Science Progress and Research

Review Article

ISSN: 2635-0955

DOI: <https://doi.org/10.52152/spr/2021.119>

Usage of social media Twitter in the Intervention of Diabetes Mellitus

Medit Leonard¹, Bethzy Williams²

¹Dept of Health information, The College of St. Scholastica 1200 Kenwood Ave., Duluth, USA

²Dept of Health Information Management, Davenport University, USA

***Corresponding Author:** Medit Leonard, Dept of Health information, The College of St. Scholastica 1200 Kenwood Ave., Duluth, USA. Email: editorjcerr867@gmail.com

Submission Date: 01/05/2021

Acceptance Date: 20/05/2021

Published Date: 28/05/2021

ABSTRACT

Social networking sites have been a common forum for exchanging health-related insights and information. This study aims to look at Twitter use in the intervention of diabetes. Specifically, utilising a prior analysis as a reference, we use a revised approach to analyse trends in the existing use of hash-tags, trending hash-tags, and the incidence of diabetes-related tweets. Our findings indicate that the diabetes population on Twitter has grown significantly over time, as well as proof that this community is becoming more capable of spreading diabetes-related health information. An enhanced system for storing, cleaning, and reviewing Twitter data relevant to diabetes, as well as the use of regular expressions to categorise subsets of tweets, are among our computational contributions. To recognise tweets from diabetic patients, we built a model focused on word embedding and long short-term memory.

Keywords: Type II Diabetes Mellitus, Social Media, Intervention, Twitter

INTRODUCTION

According to recent studies, social network data may be a reliable source of information on public wellbeing, cancer, and health-related topics. People who use social networking sites, especially Twitter, are becoming more prevalent. Local health departments, in particular, are utilising these resources to educate and spread knowledge about sickness and medical problems

to those needing awareness and social care about a variety of illnesses and diseases. Furthermore, using the Twitter site to address the big diabetes issues will help us pinpoint geographically where diabetes is more prevalent. About the fact that medical providers do not use Twitter to recognise patients, healthcare practitioners, public health authorities, and epidemiologists will be able to

better grasp the incidence of diabetes in various regional areas and prepare for different strategies. ^{[1], [2], [3]} This study focuses on diabetes, an illness that is both an increasing area of science in social network data analysis and a major public health concern. The Twitter API has grown in importance as a platform for collecting real-time data for health monitoring analysis, and as we can see, the diabetes online group is becoming increasingly involved on Twitter. As a result, finding reliable and successful ways to distinguish diabetic users or users with similar ties to diabetic patients is critical. As a consequence, we've created a system that incorporates both word embedding and long short-term memory.

Textual details from the tweet text and metadata such as the number of retweets and the date the tweet was sent are two forms of data that may be included as attributes from each tweet. Even though the metadata lacks rich semantic detail, it can be used as a feature without requiring any conversion or transformation. ^{[2], [3]} The textual data of the tweet, on the other hand, provides a large volume of linguistic knowledge about health problems. Furthermore, unusual linguistic phrases in tweet text may trigger natural language processing strategies to fail, as the sections of the speech (POS) can be wrongly tagged, causing the function data to be misled and the classifiers to perform poorly. We used an LSTM neural network with word embedding to forecast diabetic patients' tweets in addition to normal speech to classify diabetic patients' tweets. ^{[4], [5]} Since regular phrases for recognising diabetic patients, in this case, rely heavily on descriptive terms, this technique may not be suitable for identifying diabetic patients' tweets, as diabetic patients on Twitter use a variety of expressions and descriptions to convey or describe their encounters with the problems they encounter as a result of their diabetes diagnosis. Furthermore, since Twitter users are limited to 140 characters, they are more inventive in producing texts that do not often obey grammatical and spelling rules; hence, normal speech is ineffective because Twitter users prefer to use more unusual language in their tweets. ^{[5], [6]} To address this constraint, we created a system focused on word embedding and the LSTM neural network, in which tweet texts are interpreted as dense vectors, which are then fed to the LSTM neural network as sequences to predict diabetic patients' tweets. Our main goal is to provide a summary of diabetes-related Twitter behaviour by explaining tweet frequency, timing, user characteristics, and geography. This data is compared to a previous analysis that looked at changes in patterns linked to diabetes tweets over

time and used hashtags to forecast diabetes knowledge.

The following are the key results of this experiment:

- The diabetes culture on Twitter has grown over time and will continue to expand.
- Evidence of an uptick in insulin-related tweets.
- There's been an uptick in tweets about various cases of diabetes. Twitter donations from more countries to the online diabetes network.
- The use of regular expressions rather than manual execution to categorise a large data collection of tweets is a viable option.
- Implementation of the LSTM neural network and word embedding to test the model's success in predicting diabetic patient tweets.

MATERIAL AND METHODS

We used the Twitter streaming API to capture info. ^[7] According to Serrano et al. ^[8], Twitter receives roughly 8000 messages per second, of which we keep between 1 and 40% of near real-time tweets. ^[9] Since there is too much data, we built a system to store and process it directly from the API, making data cleaning more effective. Other options for pre-indexing data proposed by researchers include Elasticsearch and Kibana. ^[10] The Twitter streaming API was accessed using Python and the tweepy ^[11] libraries, and messages were targeted using keywords. The JSON answer from the Twitter API stream was saved to MongoDB using Python and the pymongo ^[4] library. ^[12] Tableau is the platform of choice for data cleaning and visualisation since it handles data quicker than other software options due to improved indexing. In addition, as seen in Fig. 1, it is directly linked to MongoDB through the BI connector application. 1 that allows for near-real-time simulation

We accumulated more than 4.5 million diabetes-related tweets over 1 year using the following API targeting question words and hashtags shown in Table 1 in comparison to the previous sample that we are replicating, which had 1,368,575 tweets over 2 years.

Data Collection

We required improved computing or a different approach to processing and cleaning our data than the previous study due to the scale of our data collection. We used Tableau Prep to optimise our data and retrieve valuable fields

from the database, such as email, usernames, dates, times, and geo-location tweets. It is important to break the time into three fields-day, month, and year-while analysing the data to obtain results; this speeds up the method of plotting frequency graphs. We also got rid of tweets that had no text area. We produced a reproduction of # dsma tweets for statistical analysis of # dsma, as well as a different copy on World Diabetes Day to review related data patterns and conduct statistical analysis on that particular day. For this package, we sorted tweets by year, i.e., on November 14th, we received 1,10,113 tweets on World Diabetes Day for study, accounting for nearly 10% of all tweets

Specific Methodology

Our scheme creates a term index vector space model (VSM) from the unlabeled tweet corpus' raw textual data, where each vector represents the index of a single term or token in the document. The VSM is then used to create a series of vector representations for each message, which is then fed into an LSTM neural network, which performs a binary classification: diabetic patient tweet or non-diabetic patient tweet.

Illustration of a tweet

Recently, strong accuracies in detecting artifacts from raw image data were achieved using a deep neural network model. As a result, we developed a method in which raw textual data is fed into a deep neural network-based classifier. The raw textual data, unlike image data, cannot be represented explicitly in the form of a dense vector since the tokens for each tweet document have varying sizes and are incompatible with traditional classifiers. As a result, we're using a distributed representation of words (word embedding) to describe individual terms in the tweet text as a dense array, and we're using a special object in our language called padding ("pad") to reflect a list of all the specific objects in the tweet text. We may obtain a fixed duration input by padding. Since there are no textual equivalents for this concept, the word cannot be expressed as vectors. To address this problem, the words in the tweets were substituted with the indices of terms in the language, resulting in a series of indices (positive integers) rather than a sequence of textual terms in each tweet. Finally, on the term index sequences, a dense vector can be created, and their representations are identical to the original tweet document. The first row of Table 2 displays the order of words (or tokens) in the tweet. The second row is the tokenized representation of indices I (term) of respective terms in the vocabulary, and the last row is the

received during that month. In total, 4,111,545 Twitter data points were included in our research.

Table 1 Hashtags used to collect the specific tweets

Hashtags
#Diabetes Mellitus
#Type I DM, #t1D
#Type II DM, #t2D
#Blood Sugar
#Diabetes
#DM
#Insulin Pump

sequence of actual term indices whose values are determined by the term positions in the vocabulary. As a result, in our approach, a vector space model was produced from the term index representation of tweets, in which a dense vector was made. We also hope that such a reflection of tweets would be able to provide classifiers with meaningful knowledge contained in each message.

We used a long short term memory (LSTM) neural network to turn our classification into a sequence classification query, which the LSTM model can handle well. We have used the distributed representation of the tweet text (word embedding) as an input to our LSTM classifier, which was trained on a collection of annotated tweets as the training set and then tested on another set of tweets as the test set.

We used the unlabeled tweets to construct the vocabulary and vector space model (VSM), which described the tweets as dense vector sequences. The tweets were then categorised using the LSTM neural network. To begin, a corpus of 600 unlabeled diabetes-related tweets was randomly selected and preprocessed to exclude retweets and non-English tweets. The preprocessed 590 tweets were used to construct a vector space model, which included removing punctuation, deleting duplicates, and removing tweets with URLs. Finally, a vocabulary of specific words was assembled, and a vector space model was generated from the preprocessed 590 tweets.

Data from Twitter

The two vocabularies and vector space models for the 490 and 1980 tweets were developed using corpora of preprocessed unlabeled tweets. Annotation guidelines identify ways to classify diabetic patients and provide samples of diabetic and nondiabetic tweets for building corpora of annotated tweets. The unlabeled tweets were divided into four sections

using this rule, allowing four annotators to independently mark each series of tweets; the annotated tweets were then re-evaluated and updated by the paper's author and the annotators. This approach was tested using a tenfold stratified cross-validation.

RESULTS

The section "Changes in Tweet Percent" displays the percentage difference in tweets from the previous sample and now. The phrase "Diabetes," which accounted for 92.5% of total tweets in the previous survey, has fallen from 42 to 62.06% of tweets, whereas the term "insulin," which accounted for 8.1% of total tweets in the

previous report, has risen to 32.11%. In addition, rather than utilising generalisations, people are being more type-specific in their tweets. Type1 tweets (including tweets with # T1d, # Type1 and # Type1diabetes) and Type2 tweets (including tweets with # T2d, # Type2 and # Type2diabetes) were 17,011 and 6117, respectively, in the previous report. These numbers have risen to 82,119 (Class 1) and 52,111 (Type 2) Fig 1. Fig 1 compares the amount of diabetes-related tweets by country of birth, as well as shifts within the prior and current studies' periods. Both our research and the previous study found that the United States generates the most tweets; moreover, we found that India has the highest rise in% in diabetes-related tweets.

Keywords	1	2	3	4
	Tweets count	Tweet count %	Users count	Users count %
Diabetes	1,200,268	87.7	748,001	89.6
Insulin	83,820	6.1	59,728	7.2
#Diabetes	165,868	12.1	67,229	8.1
Glucose	60,033	4.4	46,357	5.6
#T1d	9,040	0.7	3835	0.5
#Type1	3211	0.2	1800	0.2
#Doc	27,616	2	16,457	2
Blood glucose	10,212	0.7	6904	0.8
#Type1diabetes	1812	0.1	1248	0.1
#Type2	2905	0.2	1468	0.2
#Bloodsugar	1718	0.1	1213	0.1
#Type2diabetes	1388	0.1	1035	0.1
#T2d	935	0.1	452	0.1
Insulin pump	5179	0.4	4061	0.5
#Dsma	11,757	0.9	1424	0.2
#Dblog	5711	0.4	1132	0.1
#Wearenotwaiting	327	<(0.01)	183	<(0.01)
#Bgnow	2470	0.2	753	0.1
#Diyps	132	<(0.01)	50	<(0.01)
#Cwd2014/2018	7	<(0.01)	7	<(0.01)
#Showmeyourpump	932	0.1	645	0.1

Figure 1. Assessment of Twitter data based on used Hashtags and Keywords

Fig 2 shows that the number of diabetes-related tweets is declining in the countries outlined in red, despite an increase in total figures. In addition, two countries had major decreases in diabetes tweets: Indonesia, which went from 5355 to 890, and Venezuela, which went from 1172 to 248. There are substantial shifts, but the Dominican Republic's per cent decline is the largest, at 91.96%, while countries illustrated in oranges, such as the United Kingdom, Australia, and India, have seen an increase in numbers. Since the countries

highlighted in yellow lacked prior statistical evidence, no comparison could be made; therefore, the 'Previous Study' and 'Differences' columns in Fig 2 are identified as 'Not Accessible'. According to recent figures, 14 additional countries have been introduced to the chart, each with more than 100 tweets. These points to a possible increase in diabetes consciousness within populations across various countries. India is already becoming one of the most dangerous countries on the planet.

India	220
Australia	218
Spain	325
United Kingdom	1,897
South Africa	185
Ireland	107
Argentina	260
Canada	590
Colombia	167
United States	10,047
Nigeria	299
Brazil	816
Mexico	1,042
Netherlands	189
Philippines	439
Sweden	105
Chile	223
Malasiya	611
Singapore	147
Ghana	350
Venezuela	1,172
Indonesia	5,355
Dominican Republic	199

Figure 2. Amount of diabetes-related tweets in the various countries

Figure 3 reveals that for both sample dates, November has the highest amount of diabetes-

related tweets owing to the date of World Diabetes Day.

Figure 3. Amount of Tweets in the month of Nov 2020

DISCUSSION

The usage of the insulin hashtag has risen between 6.1 and 25.8% and may reflect increased understanding and debate on insulin online. Additionally, there is an uptick in the usage of hashtags such as # T1d and # T2d that indicates a possible increase in consumer awareness. In addition to four countries in the preventive sample, there was a spike in diabetes-related twitter participation in all other countries and some new countries contributing substantially towards the data collection. Finally, we were able to categorise 500 random tweets with diabetes through normal speech and were able to categorise tweets with diabetes efficiently, except for tweets relevant to Dsma.^{[13], [14], [15]}

Based on our findings, our model is working well, since the term embedded with the LSTM classifiers may have unearthed rich semantic data from various tweet phrases describing patients' interactions with diabetes and thereby leading to better classification. In our experiment, taking the average weight of each label increased the efficiency of reminder accuracy and f1 score relative to the findings without taking the weight of each label into account. The global diabetes association on Twitter has grown over time and employs the strength of the social network to collaborate and spread the experience of diabetes. Future research into the topic will benefit from the comparison of hashtag studies which can be expanded to include advanced text processing techniques.^[16] We can see that the usage of social networking channels is growing faster to store, pre-index, search, and query data in real-time as the exponential increase of data results. In addition, because social media spam is identified as a challenge to analyse using social media methods, a Spam Identification tool that focuses on lexical information may help improve results. We also examined a method of detecting tweets for diabetic patients that merged word embedding with neural network deep learning in our vector space model study. According to our findings, the LSTM model performed well in recognizing diabetic patients in tweets.

CONCLUSION

In the future, the LSTM model with a word embedding layer will be used to process and analyse free social network text with simplified words or phrases that do not follow grammar and orthography in either language. In contrast to the standard expression system, complicated words might not be required to search a tweeter document. The benefits of the analysis involve the treatment of a reasonably large number of

Twitter data in relatively less time, utilising the same criteria to help compare and understand pattern shifts in diabetes-focused Twitter data.

Conflict of Interest

Nil

Funding

Nil

REFERENCES

1. Eriksson-Backa K, Holmberg K, Ek S. Communicating diabetes and diets on twitter-a semantic content analysis. *Int J Networking Virtual Organ* 2016;16(1):8–24.
2. Federer LM, Joubert DJ. Providing library support for interactive scientific and biomedical visualizations with tableau. *J eSci Librariansh* 2018;7(1): e1120.
3. Gabarron E, Årsand E, Wynn R. Social media use in interventions for diabetes: a rapid evidence-based review. *J Med Internet Res* 2018;20(8):e10303.
4. German A, Salmeron S, Ha W, Henderson B. Mean web development: a tutorial for educators. In: *Proceedings of the 17th annual conference on information technology education*. ACM 2016;128–129.
5. Ghosh D, Guha R. What are we tweeting about obesity? Map-ping tweets with topic modelling and geographic information system. *Cartogr Geogr Inf Sci* 2013;40(2):90–102.
6. Liu Y, Mei Q, Hanauer DA, Zheng K, Lee JM. Use of social media in the diabetes community: an exploratory analysis of diabetes-related tweets. *JMIR Diabetes* 2016;1(2):e4.
7. Jones MT. *Process real-time big data with Twitter storm*. IBM Technical Library, New York 2013.
8. Serrano WC, Chren MM, Resneck JS, Aji NN, Pagoto S, Linos E. Online advertising for cancer prevention: Google ads and tanning beds. *JAMA Dermatol* 2016;152(1):101–102.
9. Shah N, Willick D, Mago V. A framework for social media data analytics using Elasticsearch and Kibana. *Wirel Netw*.2018;13:1–9.
10. Roesslein J. Tweepy documentation. Online <http://www.tweep yread thedo csio/en/v35>. 2009.

11. Reddy T, Giabbanelli PJ, Mago VK. The artificial facilitator: guiding participants in developing causal maps using voice-activated technologies. In: International conference on human-computer interaction. Springer 2019;111–129.
12. Kumar S, Morstatter F, Liu H. Twitter data analytics. Springer, Berlin. 2014.
13. Patel KD, Heppner A, Srivastava G, Mago V. Analyzing the use of Twitter by diabetes online community. In: ASONAM '19: international conference on advances in social networks analysis and mining, Vancouver, British Columbia, Canada, 2019; 937–944.
14. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. J Mach Learn Res 2011; 12:2825–2830.
15. Shaw RJ, Johnson CM. Health information seeking and social media use on the internet among people with diabetes. Online J Public Health Inform 2011; 3(1):1–9.
16. Taylor J, Pagliari C. Comprehensive scoping review of health research using social media data. BMJ Open 2018;8(12); e022931.