Paper 3426-2015

Application of Text Mining on Tweets to Analyze Information about Type-2 Diabetes

Shubhi Choudhary, Vijay Singh, Goutam Chakraborty, Oklahoma State University, OK, US

ABSTRACT

Twitter is a powerful form of social media for sharing information about various issues and can be used to raise awareness and collect pointers about associated risk factors and preventive measures. Type-2 diabetes is a national problem in the US. We analyzed twitter feeds about Type-2 diabetes in order to suggest a possible use of social media with respect to a study of any ailment. To accomplish this task, 3.488 tweets were collected using Twitter API v1.1 in a customized Python script. Tweets, follower counts, and user information were extracted via these scripts. The tweets were segregated into different groups on the basis of their annotations related to risk factors, complications, preventions and precautions, and so on. We then used SAS® Text Miner to analyze the data. We found that 70% of the tweets stated the status quo based on marketing and awareness campaigns. The remaining 30% of tweets contained various key terms and labels associated with type-2 diabetes. It was interesting to find that influential users tweeted more about precautionary measures whereas non-influential users offered suggestions about treatments as well as preventions and precautions.

INTRODUCTION

Popular online social media sites are excellent platforms that for providing healthcare education and support to ever growing internet users. Fox & Fallows (2003) in their research found that third most important use of Internet technology was searching out for health related information. They estimated that more than two-third of Internet users in US seeks online help to obtain health information. There is a plethora of information available about health information in this digitized world but there needs to be a medium where people could share or express any sort of information that could help others. Social media provides such type of a platform and can be used as an application that allows creation and exchange of user generated content as defined by Kaplan & Haenlein (2010). Diabetes is a very common and costly chronic illness affecting people of all ages and has been found out to be the seventh leading cause of death in the US as pointed out in National Diabetes Statistics Report (2014). In our study, we try showcase the ways in which a popular microblogging social networking website 'Twitter' can be used to provide useful information about Diabetes.

The emergence of Web 2.0 in 2004 facilitated improved communication and collaboration between people via social networking. The terms such as eHealth, Health 2.0 or Medicine 2.0 were coined which facilitated social networking, participation and openness in groups as described by Eysenbach (2008). Bandura (1986) in his research explains the social cognitive theory as changes in one's lifestyle and health related behaviors by observing or following others. Abundant information is available on the web about healthcare in various forms and via various sites and mediums. Online communities such as chat rooms, bulletin boards, news groups and web forums have long served as mediums to share health issues and gain information. Social networking sites such as Twitter and Facebook differ from these traditional online communities in a way that it is built based on existing social ties and interpersonal relationship and tends to me more intense by bringing together people who share common interests (Rau, Gao & Ding (2008)). As a result of this, healthcare providers are turning to Social networking sites to circulate and provide health information to general public (Holt 2011). Hwang, Ottenbacher & Green (2010) in their recent study have demonstrated that social networking has been effective in improving users' access to health care information. Considering the growing number of Twitter users which has more than doubled in the past few years, we believe that Twitter's effective utilization will enable users to share health information and thus perhaps take preventive and precautionary measures.

Social media has become one of the most common mode of not just for communication but also to help people deal with various issues they are facing in their day to day lives. Diabetes also known as hypoglycemia is a medical condition in which the body is not able to absorb excess glucose present in the blood stream. The food we eat is converted into glucose which runs in our blood stream and is the only source of energy for the cells. The extra glucose is converted into glycogen and stored in the body. Type 2 diabetes is one of the most common conditions in diabetes where the pancreas cannot secrete enough insulin to support the absorption of blood sugar. Various procedures are present to cope diabetes; the most common ones being insulin shots and dialysis. Internet is a vast repository for a variety of information about the cure to diabetes however people tend to go for personalized advices and therefore turn to social media to guide them about it.

DATA

The primary source of the data used in this study are the twitter feeds that the users post. We collected the data using the twitter streaming API version 1.1. We wrote a phython script to extract the twitter feeds by using Tweepy which is a python library used to access the twitter API. We found tweets related to all types diabetes e.g., pre-diabetes, type-1 diabetes and type-2 diabetes. Since the study concentrates only on type-2 diabetes, we restricted the keywords which directly relates to type-2 diabetes. The list of keywords that we used are 'Hyperglycemia+diabetes', 'Lifestyle+diabetes', 'Glycemic control', 'T2D', 'Type 2 diabetes' 'Genetic+diabetes'. These keywords were suggested by domain experts in the field of medicine. The reason we used both the scientific and general names is because we were interested in collecting tweets from professionals in the field of medicine as well as common users. Since diabetes is not an outbreak or an epidemic, the number of tweets does not vary much throughout the year. Therefore, we collected the data over a period of time from August 21st - August 25th, September 2nd - September 6th and October 9th – October 14th in the year 2014. The number of tweets we were able to collect over the above mentioned time span were 4,374. However, during the exploratory analysis, we found duplicate tweets which occurred because a user might have copied the tweet rather than retweeting. Thus, we kept only one instance of a tweet and removed the duplicates. We also filtered the tweets on the basis of language and used only those tweets which were in English. The above filtering process led us to remove 886 tweets and so we were left out with 3,488 unique tweets for analyses.

Other than the tweets, we extracted tweet id which is a unique identifier of a tweet along with the user name which is the twitter handle of a user. We collected the number of followers a user had and a short introduction provided by the user. The variable named language is a variable denoting the language in which the user tweeted. We also collected information about the location, name and number of friends of the user. The data dictionary is provided in the table below.

Variable	Data type	Description
Tweet ID	Number	Every tweet has a unique identifier.
User-name	Character	The display name of the user who tweeted/retweeted.
Tweet	Character	Tweet
Description	Character	A short introduction provided by the user
Follower Count	Number	The number of followers for the user
Language	Character	Language of the tweet

Table 1: Data dictionary

ANALYSES & RESULTS

We performed text mining on the tweets to understand the actual content. We followed a standard text mining approach (as discussed in Chakraborty, Pagolu and Garla, 2014) by parsing which helped us to identify unique terms in the tweets. We used text miner's default setting in doing so and analyzed the terms in the term by document matrix. The term by document matrix helps us to understand the terms in tweets along with the frequency of its occurrence in each document or a tweet. Using interactive filter property of text miner, we filtered out terms which were not relevant to our study and could have possibly added noise in the analyses. We found six interesting clusters using text cluster node that dealt about dietary habits, statistics that twitter users' post, suggestions from insurance providers, current research and study about diabetes from various universities and organizations as well as various health awareness programs initiated by non-profit organizations. The typical text mining process we followed and the clusters obtained are provided below.

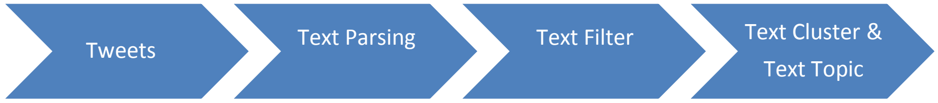


Figure 1. Text mining process

ID	Descriptive terms	Percent	Number of tweets
1	advice diabetes diabeteshelp helpful info la life living	(27)	(245)
2	'healthy food guide' +food +sweet arden diabetes endocrinology gt guide healthy http://t.co/hglgfpmzy9	(21)	(191)
3	+symptoms +work thirst control cuts happy http://t.co/p4mzvnyesr http://t.co/zedl8gbami	(19)	(172)
4	'heart disease' +blood +high +know +sugar amp heart insulin levels sept	(16)	(145)
5	+diet +fact +help +type announces doc health high-risk ids model	(10)	(91)
6	'reverse diabetes today' +body +hospital +story diabetesuk estimated government latest million population	(7)	(64)

Table 2: Text clusters of selected terms from the tweets

Definition of clusters:

- Cluster 1 contains 'advices/tips' given about controlling diabetes
- Cluster 2 describes various diet plans prescribed for diabetic and pre-diabetic partients
- Cluster 3 describes the 'symptoms' of diabetes- for example being constantly thirsty is a symptom of type-2 diabetes
- Cluster 4 talks about 'blood sugar' and 'insulin'
- Cluster 5 describes various 'diabetes facts' tweeted
- Cluster 6 talks about 'reversing diabetes'

During the text mining process we observed various tweets which contained information related to awareness campaigns and marketing. So, we extracted keywords from tweets specific to type-2 diabetes dealing with causes, precautions, treatments, etc. We found that 26% of the tweets (901 tweets) consisted of these keywords whereas the rest of the 74% (2,587) tweets consisted of awareness campaigns and marketing as shown in the figure below.

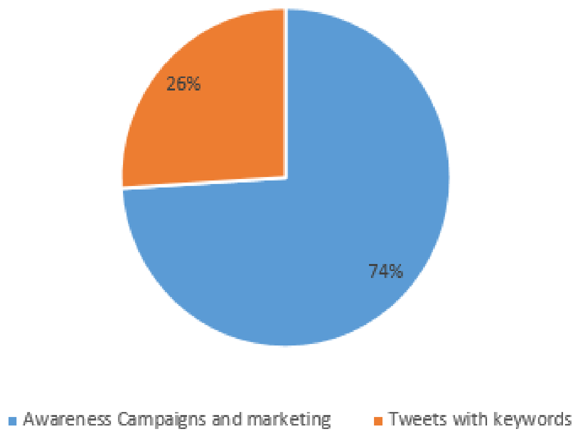


Figure 2. Distribution of tweets with keywords specific to type-2 diabetes and tweets about awareness campaigns and marketing related to diabetes.

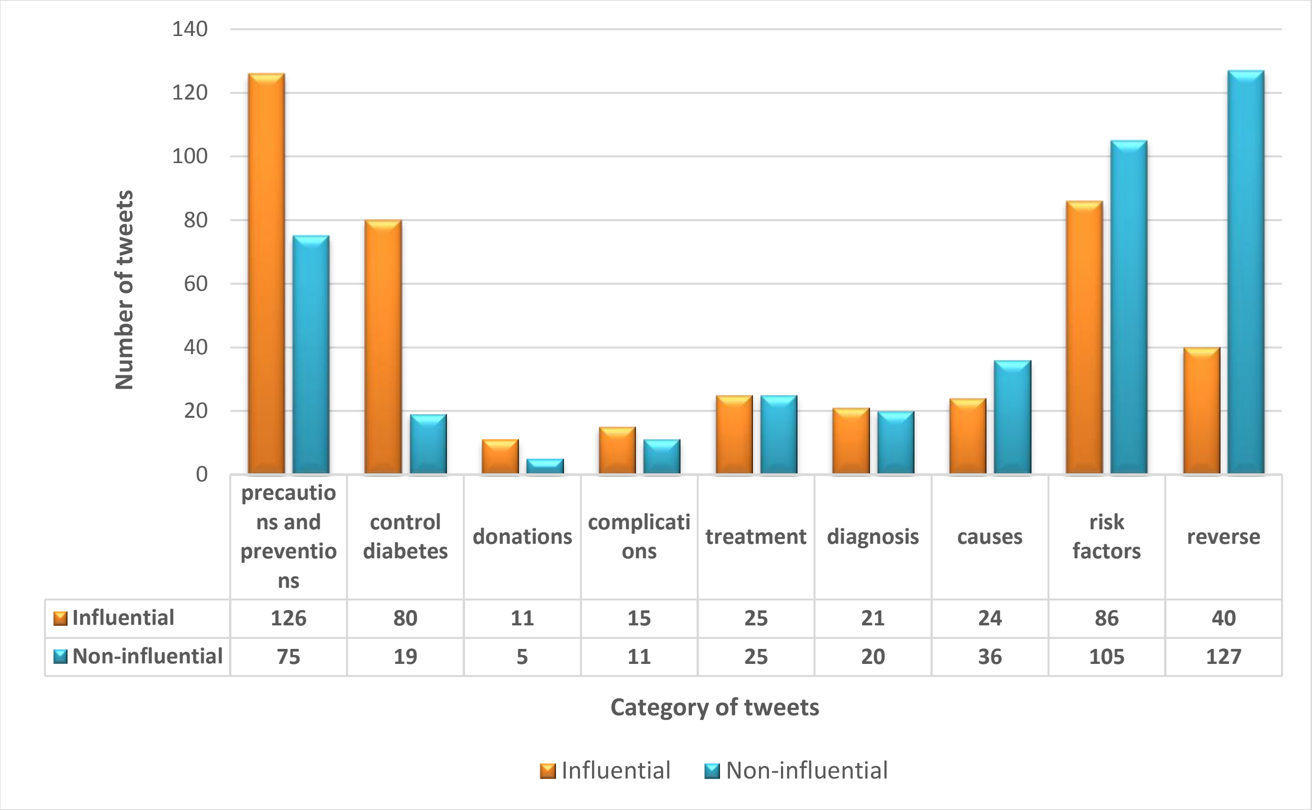
Focusing on with these 26% tweets, we tried to evaluate the category each tweet belongs to. Working with domain experts, we categorized them as risk factors, complications, diagnostics, precautions and preventions, treatment, causes, reverse and control of diabetes. Tweet segmentation provided us with interesting patterns in the nature of tweets. The trend in tweets that we observed were either suggestions to avoid diabetes or advices to those suffering from diabetes. Majority of the tweets expressed precautionary and preventive measures and the risk factors associated with diabetes. Other tweets dealt with ways to treat diabetes, the complications that could occur and the way diabetes could be diagnosed. The most interesting patterns of tweet that we observed are the ones where users tweeted about the ways by which diabetes could be reversed which according to domain experts is medically impossible. The distribution of all the categories of tweets with their frequency, percentage and description is provided below.

Category	Frequency	Percent	Description
Precaution & Prevention	**201**	**22**	This category of tweets describe all the safety measures to be taken by any person who is at a risk of developing type-2 diabetes.
Risk Factors	**191**	**21**	This category of tweets describe all the factors that may jeopardize a person into developing diabetes.
Reverse	**167**	**19**	This category of tweets describes- completely reversing or curing type-2 diabetes.
Control Diabetes	**149**	**17**	This category of tweets describe all the practices that can help maintaining blood sugar and hence control diabetes.
Causes	**60**	**7**	This category of tweets describe all the factors that trigger type-2 diabetes.
Treatment	**50**	**6**	This category of tweets describe different treatments available for diabetes.
Diagnosis	**41**	**5**	This category of tweets describe all the tests that help in the diagnosis of type-2 diabetes.
Complications	**26**	**3**	This category of tweets describes various consequences of diabetes.
Donations	**16**	**2**	This category of tweets relates to all the awareness and research campaigns that call for donations on twitter.

Table 2: Distribution of category of tweets

Extending our study, we examined the type of users who tweeted. During the exploratory analysis, we found tweets from health organizations, forums as well as from individuals with high follower counts. Later, we tried to divide them into influential users and non-Influential ones based on the follower counts of that user.

The users with follower counts less than 180 were classified as non-influential users while the ones with a follower count of 180 or more were treated as influential users. We observed that the influential users talked more about precautions and prevention of diabetes and ways to control diabetes. On the other hand, non-influential users tweeted about medically impossible phenomena of reversing diabetes. The distribution of various categories of tweets by influential and non-influential users is given in the figure below.



	precautions and prevention	control diabetes	donations	complications	treatment	diagnosis	causes	risk factors	reverse
Influential	126	80	11	15	25	21	24	86	40
Non-influential	75	19	5	11	25	20	36	105	127

Figure 3. Distribution of categories of tweets across influential and non-influential users.

CONCLUSION

WHEN THE DAMAGE IS DONE

In this case the target audience are the ones already suffering from diabetes. Twitter users provided range of suggestions such as consumption of bitter-gourd juice as well as taking insulin shots. Triple cure was the most widely discussed topic for the cure of diabetes among the twitter users. We also observed users posting links about dietary plans as prescribed by different dieticians and are customized according to different sugar levels and stages of diabetes. Various advertisements of sugar free products and artificial sweeteners are marketed all around the Internet. Tweets about the amputations as a treatment option was also observed in this segment. We also observed some false claims. These claims were hyperlinks to videos pretending to be suggestions about complete treatment or reversal of type 2 diabetes. However, in reality it is not possible to completely cure diabetes.

PREVENTION IS BETTER THAN CURE

Since Type 2 Diabetes is an acquired medical condition caused because of lifestyle changes, genetics and medical condition, the number of awareness campaigns are soaring to educate people about the onset of diabetes, symptoms, risk factors and complications associated with it. For example- the cycle marathon in various parts of USA which was an awareness campaign to promote healthy eating habits. Different statistics about the population suffering from diabetes and pre diabetes were tweeted which underscores the need to take precaution before falling into the trap of this ever growing disease. Interestingly, simple tips to have a check on diabetes for example how a can of soda a day can increase the risk of diabetes and on the contrary how an alcoholic drink a day can reduce the chances of diabetes were also observed. About 40 percent of the posts contained workout motivation- getting back to shape, eliminating obesity and thus reducing the chances of diabetes. Many individuals share their workout videos and their stories of how they fought diabetes by losing statistically disproportionate weights which is definitely inspiring for any new diabetic patient.

REFERENCES

Fox, S., & Fallows, D. (2003). *Internet health resources: health searches and email have become commonplace, but there is room for improvement in searches and overall Internet access*. Pew Internet & American Life Project.

Shaw, R. J., & Johnson, C. M. (2011). Health information seeking and social media use on the Internet among people with diabetes. *Online journal of public health informatics*, *3*(1).

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, *53*(1), 59-68.

Eysenbach, G. (2008). Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. *Journal of medical Internet research*, *10*(3).

Bandura, A. (1986). Social foundations of thought and action (pp. 5-107).

Hwang, K. O., Ottenbacher, A. J., Green, A. P., Cannon-Diehl, M. R., Richardson, O., Bernstam, E. V., & Thomas, E. J. (2010). Social support in an Internet weight loss community. *International journal of medical informatics*, *79*(1), 5-13.

Rau, P. L. P., Gao, Q., & Ding, Y. (2008). Relationship between the level of intimacy and lurking in online social network services. *Computers in Human Behavior*, *24*(6), 2757-2770.

Holt, C. (2011). Emerging technologies: web 2.0. *Health Information Management Journal*, *40*(1), 33.

Centers for Disease Control and Prevention. (2014). National diabetes statistics report: estimates of diabetes and its burden in the United States, 2014. *Atlanta, GA: US Department of Health and Human Services*.

Chakraborty, G., Pagolu, M., & Garla, S. (2014). *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*. SAS Institute.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Shubhi Choudhary, Oklahoma State University, Stillwater, OK, Email: shubhi.choudhary@okstate.edu
Shubhi Choudhary is a second year graduate student pursuing Maters in Management Information Systems at Oklahoma State University. She has two years of experience of using SAS® tools for Data Mining, Text Mining, and Sentiment Analysis projects. She is a SAS® certified Base Programmer and Business Analyst. In May 2014, she received her SAS® and OSU Data Mining Certificate.

Vijay Singh, Oklahoma State University, Stillwater, OK, Email: vijayms@okstate.edu
Vijay Singh is a second year graduate student majoring in Management Information Systems at Oklahoma State University. He has two years of experience of using SAS® tools for Data Mining, Text Mining, and Sentiment Analysis projects. He is a SAS® certified Base Programmer and Business Analyst. In May 2014, he received his SAS® and OSU Data Mining Certificate and he has been awarded SAS Global Forum 2015 student scholarship award.

Goutam Chakraborty, Oklahoma State University, Stillwater OK, Email:goutam.chakraborty@okstate.edu
Dr. Goutam Chakraborty is Ralph A. and Peggy A. Brenneman professor of marketing and founder of SAS® and OSU data mining certificate and SAS® and OSU marketing analytics certificate at Oklahoma State

University. He has published in many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has chaired the national conference for direct marketing educators for 2004 and 2005 and co-chaired M2007 data mining conference. He has over 25 years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.