach to identify relevant tweets based on a set of pre-defined keywords. Sentiment analysis is applied to obtain an indication of people's satisfaction and/or feelings. In addition, the collected tweets are parsed and analyzed to extract the most shared videos, images, and links in these tweets. Finally, a comprehensive report about the extracted information is generated and presented in a web interface to allow experts or care givers to monitor the public health in specific locations. This report includes: most circulated tweets, videos, images, and articles; tag-cloud of top used terms; and sentiment graph. We test this platform on three different locations: London, Boston and Dublin. In addition, it is tested on English and Arabic tweets coming from unspecified locations.

2. HEALTHCARE AND SOCIAL MEDIA

Social network started to have visible presence in health care in 2008. "*Hello-Health*" [5] is an initiative to think of the Electronic Health Record (EHR) as a social network, which acts as a concierge practice. The main shift is from large hospital networks to patient support groups and news media tools, such as weblogs, instant messaging, video chat and social networking. Patient could use the *Hello-Health* network to send a message to the physician describing the symptoms and asking for advice. A quick e-mail from the physician to follow up, and if needed a guaranteed hospital visit scheduled in 24 hours. Another example, *PatientsLikeMe* [6] is a platform to use online personal information for patients to share their experience using patient-reported outcomes, find other patients like them matched on demographic and clinical characteristics, and learn from the aggregated data reports of others to improve their outcomes. The main goal for *PatientsLikeMe* is to help patients answer the question: "Given my status, what is the best outcome I can hope to achieve, and how do I get there?".

Patients have moved beyond searching and asking towards sharing and interacting. The *PewInternet&American 2011* study showed that 23% of those with chronic health issues, such as cancer, diabetes, or heart disease, have gone online to look for patients with similar conditions; while only 15% of patients with no chronic conditions have sought peer-to-peer information [7].

The Centers for Disease Control (CDC) has been leading the efforts to use twitter as a channel to reach out to the public to deliver more information about infection, disease. During the H1N1 outbreak of 2009, the CDC decided to communicate with patients and caregivers across the US though Twitter. The CDC got more than 1.2 million followers for emergency information and 46,000 following the flu [8, 9].

HealthVault is a Microsoft ongoing initiative to integrate health records into a unified solution, enabling patients to store, and share health information online[10]. Google Health was a similar

HSD 2013, August 1, 2013, Dublin, Ireland.

platform launched in 2008 [11]. Both HealthVault and Google Health were developed as platforms to facilitate easy access to and management of the Electronic Public Health Record EPHR and sharing details amongst user specified networks.

In [12], it was shown that the shift of accessibility in healthcare is no longer just about getting an appointment with a physician or scheduling a treatment; it goes to community outreach, since social network sites can help hospitals to communicate with patients they serve. However social media in healthcare is still in its infancy. A study carried out in the US including 5,800 hospitals, showed that only 965 hospitals, which is less than 17%, are using social media to reach out to patients [13]

Several studies investigated the use of Twitter to analyze ongoing health related events. A study for detecting the influenza epidemic was carried out over 300 million tweets [14]. They used support vector machines (SVM) and showed the feasibility of their approach with a correlation of 0.89 to the gold standard from the Infection Disease Surveillance Center IDSC, thus outperforming the Google flu trend. Another recent study used Twitter to track the public concern in the US during the influenza A H1N1 [1]. It showed the capability of Twitter feeds not only to describe, but to track users' interest and concern about the development of the H1N1 epidemic and track disease activity. The study used a data set of 5 million tweets.

The aforementioned studies focused on sampling and analyzing healthcare related data on social network to have representative sampling with enough confidence in the given results. However, to the best of our knowledge, no attention was directed to providing a generic healthcare platform for detecting health related social posts, analyzing them, then presenting them in a comprehensive way for capturing trends in people's perception and reaction towards health related issues.

3. TWEETS MONITORING METHODOLOGY

In this section, the method underlying our study is presented. First we describe the data collection process, then the analysis criteria are introduced, and finally we describe how the healthcare tweets are modeled in both Arabic and English tweets.

3.1 Microblog Data Streams

Our platform is monitoring tweets coming from 5 streams; three streams coming from specific locations: London, Dublin, and Boston; and two language streams that are not necessary geo-tagged: English stream and Arabic stream. The purpose of the location-tweets streams is to monitor health related tweets in specific regions, while the purpose of language-specific streams is to monitor the trend in health related tweets globally over Arabic and English tweets in general.

Location-tweets are streamed based on geo-location of tweets by specifying longitude, latitude, and radius, which are specified to cover the targeted cities. Language-tweets are streamed by searching Twitter for "lang:en" and "lang:ar" for the English and Arabic tweets respectively. For each stream, health related terms were used to identify potentially health-related tweets. Table 1 shows the details of how each stream is collected and the average number of healthcare tweets identified per day. Table 2 presents an example of the terms used to identify the potentially health related tweets for English and Arabic.

Text of streamed tweets is pre-processed using state-of-the-art normalization techniques for microblogs in English [15] and Arabic [16] .

Table 1: Tweets Streams

Stream	English	Arabic	Boston	London	Dublin
Location	Any	Any	Boston	London	Dublin
Language	English	Arabic	English	English	English
Longitude	Null	Null	42.36,	51.51,	53.25,
Latitude			-71.05	-0.1241	-6.36
Radius			50 km	50 km	50 km
avg. # tweets collected/day	50,000	20,000	3000	7000	800

Table 2: Example of terms used to identify potential health-related tweets

English	*hospital clinic disease infection virus #health #healthcare #nhs #disease #emergency #medicine #ObamaCare*
Arabic	مستشفي عياد مريض مرض الم فيرس فيروس وجع تعبان مصاب

3.2 Modeling Healthcare in Twitter

Terms used in Table 2 lead to the collection of tweets that are potentially related to healthcare. However, we noticed that some of the collected tweets that contain any of these terms are irrelevant to healthcare because of the multiuse of these terms in different domains. For example, in English word 'virus' has a potential to exist in healthcare tweets, however it occurs more often in the context of computer virus. Similarly, in Arabic we found out that the word "**كورونا**" (corona) appears so often in the first 2 weeks of May 2013 and this is due to virus corona in Saudi Arabia, and such a word is unlikely to stay as a healthcare top frequent keyword after the virus outbreak. Therefore, a more reliable classification is required to distinguish between healthcare relevant and irrelevant tweets. For this purpose, we used more restricted queries to select the tweets among the potentially identified healthcare ones that are very likely to be relevant healthcare. Then we used the narrow identified tweets to train a support vector machine (SVM) with linear kernel to classify the remaining identified tweets as relevant or irrelevant to healthcare.

The SVM models are trained and updated automatically in an unsupervised manner. Figure 1 shows our classification algorithm. Using the same tweets stream we build three data sets, a positive and a negative set for the training the models, and the set of potentially relevant tweets, which would be classified by the SVM classifier. Our approach works as follows:

- Tweets Set$_P$ P (positive): the positive tweets are tweets related to health and healthcare and we select them by having a restricted precise Boolean regular expression. For example in English, we search for tweets with the most frequent healthcare hash tags such as #healthcare #ObamaCare #nhs and the tweets itself have at least one of the healthcare keywords such as: infection, virus, hospital, clinic ... etc. Out of theses tweets, we select the most retweeted 1000 tweets and use them as positive training examples.
- Tweets Set$_T$ (potentially relevant tweet): This set contains potentially relevant tweets. They are identified using the general health related terms shown in Table 2. Intuitively, this set is larger than the Set$_P$ since it includes any tweet, which matches one or several of the given keywords. However the precession is low, since this set contains irrelevant tweets such as the computer virus tweets mentioned earlier. A similar pattern happens in Arabic tweets, where religious tweets often have one or more health-related keywords. The Tweet set T suffers from low precision, and this database should be filtered to identify the truly healthcare related tweets.
- Tweet Set$_N$ (negative): This set is used as negative examples in the SVM training process. We select randomly 1000 tweets that

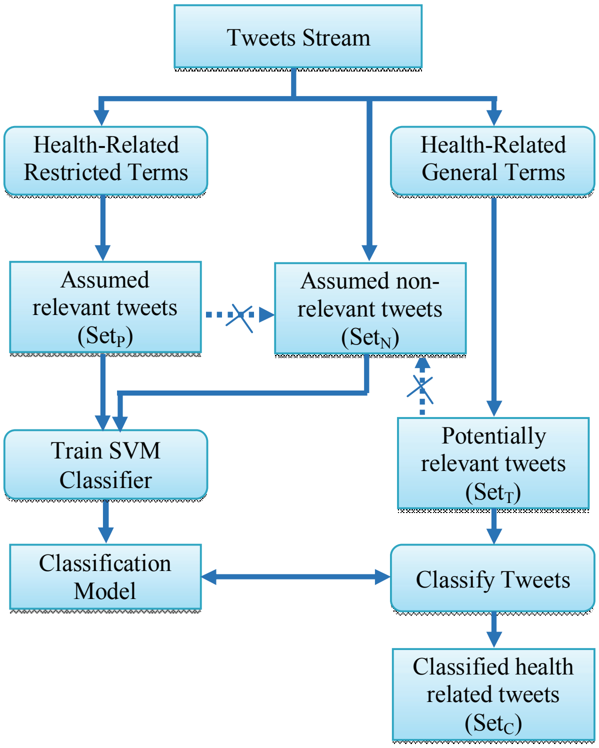
Tweets Stream
Health-Related Restricted Terms
Health-Related General Terms
Assumed relevant tweets (Set$_P$)
Assumed non-relevant tweets (Set$_N$)
Train SVM Classifier
Potentially relevant tweets (Set$_T$)
Classification Model
Classify Tweets
Classified health related tweets (Set$_C$)

Figure 1: Healthcare tweets expansion using SVM classifier

do not match any of the healthcare terms. Set$_P$ and Set$_N$ are used together for training the SVM model, which is then used to classify the tweets of Set$_T$ to either relevant or irrelevant tweets to healthcare.

The process of training the SVM model is applied every 30 minutes using the identified tweets from the past 48 hours. Classified tweets as relevant from Set$_T$ are added to Set$_P$ and a comprehensive report is generated from them and presented in a web interface to users.

3.3 Healthcare Tweets Identification Results

The healthcare tweets retrieval is done for each of the tweet feeds separately. Table 3 shows the results studied over 3 feeds; Arabic, London and Boston. For each one of the 3 feeds, 200 tweets randomly chosen over 2 days to measure the robustness of the selection algorithm, the evaluation has been done manually by labeling the tweet either relevant or irrelevant to healthcare.

As shown in Table 3, the first step is to start by Set$_P$ in the first row. It has high precision and low coverage, which is used as positive examples to train the SVM models. The second set (Set$_T$) is larger and potentially relevant as shown in the second row. This set is the one that matches the general health terms, and which is classified later by the SVM model. Finally, the classified set (Set$_C$) in the third row is used for monitoring and analysis. The results show the robustness the classification approach, where large number of healthcare tweets is identified with high precision.

Table 3: Accuracy for healthcare related tweets

Tweets Set	Arabic Stream		Boston Stream		London Stream	
	#tweets	Prec	#tweets	Prec.	#tweets	Prec.
Set$_P$	3000	0.90	400	0.95	550	0.97
Set$_T$	20,000	0.65	3000	0.71	7000	0.75
Set$_C$	12,000	0.82	1900	0.89	4500	0.87

4. HEALTH-RELATED TWEETS ANALYSIS AND VISUALIZATION

Extensive analysis is applied to the identified health-related tweets from each stream. A web interface has been developed to select any of the data streams and display the results of the analysis. The presentation of content is inspired from previous work in [17, 18], which presents a comprehensive report about relevant tweets to a given topic or entity. However, we use additional visualization tools that relate more to the analysis we use for our task. The analysis is applied to the identified tweets in the last 48 hours and the presented reports are updated every 30 minutes.

The analysis applied and presented consists of:

- **Extracting Most Popular Tweet:**

Identified health-related tweets of each stream are grouped by aggregating similar tweets into the same group. For a fast and robust matching between tweets, we keep only the main content text of tweets by removing all hashtags, name mentions, URLs, punctuations, symbols, emoticons, and retweet symbols. Tweets that match exactly in their main content are grouped together. Groups are presented in ranked (descending) order by their size with the most common tweet form as the representative of the group along with the number of tweets in the group.

- **Extracting Top Circulated Videos, Images, and Links:**

Since URLs in tweets are typically shortened and URLs may have multiple shortened forms, we expand all URLS to find the original URLs. We used URLs pointing to YouTube videos to obtain a ranked list of the most popular videos and embed them in the report separately. Also links pointing to Twitter images are presented for the most circulated images. Other URLs are extracted and their titles are presented with links to the pages. Links are ranked by the number of appearances in tweets.

- **Sentiment Analysis:**

Applying sentiment analysis on the healthcare related tweets can be used as an indicator about public satisfaction or feelings towards health. We use the SentiStrength [19] tools to estimate the strength of the positive and the negative sentiments in each tweet using the scale from -5 (extremely negative) to +5 (extremely positive). Tweets are then sorted according to the sentiment score to identify the most negative and most positive tweets. Also, we show the sentiment score on the top frequent tweets, which indicates the public view especially with the most popular (re-tweed) posts.

- **Tag Cloud:**

The top frequent words, terms, and hash tags (excluding the stop words) are used to draw a tag cloud, where font size used for the different terms indicates their frequencies. The tag-clouds helps to summarize the most popular terms in the tweets, which in turn indicate the most popular topics people are interested in.

5. SAMPLES OF THE EXTRACTED INFORMATION FROM TWEETS

In this section we present examples of the information that was extracted from the identified healthcare tweets to demonstrate the benefit of our system.

The top frequent tweets show the public awareness and concern with healthcare services. For example the third line in the most popular tweets is very important to professionals in healthcare sectors, government, and probably insurance companies to be aware with the concern of such a big difference in price for the

Table 4: Examples of the extracted top tweets, articles, videos, and images related to healthcare from different streams

Most Popular Tweet	Stream
nurses fighting to save nhs for patients, 150,000-strong london march largely ignored by uk media	London
it is so painful. 135,000 health professionals applied for 1,760 jobs in alicante some months ago	Dublin
how can a procedure cost $297,000 at one hospital and $84,000 at another? disparities in #healthcare	Boston

Most Circulated Article title	Stream
Cap on number of GP visits being considered by Tories	London
IMF to have power over Irish healthcare spending	Dublin
5 Jedi Mind Tricks to Help Yourself Get Healthy	Boston
فايروس كورونا الجديد وخطر واعراضه (lecture on virus corona)	Arabic

Most Circulated Videos (showing titles only)	Stream
My Spinal Cord Injury Story in pictures - Motivation -Bradley Hill Fitness	London
The Epidemiology of In-Flight Medical Emergencies	Dublin
Corona virus and the risk of iron and its symptoms	Arabic

Most Circulated Images (showing caption only)	Stream
Think again about going barefoot, a viewer got this nasty virus http://t.co/9rreT7Deys	English
كيف الإصابة بـ #فيروس_كورونا ؟وماهي أعراضه ؟ وكيف تقي نفسك؟وماهو العلاج ؟ http://t.co/3JBiLkgqhj	Arabic

same procedure in two different hospitals. The first post is important for the hiring in healthcare sector.

The example of most circulated articles show the public concern about both healthcare (first two examples) and public health (in the last 2 examples). Interesting findings result from comparing the different data streams: while common health related topic like healthy life style and cuts for the healthcare are discussed in Dublin, London, and Boston,, we find that in the Arabic tweets, mainly in Saudi Arabia, the virus corona was the hot topic at the same time, due to the outbreak in. So, it is really useful to get the proposed automatic update about the health public concern.

The examples of videos are mainly focusing on public health issues, such as diseases outbreaks, healthy life, and lecture on how to deal with a certain virus (corona) … etc.

The section displaying the most circulated images is a very dynamic part in the generated report, as people are sharing images very frequently. For example the tweet about a woman contracting a virus after walking barefoot has been posted more than 1,100 times, but disappeared again within 2 days. Being able to address these findings will be quite useful for caregivers.

6. CONCLUSION AND FUTURE WORK

This paper presented a platform for monitoring and analyzing public tweets that relate to healthcare. We proposed an automatic approach for collecting relevant tweets to healthcare. We showed that our approach leads to the retrieval of reasonable number of healthcare tweets with high precision. The platform analyzes these collected tweets and extracts some useful information to be presented in a web interface for healthcare experts. Our platform was tested on 5 streams of tweets from specific and general location in two different languages. We believe that this platform is considered a good start for similar tools that helps in monitoring flowing information about health in social media in an attempt to utilize this information for improving the health services and the detection of public health threats.

For future work, we plan to apply different user studies on our platform to get feedback about its practicality for healthcare institutes. Furthermore, automatic alerting systems could be developed within the system to trigger alerts without the need for checking the web interface. A simple alert can be the number of identified health related tweets in a given window of time. Also, it can be a more advanced alert based on analyzing the appearance of new terms that indicate a healthcare disasters or threats.

7. REFERENCES

1. Signorini, A., A. Segre, and P. Polgreen, *The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S during the Influenza A H1N1 Panademic.* PLoS ONE, 2011. **6**(5): p. 1-10.
2. Prier, K.W., et al., *Identifying Health-Related Topic on Twitter An Exploration of Tobacco-Related Tweets as a Test Topic*, in *Springer-Verlag Berlin Heidelberg*2011. p. 18-25.
3. Paul, M.J. and M. Dredze. *You Are what you Tweet: Analyzing Twitter for Public Health.* in *The Fifth International AAAI conference on Weblogs and Social Media*. 2011.
4. Adams, S.A., Blog-based applications and health information: Two case studies that illustrate important questions for Consumer Health Informatics (CHI) research. International Journal of Medical Informatics, 2010. 79: p. e89-e96.
5. Hawn, C., Take Two Aspirin And Tweet Me In the Morning: How Twitter, Facebook, And Other Social Media Are Reshaping Helath Care. Health Affairs, 2009. **28**: p. 361-368.
6. Wicks, P., et al., *Sharing Health Data for Better Outcomes on PatientsLikeMe.* Journal of Medical Internet Research, 2010.
7. Fox Susannah, *The social life of Health Information, 2011*, in *Pew Internet & American Life Project*2011: http://pewinternet.org/Reports/2011/Social-Life-of-Health-Info.aspx. p. 1-45.
8. Aikin, A. *Case Study: Social Networking at the CDC*. 2010.
9. Eytan, T., et al., *Social Media and the Health System.* The Permanente Journal, 2011. **15**: p. 71-74.
10. Mircrosoft, *Healthvault: Connected Continuous Care How technology will transform chronic disease management*, 2011.
11. Aramaki, E., S. Maskawa, and M. Morita. *Twitter catches the flu: detecting influenza epidemics using Twitter.* in *EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2011.
12. Williams, J., *A New Road Map for Healthcare Business Success.* Healthcare Financial Management, 2011: p. 63-69.
13. Bennett, E., *A New Home for the Hospital Social Network List.* 2012.
14. ARAMAKI, E., S. MASKAWA, and M. MORITA, *Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter*, in *EMNLP*2011. p. 1568–1576
15. Han, B. and T. Baldwin, *Lexical Normalisation of Short Text Messages: Makn Sens a #twitter*, in *ACL-HLT*2011. p. 368–378.
16. Darwish, K., W. Magdy, and A. Mourad, *Language Processing for Arabic Microblog Retrieval*, in *CIKM*2012.
17. Bennett, E., *Hospital Social Network List.* 2011.
18. Magdy, W., A. Ali, and K. Darwish. *A summarization tool for time-sensitive social media*. in *CIKM*. 2012.
19. Yerva, S.R., et al., *TweetSpector: Entity-based retrieval of Tweets*, in *SIGIR* 2012.
20. Thelwall, M., et al., *Sentiment strength detection in short informal text.* Journal of the American Society for Information Science and Technology, 2010. **61**(12): p. 2544–2558.