3-22-2019

Exploring Diabetes and Users' Lifestyle Choices in Digital Spaces to Improve Health Outcomes

George Shaw Jr.
UNC Charlotte, gshaw11@uncc.edu

Tarun Sharma
UNC Charlotte, tsharma4@uncc.edu

Sthanu Ramakrishnan
UNC Charlotte, sramak11@uncc.edu

EXPLORING DIABETES AND USERS' LIFESTYLE CHOICES IN TWITTER TO IMPROVE HEALTH OUTCOMES

George Shaw, Jr.
University of North Carolina at Charlotte
gshaw11@uncc.edu

Tarun Sharma
University of North Carolina at Charlotte
tsharma4@uncc.edu

Sthanu Ramakrishnan
University of North Carolina at Charlotte
sramak11@uncc.edu

ABSTRACT

The information derived from social media analytic studies provides valuable sources of information for healthcare stakeholders. However, there is still a lack of research with using social media to identify the lifestyle choices of those dealing with diabetes in order to better understand and design impactful health interventions. This exploratory study aims to demonstrate how social media can be leveraged as a data source to help us understand the lifestyle choices of those dealing with diabetes. Using two text mining approaches - sentiment analysis and unsupervised topic modeling – food and physiology were topics expressed in both sentiments. Overall, lifestyle related topics accounted for nearly 25% of the topics identified in the corpus of data. There is a pressing need for incorporating predictive modelling approaches to this study in order to quantify our findings and how this knowledge can improve health outcomes from a population perspective.

Keywords

Natural language processing, topic modeling, diabetes, text-mining, diabetes

INTRODUCTION

The Centers for Disease Control and Prevention (CDC) defines diabetes as a chronic disease in which the body does not properly process food for use as energy; diabetes affected 9.4% of the US population and was the seventh leading cause of death in the United States as of 2015 (CDC, 2018). With increased inflation rates and people opting for unhealthy food, the risk of diabetes is at a constant high. Diabetes affecting the vast number of the US population also induces a major healthcare cost. The total estimated cost of diagnosed diabetes - as of 2017 - was $327 billion (American Diabetes Association, 2018). This represents 9.3% of the total healthcare spending from 2017. Improperly managed diabetes can lead to the onset of other chronic conditions like heart disease, stroke, hypertension and skin complications that have a combined estimated cost of $283 billion (CDC, 2018). This makes the research into diabetes crucial from a public health and healthcare system viewpoint.

With the ever-increasing use of social media, the amount of data continues to grow exponentially. Researchers have made use of this unstructured data to analyze the impact of diabetes on people's lives. Researchers have studied the impact of social media on diabetes using various techniques like exploratory analysis (Liu, Mei, Hanauer, Zheng and Lee, 2016; Sinnenberg et al., 2018), retrospective analysis (Tufts, Polsky, Volpp, Groeneveld, Ungar, Merchant and Pelullo 2018), sentiment analysis (Gabarron, Dorronzoro, Rivera-Romero and Wynn, 2018), and machine learning approaches to make predictions on prediabetic patients (Xu, Litchman, Gee, Whatcott, Chacon, Holmes and Srinivasan, 2018). Facebook and Twitter are amongst the major social media platforms used.

Information derived from social media data analytics' studies proves to be extremely valuable. Harris, Mueller, Snider, and Haire-Joshu (2013) examined the use of Twitter to disseminate diabetes information by local health departments. This would lead to a major reduction in the economic impact of diabetes by saving healthcare cost spent on educating and informing individuals. However, there is still a lack of research in identifying the lifestyle patterns of those dealing with diabetes in order to better understand and improve interventions before an extremity like death occurs due to diabetes. This preliminary study aims to demonstrate how social media can be leveraged as a data source to help us understand the lifestyle choices of those dealing with diabetes, which can help with an intervention design aimed at improving diabetic patients' outcomes and additional chronic conditions experienced by these patients.

BACKGROUND

Probabilistic models are important when wrangling a large corpus of data. This approach is appropriate when there is no preconceived understanding of the information that may exist within that data corpora. In probabilistic modeling – also known as generative probabilistic modeling - data is treated as derived from a generative process that also includes hidden variables (Blei, 2012). Topic modeling is a subdomain of the larger probabilistic modeling field. Topic models are algorithms used to discover central themes from an unstructured collection of documents (Sivic, Russell, Zisserman, Freeman and Efros, 2008). Topic models can be applied to multiple types of data, this includes genetic, medical images, and social media data sources (Blei, 2012).

Within the context of social media, Twitter is a common social media platform that has been used by researchers to conduct computational content analysis research. For communicable related diseases, Twitter has been used to collect people's sentiments regarding the H1N1 outbreak (Chew and Eysenbach, 2010). From a non-communicable disease perspective, multiple chronic health issues have been detected based on users' Tweets (Yin, Fabbri, Rosenbloom and Malin, 2015). There are several probabilistic topic models that exist, but the well-known and studied Latent Dirichlet Allocation (LDA) topic model approach was used in this study. LDA has proven to be a successful model for disclosing hidden topics in a Twitter data corpus (Guo, Vargo, Pan, Ding and Ishwar, P, 2016; Webb, Karami and Kitzie, 2018).

Food behavior and lack of physical activity are common lifestyle behaviors discussed in Twitter that are attributed to diabetes and childhood obesity (Harris et al., 2014). Improper dietary habits that include food with added caloric sweeteners and a decline in physical activity are linked to diabetes (Popkin, 2015). Twitter allows individuals representing the diabetes community to express their sentiments but also provide nurses, dietitians, and public health community educators with an opportunity to understand people's lifestyle choices (Liu, Mei, Hanauer, Zheng and Lee, 2016). As health facilities begin to shift their focus to population health analytics, local public health departments also use Twitter to disseminate diabetes information regarding lifestyle behaviors (Harris et al., 2013). The inclusion of two text mining methods - topic modeling and sentiment analysis - allows us to explore the lifestyle behaviors of diabetes in a unique way. More importantly, this computational approach lends itself to the discovery of latent topics.

METHODOLOGY

Data Collection and Cleaning

The Twitter data used for this study was collected using the Application Program Interface for Twitter. The twitteR package within the R program software suite was used to collect the data. This Twitter streaming approach has been used in other studies (Collins and Karami, 2018; Paul and Dredze, 2014). The data were collected from December 2016 to February 2017. These months were chosen as they reflect people's behavior change with the connotation involving a "new year" to incorporate healthy eating and increase physical activity (Turner-McGrievy and Beets, 2015). Diabetes - the traditional and hashtag spelling - was the query term used for data collection. Cleaning the data is an essential and often overlooked aspect of computational content analysis. Properly prepared data impacts the knowledge derived from the analytical process. It also reduces the level of noise and text that has no meaning (Leetaru, 2012). Stop words (*of*, *the*, *&*, and *in*), retweets, and tweets with an URL were removed during the data cleaning process.

Sentiment Analysis

To determine if the tweets were positive or negative, we used the sentiment analysis text mining method. As noted in the previous section, sentiment analysis is crucial with identifying positive and negative sentiments concerning entities (Salas-Zárate et al., 2017). A learning-based and lexicon-based approach is commonly used to conduct sentiment analysis. The lexicon-based method uses a pre-established dictionary of positive and negative words to find the positive and negative frequency of words within a corpus of data (Medhat, Hassan and Korashy, 2014). This study used the Linguistic Inquiry and Word Count (LIWC) software to perform the sentiment analysis (see example tweets in Figure 1). The lexicon-based approach used within LIWC has effectively identified the positive and negative sentiments in similar studies (Webb et al., 2018; Tumasjan, Sprenger, Sandner and Welpe, 2010). This portion of the methodology allows us to identify the positive and negative concerns regarding diabetes and lifestyle choices.

Clustering Topics and Characterizing the Data

This research study uses an unsupervised machine learning approach to characterize the topics (perform topic modeling). There are other unsupervised machine learning models used on similar unstructured health data (Paul and Dredze, 2012); however, LDA is a well-known and a proven topic model approach to characterize topics (Blei et al., 2003, Comito, Pizzuti and Procopio, 2018). LDA makes the probabilistic assumption that a corpus of data reflects latent topics. As a result, these unknown categories - or messages - are grouped together based on their probabilistic similarity. Characterization (labeling) of the topics was conducted by analyzing the overall topic results as being related or unrelated to diabetes. Once those topics were removed, two researchers characterize each topic as being related or unrelated to diabetes. Lastly, the specific characterization was provided for the tweets that were agreed upon in the previous step.

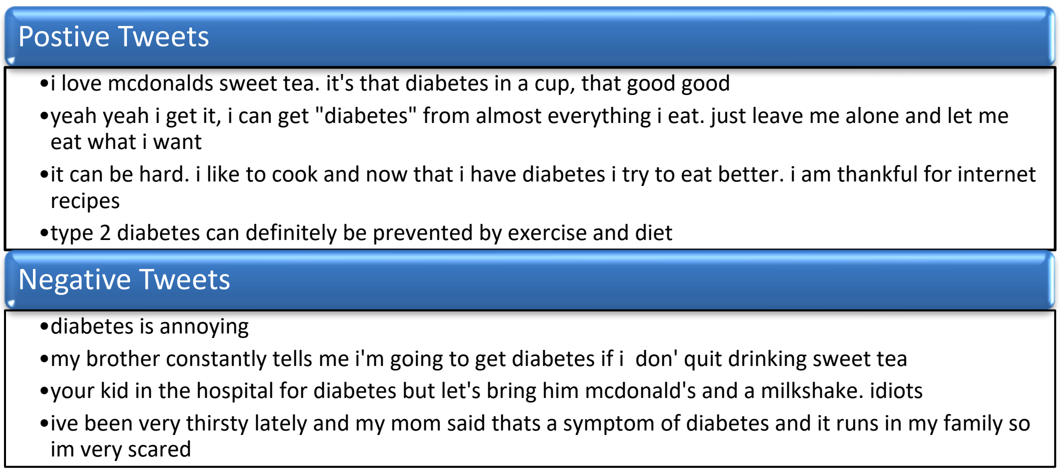

Figure 1: A Sample of Diabetes-Related Tweets for each Sentiment

RESULTS & DISCUSSION

While previous studies have incorporated prior knowledge in the development of various topic models, we chose to derive topics without prior knowledge using the default settings within Mallet[1]. This provides a more inductive approach for topic discovery and allows an exploratory analysis of the identified topics for insightful patterns (Chang, Boyd-Graber, Gerrish, Wang & Blei, 2009). A total of 121, 827 diabetes-related tweets were collected during the three-month period. The number of topics was set at 50, 75, and 100 to examine the topic composition. Ultimately, we decided that 100 topics provided an adequate representation of the topics. After the removal of topics deemed incomprehensive, 89 negative topics and 81 positive topics were used for binary classification of being related or unrelated to diabetes. If researchers agreed on a topic, it was designated with one; topics were denoted with a zero if there was disagreement. Results show that 58.4% of the positive topics were agreed upon as being related to diabetes – 55.6% of the negative topics were agreed upon as being related to diabetes. When we factor in chance agreement, our Cohen's kappa show there was not much agreement; the positive topic is k = .09 and k =.19 among the negative topics.

Before we explain the overall results and lifestyle behavior related results, it is imperative that we provide our definition of lifestyle. We define lifestyle behaviors as those behaviors associated with dieting (Diet) and exercising (Exercise) that can be attributed to diabetes but are modifiable to reduce peoples' risk of being diagnosed with type II diabetes (Flegal, Carroll, Kit, and Ogden 2012; Wing, Goldstein, Acton, Birch, Jakicic, Sallis, Smith-West, Jeffery and Surwit, 2001). Therefore, they can be categorized as:

- Diet = Topics were considered part of this category that pertains to food, beverages, or eating. For example, T1 (Topic number one) in Table 1 - *diabetes, syrup, corn, fructose, artificial* – words represent food items and ingredients that can be found in food items with high sugar calorie consumption. T16 in Table 2 references a type of diet or low carb diet based on the words of l*ow, carb, fat, intake, calories.*

[1] http://mallet.cs.umass.edu/

- Exercise = This category describes the lifestyle within the context of exercising that includes temporal dynamics and active behavior such as an active lunch break.

When examining the five most frequent positive topics, Food, Diabetes Care, Chronic Conditions, Diabetes Outcome, and Physiology were the most frequent characterized positive topics. As seen in Table 1, many of the foods and drinks representing the food topic lack nutritional value. It is important to maintain proper cardiovascular health; however, the American Heart Association has discussed the association between sugar-sweetened beverages and obesity (Lloyd-Jones et al., 2010). Diabetes care appears to be positively related with user feelings toward managing chronic conditions and improving quality.

Positive Topic Labels	Topic Words (characterizing words)	
Food (17.5%)	**T1**	*Surprised, sweets, sugar, good, cake*
	T2	*Diabetes, syrup, corn, fructose, artificial*
	T3	*Sugar, drink, water, rice, egg*
Diabetes Care (12.5%)	**T8**	*Care, supplies, diabetes, insulin, health*
	T9	*Insulin, meetings, care, talking, stigma*
	T10	*Pump, insulin, treatment, medicine, dexcom*
Chronic Condition (10%)	**T13**	*Disease, cancer, health, obesity, kidney*
	T14	*Free, high, blood, pressure, cholesterol*
	T15	*Cancer, aids, cure, treatment, lupus*
Diabetes Outcome (7.5%)	**T17**	*Time, amputated, foot, legs, fatter*
	T18	*Care, health, patients, outcomes, obesity*
	T19	*People, healthcare, quality, system, management*
Physiology (7.5%)	**T20**	*Hope, family, good, feel, happy*
	T21	*Type, years, diagnosed, today, happy*
	T22	*Eat, sugar, sweet, blood, anemic*

Table 1: Sample of Positive Diabetes Topics

The five most frequent negatives topics discovered includes Physiology, Diabetes Outcome, Chronic Condition, Food, and Diet (see Table 2 below). Physiology for this research incorporates metabolism, emotions, mental health, endorphins, and the general wellness of an individual. The high frequency of the chronic condition label with the negative sentiments regarding diabetes also supports prior researcher and the health consequences attributed with diabetes, arthritis, and increase in high cholesterol (Kim and Basu, 2016, p. 603). Dalen et al. (2010) work examined the impact of negative emotions (physiology) experienced by people. Negative emotions are used as an impulsive style of coping that can lead to an unintended lifestyle choice of consuming excess calories in an automatic and dissociative fashion.

We also examined the distribution of characterizations that were identified for understanding lifestyle behaviors and its contribution to diabetes. Overall, lifestyle related topics accounted for roughly 25% of the topics identified in the corpus of data. While the lifestyle topics were not a significant percentage of the topics, the results show that information is being shared within Twitter that provides insightful information to peoples' lifestyle choices. More importantly, this allows an alternative approach to understanding lifestyle choices that enable or prohibit people's ability to engage in proper dieting or exercising. Categorizing the topics based on diet and exercise, lifestyle choices for diet represented 75% of the topics. These findings are consistent with a similar study that identified diet having the highest number of topics (Karami, Dahl, Turner-McGrievy, Kharrazi, and Shaw, 2018).

People did not share as much information about their exercise lifestyle. However, despite this low number, Twitter health conversations and the CDC statistics have been documented with a strong correlation between the two data sources (Prier, Smith, Giraud-Carrier and Hanson, 2011). Health economics was a negative latent topic identified. Characterizing words such as *insulin, price, cost, metformin, cuts, prices* may speak to the concern users have with medical cost and managing their diabetes. Self-monitoring was identified as a positive latent topic. Glucose monitoring systems that communicate with other wireless diabetes management devices are technological advances that people feel positive about.

Negative Topic Labels	Topic Words	
Physiology (18.8%)	**T1**	*Bad, day, feel, insulin, hate*
	T2	*Insulin, resistance, beta, pancreas, mellitus*
	T3	*Blood, pressure, depression, diseases, family*
Diabetes Outcome (12.5%)	**T7**	*Weight, lost, foot leg, pounds*
	T8	*Long, term, complications, fatal, memory*
	T9	*Failure, kidney, damage, liver, blindness*
Chronic Condition (9.4)	**T11**	*Insipidus, trauma, urine, nervous, infection*
	T12	*Risk, type, obesity, cancer, hypertension*
	T13	*Cancer, sugar, sickle, arthritis, inflammation*
Food (6.3%)	**T14**	*Sweet, rice, drinks, candy, sugary*
	T15	*Eating, lifestyle, habits, choices, unhealthy*
Diet (6.3%)	**T16**	*Low, carb, fat, intake, calories*
	T17	*Type, diet, horrible, fat, obesity*

Table 1: Sample of Negative Diabetes Topics

Lifestyle Issue	Distribution Among Lifestyle Topics (%)	Distribution Among All Topics (%)
Food	56.25%	12.5%
Diet	18.75%	4.2%
Physical Activity	12.5%	2.8%
Weight Management	12.5%	2.8%

Table 2: Collective Distribution of Lifestyle Topics

CONCLUSION

This research has limitations. First, it does account for geographic locations with respect to tweet volume and geographical differences. The analysis is based on tweets from one seasonal period and does not fully capture the sentiments over an extended period. Twitter data is noisy – we did not account for health care organizations and bots that may be represented in the dataset. Lastly, the identification of topics was conducted by non-clinical and allied health professionals. Future research will incorporate experts from the medical domain to classify the tweets.

This exploratory study demonstrates that Twitter can be used as a complementary data source to study diabetes and lifestyle behaviors. Unstructured social media analysis in identifying lifestyle behaviors of diabetic patients using topic modelling is helpful; however, further research is required to increase the effectiveness of the existing knowledge. Due to the 'unquantifiable' limitation, there is a pressing need for incorporating predictive modelling approaches to this study in order to quantify our findings and make informed decisions to help reduce the impact of diabetes on the population. This would involve combining unstructured data with structured data about diabetic patients and using a discriminative classification technique such as Support Vector Machine (SVM).

REFERENCES

1. American Diabetes Association. (2018) *The cost of diabetes*, Retrieved from http://www.diabetes.org/advocacy/news-events/cost-of-diabetes.htm
2. Blei, D. M., Ng, A. Y and Jordan, M. I. (2003) Latent Dirichlet Allocation. *Journal of machine Learning research*, 3, 993-1022.
3. Blei, D. M. (2012) Probabilistic topic models, *Communications of the ACM, 55*(4), 77-84.
4. Centers for Disease Control and Prevention (CDC). About chronic diseases. Retrieved from https://www.cdc.gov/chronicdisease/about/index.htm
5. CDC. (2018) *Diabetes*, Retrieved from https://www.cdc.gov/media/presskits/aahd/diabetes.pdf
6. CDC. (2018) CDC *Diabetes Report*, Retrieved from https://www.cdc.gov/media/releases/2017/p0718-diabetes-report.html
7. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L. and Blei, D. M. (2009) Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems* 22, 288-296. https://doi.org/10.1.1.100.1089
8. Chew, C. and Eysenbach, G. (2010), Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PloS one, 5*(11), e14118.
9. Collins, M. and Karami, A. Social media analysis for organizations: Us Northeastern public And state libraries case study, *SAIS 2018 Proceedings*, 2018, 31.
10. Comito, C., Pizzuti, C. and Procopio, N. (2018) How people talk about health? Detecting health topics from Twitter streams. In *BDIOT*, 1-6.
11. Dalen, J., Smith, B. W., Shelley, B. M., Sloan, A. L., Leahigh, L. and Begay, D. (2010), Pilot study: Mindful Eating and Living (MEAL): weight, eating behavior, and psychological outcomes associated with a mindfulness-based intervention for people with obesity. *Complementary Therapies in Medicine*, 18(6), 260-264.
12. Flegal, K. M., Carroll, M. D., Kit, B. K. and Ogden, C. L. (2012) Prevalence of obesity and trends in the distribution of body mass index among US adults, 1999-2010, *The Journal of the American Medical Association, 307*(5), 491-497.
13. Gabarron, E., Dorronzoro, E., Rivera-Romero, O. and Wynn, R. (2018) Diabetes on Twitter: A sentiment analysis, *Journal of Diabetes Science and Technology*.
14. Guo, L., Vargo, C. J., Pan, Z., Ding, W. and Ishwar, P. (2016) Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling *Journalism & Mass Communication Quarterly*, 93(2), 332-359.
15. Harris, J. K., Moreland-Russell, S., Tabak, R. G., Ruhr, L. R. and Maier, R. C. (2014) Communication about childhood obesity on Twitter, *American Journal of Public Health, 104*(7), e62-e69.
16. Harris, J. K., Mueller, N. L., Snider, D. and Haire-Joshu, D. (2013) Local health department use of Twitter to disseminate diabetes information, United States, *Preventing chronic disease*, 10.
17. Karami, A., Dahl, A. A., Turner-McGrievy, G., Kharrazi, H. and Shaw, G. (2018) Characterizing diabetes, diet, exercise, and obesity comments on Twitter, *International Journal of Information Management, 38*(1), 1-6.
18. Kim, D. D., and Basu, A. (2016) Estimating the medical care costs of obesity in the United States: Systematic review, meta-analysis, and empirical analysis, *Value in Health, 19*(5), 602-613.
19. Leetaru, K. H. (2012) *Data mining methods for the Content Analyst: An introduction to the computational analysis of content*, Taylor and Francis Group, New York, New York.

20. Lloyd-Jones, D. M., Hong, Y., Labarthe, D., Mozaffarian, D., Appel, L. J., Van Horn, L., ... & Arnett, D. K. (2010) Defining and setting national goals for cardiovascular health promotion and disease reduction: the American Heart Association's strategic Impact Goal through 2020 and beyond, *Circulation*, 121,4, 586-613.
21. Liu, Y., Mei, Q., Hanauer, D. A., Zheng, K. and Lee, J. M. (2016) Use of social media in the diabetes community: An exploratory analysis of diabetes-related tweets, *JMIR Diabetes*, *1*,2, e4.
22. Medhat, W., Hassan, A. and Korashy, H. (2014) Sentiment analysis algorithms and applications: A survey, *Ain Shams Engineering Journal*, *5*, 4, 1093-1113.
23. Paul, M. J., and Dredze, M. (2012) A model for mining public health topics from Twitter, *Health*, 11, 16-6.
24. Paul, M. J., and Dredze, M. (2014) Discovering health topics in social media using topic models. *PloS one*, *9*,8.
25. Popkin B. M. (2015) Nutrition Transition and the Global Diabetes Epidemic. *Current Diabetes Reports*, *15*,9, 64.
26. Prier, K. W., Smith, M. S., Giraud-Carrier, C. and Hanson, C. L. (2011) Identifying health-related topics on twitter. In *International conference on social computing, behavioral-cultural modeling, and prediction*, Berlin, Heidelberg.
27. Salas-Zárate, M. D. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodriguez-Garcia, M. A. and Valencia-García, R. (2017) Sentiment analysis on tweets about diabetes: an aspect-level approach, *Computational and mathematical methods in medicine*, 2017. Retrieved from https://doi.org/10.1155/2017/5140631
28. Sinnenberg, L., Mancheno, C., Barg, F., Asch, D., Rivard, C., Horst-Martz,….Merchant, R. (2018) Content analysis of metaphors about hypertension and diabetes on Twitter: Exploratory mixed-methods study, *JMIR Diabetes*, *3*, 4.
29. Sivic, J., Russell, B. C., Zisserman, A., Freeman, W. T. and Efros, A. A. (2008) Unsupervised discovery of visual object class hierarchies. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. *IEEE Conference*.
30. Tufts, C., Polsky, D., Volpp, K., Groeneveld, P., Ungar, L., Merchant, R. and Pelullo, A. (2018) Characterizing tweet volume and content about common health conditions across Pennsylvania: Retrospective analysis. *JMIR Public Health and Surveillance*, *4*, 4.
31. Tumasjan, A., Sprenger, T. O., Sandner, P. G. and Welpe, I. M. (2010) Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth International AAAI Conference on Weblogs and Social Media*, 178-185.
32. Webb, F., Karami, A. and Kitzie, V. (2018) Characterizing diseases and disorders in gay Users' tweets, *SAIS 2018 Proceedings*, 31.
33. Wing, R. R., Goldstein, M. G., Acton, K. J., Birch, L. L., Jakicic, J. M., Sallis, J. F., ...and Surwit, R. S. (2001) Behavioral science research in diabetes: lifestyle changes related to obesity, eating behavior, and physical activity. *Diabetes care*, *24*, 1, 117-123.
34. Xu, X., Litchman, M., Gee, P., Whatcott, W., Chacon, L., Holmes, J. and Srinivasan, S. (2018) Predicting prediabetes through Facebook postings: Protocol for a mixed-methods study. *JMIR Research Protocols*, *7*, 12.
35. Yin, Z., Fabbri, D., Rosenbloom, S. T. and Malin, B. (2015) A scalable framework to detect personal health mentions on Twitter, *Journal of Medical Internet Research*, 17, 6.