Pranav Shankar (psv6974)
Jayalakshmi Jain (jjk6400)
Prithvi SS (psx2851)

Introduction

We live in a age where we are more likely to look for answers online than go to our Doctor. Surveys have shown that social media affects the way consumers deal with their health (more than 40%). One in every five people have a healthcare app installed on their phone. An increasingly number of people have started sharing their health information online. This helps us monitor diseases, spread of diseases, symptoms, outbreak and more importantly factors contributing to the spread o the disease.

Traditionally the CDC and WHO collect this information for hospitals and other healthcare institutions but this is time consuming and expensive. However through social media, we can instantly at any given point in time, get data and process it in a matter of hours. We have chosen to use twitter as a platform for data. This is done because of twitter RESTful API which gives us an easy way to find tweets based on keywords.

We have chosen to work on diabetes which affect 1 in every 4 people in America. Whats more worrying is that 1 in every 3 people have pre diabetes. Diabetes is cause when the body simply cannot use or produce insulin. It is observed by raised glucose levels in the blood. This high level of glucose is what cause diseases and health complications. Diabetes requires constant doctors visit. This is why chronic diseases have gained increasingly popularity on social media where sufferers can find support groups and solutions to their diseases.

Related work

Analysis of health information through social media is relatively new. Although sentiment analysis and such has been there for quite some time, health information and social media is a relatively new field.

'Sentiment Analysis on Tweets about Diabetes' published in 2017 works on classifying tweets related to diabetes according to their sentiment.

Northwestern's own Kathy Lee has written a paper on detecting the outbreak of flu in the United States using twitter data. 'Read time Digital Flu surveillance using Twitter Data'

'Zika in Twitter', published in 2017 talks about the Zika virus outbreak. They analyse the geographical footprint, actors participating in the discourse as well as emerging concepts associated with the issue.

'A pattern-matched Twitter analysis of US cancer-patient sentiments', published 2016, talks about computing average happiness of patients suffering from cancer.

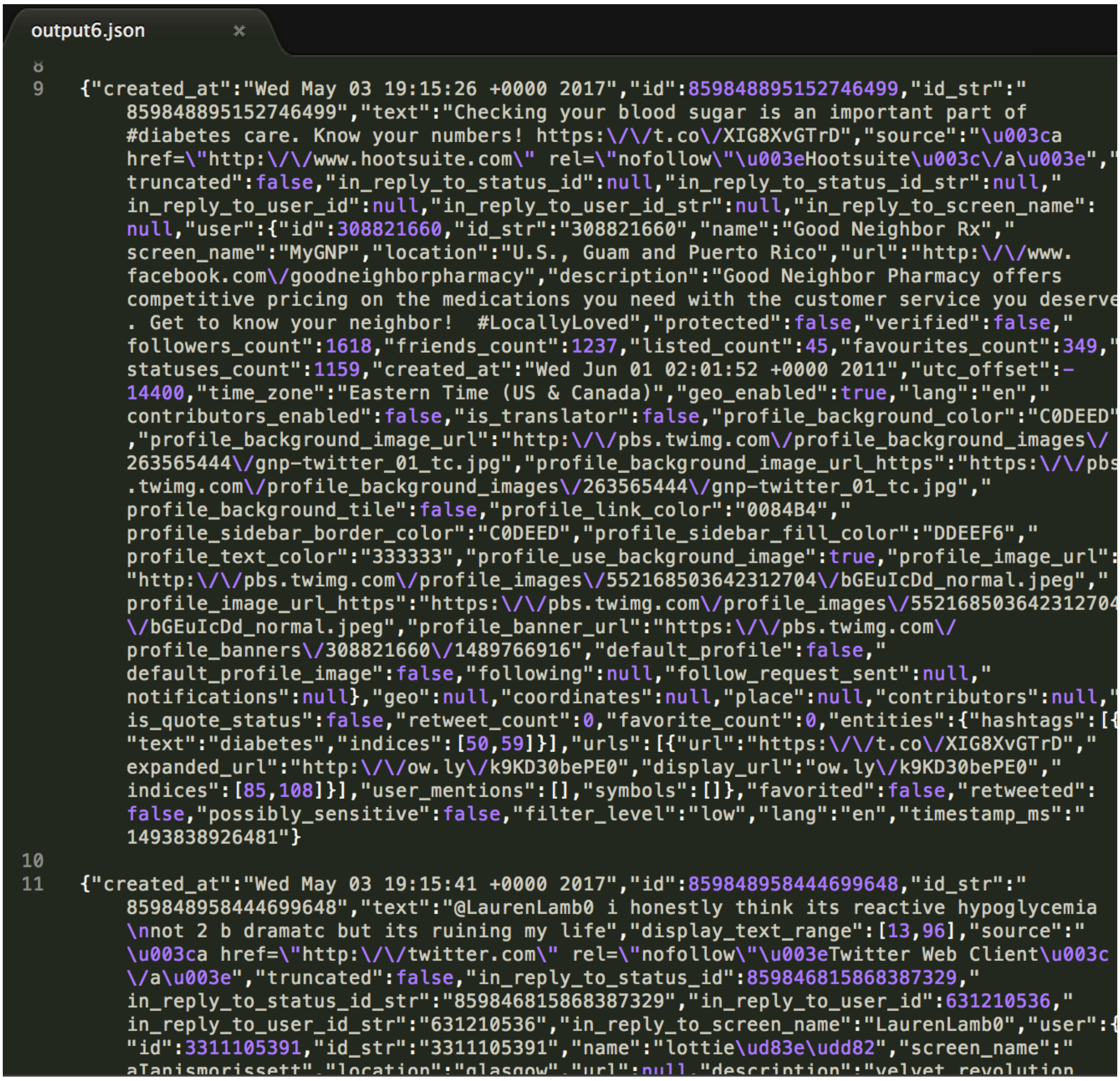
Most of the papers we have come across that track diseases on social media have been published very recently. There has been some sentiment analysis done on tweets that refer to diabetes. Some papers have tried to track diseases on social media but as far as we are aware no work has been done that closely reflects our own (diabetes).

Approach

There are 3 main states that involves analysis of diabetes on twitter. The first is Data Collection, Data Preprocessing and finally Analysis of the data.

Data Collection:

Tweets are primarily collected using the twitter Developer RESTful APIs. We are collecting tweets with certain keywords. Theres tweets are returned to us in a JSON format (refer Figure 1). The JSON includes the tweet as well as many attributes associated with the tweets. Some of the attributes are geo-tag, created_at, unique_id, time_zone, retweets_count, language etc. In a period of 1 month May 1st to may 31st we have collected millions of tweets. After segregating the tweets (using pandas) we ended up with 845,418 useful tweets.



```
{"created_at":"Wed May 03 19:15:26 +0000 2017","id":859848895152746499,"id_str":"
    859848895152746499","text":"Checking your blood sugar is an important part of
    #diabetes care. Know your numbers! https:\/\/t.co\/XIG8XvGTrD","source":"\u003ca
    href=\"http:\/\/www.hootsuite.com\" rel=\"nofollow\"\u003eHootsuite\u003c\/a\u003e","
    truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"
    in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":
    null,"user":{"id":308821660,"id_str":"308821660","name":"Good Neighbor Rx","
    screen_name":"MyGNP","location":"U.S., Guam and Puerto Rico","url":"http:\/\/www.
    facebook.com\/goodneighborpharmacy","description":"Good Neighbor Pharmacy offers
    competitive pricing on the medications you need with the customer service you deserve
    . Get to know your neighbor!  #LocallyLoved","protected":false,"verified":false,"
    followers_count":1618,"friends_count":1237,"listed_count":45,"favourites_count":349,"
    statuses_count":1159,"created_at":"Wed Jun 01 02:01:52 +0000 2011","utc_offset":-
    14400,"time_zone":"Eastern Time (US & Canada)","geo_enabled":true,"lang":"en","
    contributors_enabled":false,"is_translator":false,"profile_background_color":"C0DEED"
    ,"profile_background_image_url":"http:\/\/pbs.twimg.com\/profile_background_images\/
    263565444\/gnp-twitter_01_tc.jpg","profile_background_image_url_https":"https:\/\/pbs
    .twimg.com\/profile_background_images\/263565444\/gnp-twitter_01_tc.jpg","
    profile_background_tile":false,"profile_link_color":"0084B4","
    profile_sidebar_border_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","
    profile_text_color":"333333","profile_use_background_image":true,"profile_image_url":
    "http:\/\/pbs.twimg.com\/profile_images\/552168503642312704\/bGEuIcDd_normal.jpeg","
    profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/552168503642312704
    \/bGEuIcDd_normal.jpeg","profile_banner_url":"https:\/\/pbs.twimg.com\/
    profile_banners\/308821660\/1489766916","default_profile":false,"
    default_profile_image":false,"following":null,"follow_request_sent":null,"
    notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"
    is_quote_status":false,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[{
    "text":"diabetes","indices":[50,59]}],"urls":[{"url":"https:\/\/t.co\/XIG8XvGTrD","
    expanded_url":"http:\/\/ow.ly\/k9KD30bePE0","display_url":"ow.ly\/k9KD30bePE0","
    indices":[85,108]}],"user_mentions":[],"symbols":[]},"favorited":false,"retweeted":
    false,"possibly_sensitive":false,"filter_level":"low","lang":"en","timestamp_ms":"
    1493838926481"}

{"created_at":"Wed May 03 19:15:41 +0000 2017","id":859848958444699648,"id_str":"
    859848958444699648","text":"@LaurenLamb0 i honestly think its reactive hypoglycemia
    \nnot 2 b dramatc but its ruining my life","display_text_range":[13,96],"source":"
    \u003ca href=\"http:\/\/twitter.com\" rel=\"nofollow\"\u003eTwitter Web Client\u003c
    \/a\u003e","truncated":false,"in_reply_to_status_id":859846815868387329,"
    in_reply_to_status_id_str":"859846815868387329","in_reply_to_user_id":631210536,"
    in_reply_to_user_id_str":"631210536","in_reply_to_screen_name":"LaurenLamb0","user":{
    "id":3311105391,"id_str":"3311105391","name":"lottie\ud83e\udd82","screen_name":"
    aIanismorissett","location":"glasgow","url":null,"description":"velvet revolution
```

Figure 1: JSON file with tweets

The keywords we used to obtain these tweets include:-

- Diabetes
- Diabetic
- Type I
- Type II
- T1D
- T2D
- Blood Sugar
- Insulin
- Glucometer
- Glycemia
- Hypo glycemia
- Hyper glycemia
- Hypoglycemia
- Hyperglycemia
- #Diabetic
- #Diabetes
- Diabetic Diet

We have also collected data from specific users (organisations) whose concentration centers around diabetes. These users are

- World Diabetes Day (@WDD)
- Int. Diabetes Fed. (@IntDiabetesFed)
- Medtronic Diabetes (@MDT_Diabetes)
- Diabetes Hands Fndn (@diabeteshf)

Data preprocessing:

This is an important step towards obtaining a clean database. The results are only as good as the quality of the tweets

The tweets obtained in JSON are read using python and stored in a pandas data frame. This was done in batches of 20,000 tweets (in case something breaks). By using the ‘lang’ attributes we ensured that we kept only tweets in English.

For the purposes of geo locations, we stored tweets that did not have ‘null’ as the ‘place’ attribute for further analysis.

In order to use the ‘bag of words’, we had to remove retweets to prevent skewing the distribution. This was done by removing tweets that started with ‘RT’.

Methodology

Text modeling

The tweets we have collected can be used to analyse a number of metrics. We have successfully analysed the following metrics.

- Symptoms of diabetes
- Preferred Treatment for diabetes
- Preferred diabetic drugs (western medicine)
- Diabetic Diet
- Terms associated with diabetes

We obtain these by using ‘bag of words’ algorithm. In bag of words a text is represented as ‘bag’ of words. When this is done over thousands of tweets, we can obtain frequencies of word use.

An important concept here is the idea of stemming. In stemming we convert every word to its root word. This is done in order to avoid repetition of words that convey the same idea. For example in our data set, we used stemming in order to classify words such as ‘patients’ and ‘patient’ under the same category. Both these words are converted to their root forms and hence counted under one category.

We then run these words again known databases such as ‘symptoms of diabetes’, ‘treatment of diabetes’, ‘diabetes diet’ and so on. This helps us classify the most frequently occurring words into a given category.

Geographical modelling

We use our dataset of tweets that have a geotag. We use this geo tag information ‘Full name’ which includes the city and state of the tweets. We use this information to label how ‘concerned’ each state is about diabetes.

We are using tableau to do this. However an important factor to consider is the fact that the number of tweets coming out of a state is directly dependent on the number of people in the state. Thus we divide the value we get by the number of people in the state. This yields really small values which we cannot comprehend. In order for this data to make more sense, we normalise all the data to fit between 0 to 1.

Sentiment Analysis

We are using an API, ‘text-processing’. We pass the tweet through the API and it returns 2 attributes, label and probability. Label can be positive, negative or neutral and the likelihood of it being so (probability).

In order to check the accuracy we have written a test in python. We have manually labelled tweets with a sentiment. We send this tweet through the API and cross check with the results we get. This helps us determine the accuracy of the sentiment analysis of the tweet.

Temporal Analysis

Using pandas we have an attributed ‘created_at’. We are taking the length of the number of tweets (and retweets) on a particular day for each day. We then plot the number of tweets against time.

Results

Text modeling:

- Symptoms of diabetes: We found that weight changes is a very significant symptom of diabetes accounting for more than 50% of the symptoms (refer Figure 2)

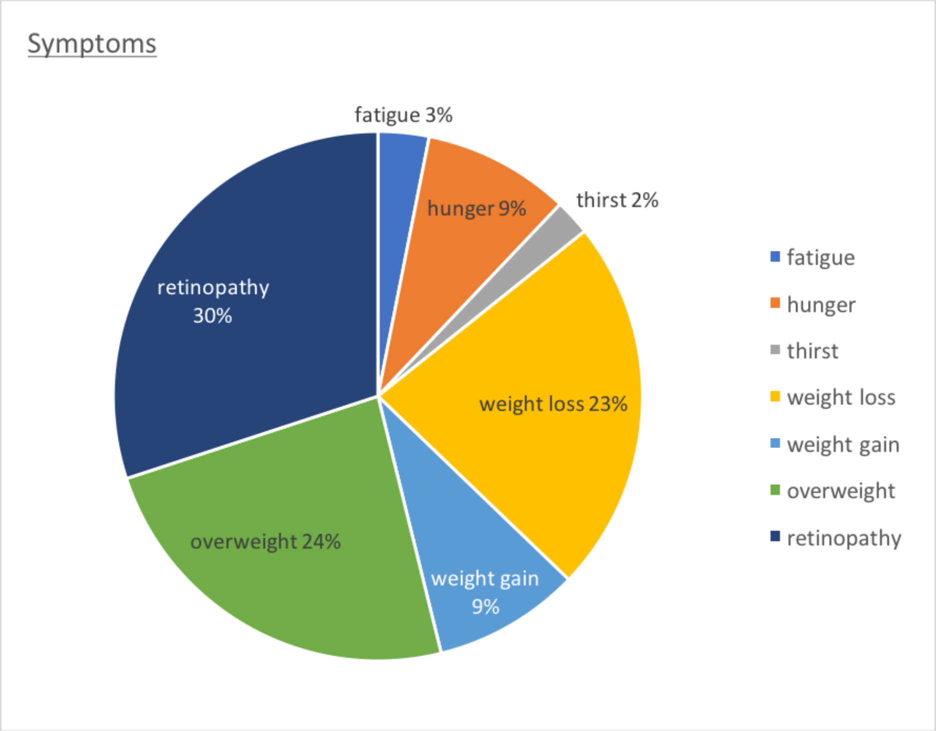


Figure 2: Symptoms of diabetes

- Preferred Treatment for diabetes: In our findings we found that while a majority of people preferred insulin as the mode of treatment, others forms of treatment such as natural remedies play a significant role as well (refer Figure 3). We have also expanded the other category in Figure 4.

- Preferred diabetic drugs: This pie chart shows us the generic drug that people prefer to combat diabetes. It is obvious that 1 drug ‘Novo Nordisk’ dominates the entire market (refer Figure 5)

- Diabetic Diet: The positive and negative diets associated with diabetes. Soda is most definitively negative while low carb and vegan is positive (refer Figure 6)

- Terms associated with diabetes: This is a list of the most popular terms associated with diabetes (refer Figure 7 and Table 1)

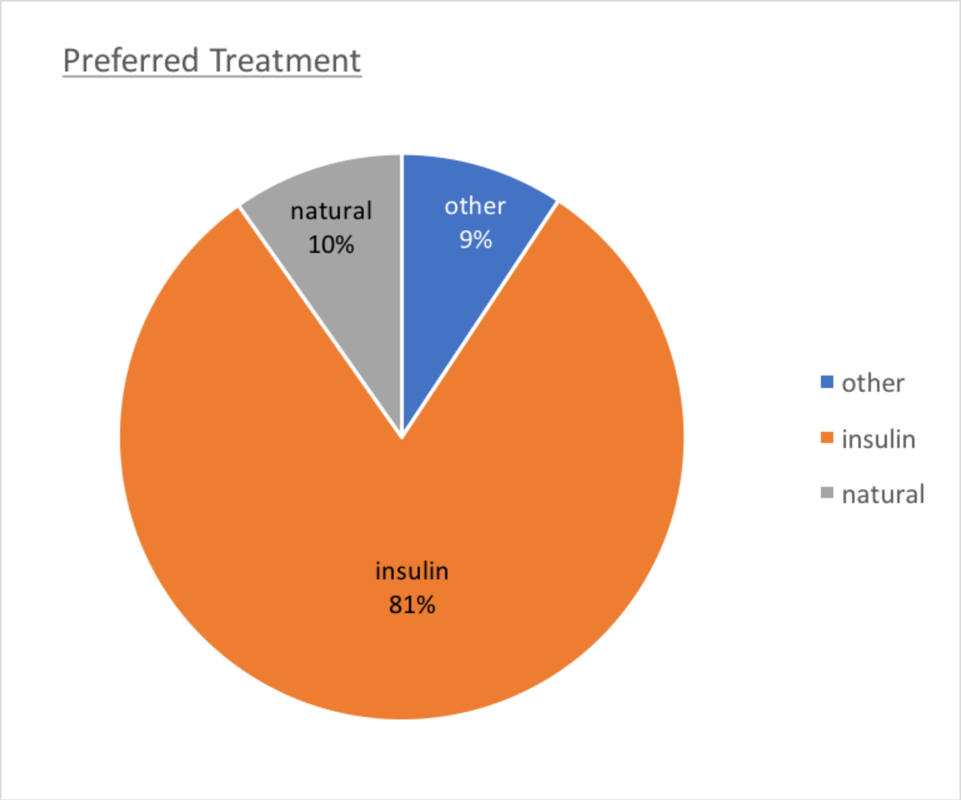


Figure 3: Insulin is the treatment of choice

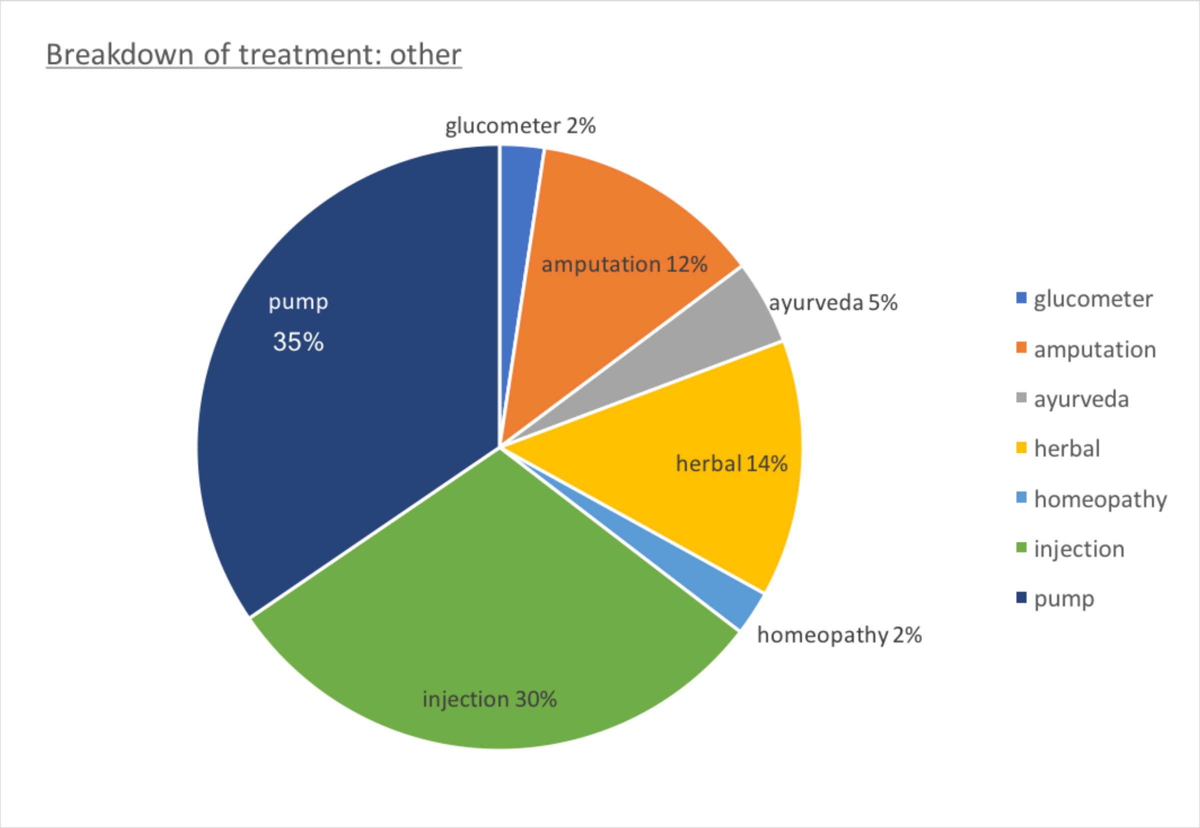


Figure 4: Breakdown of ‘other’ forms of treatment

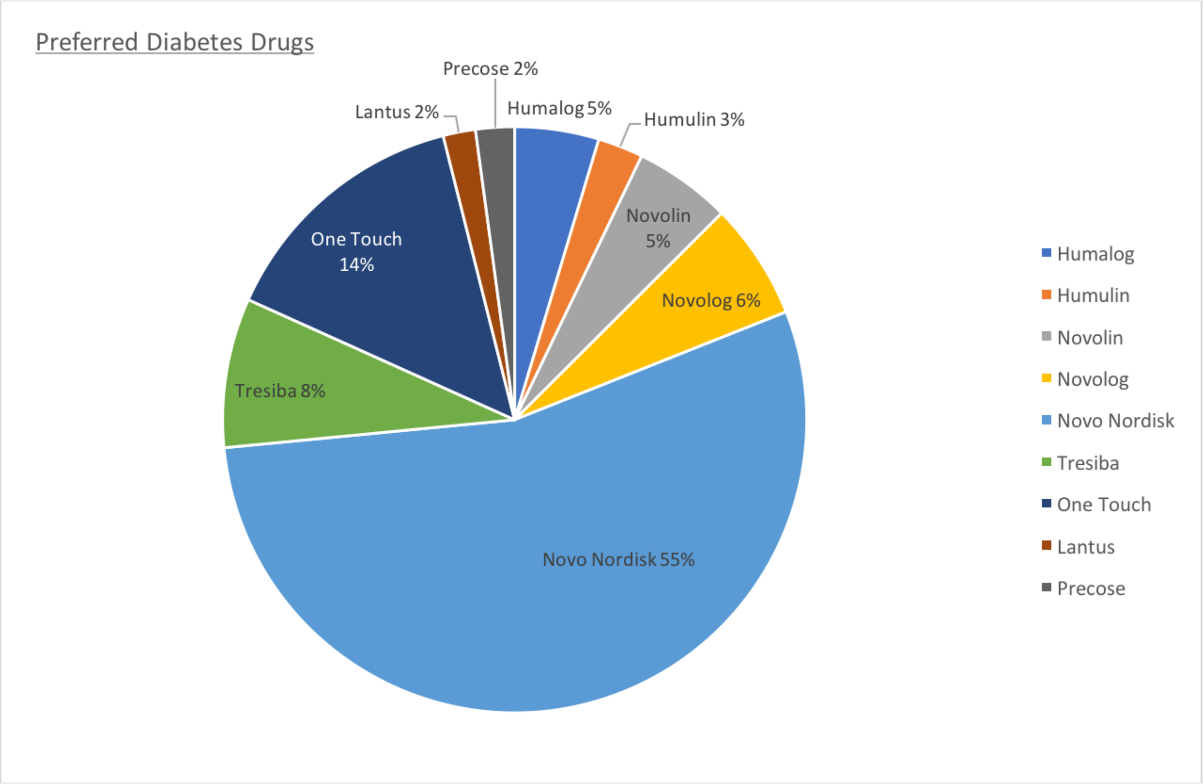


Figure 5: Novo Nordisk is the preferred drug for diabetes

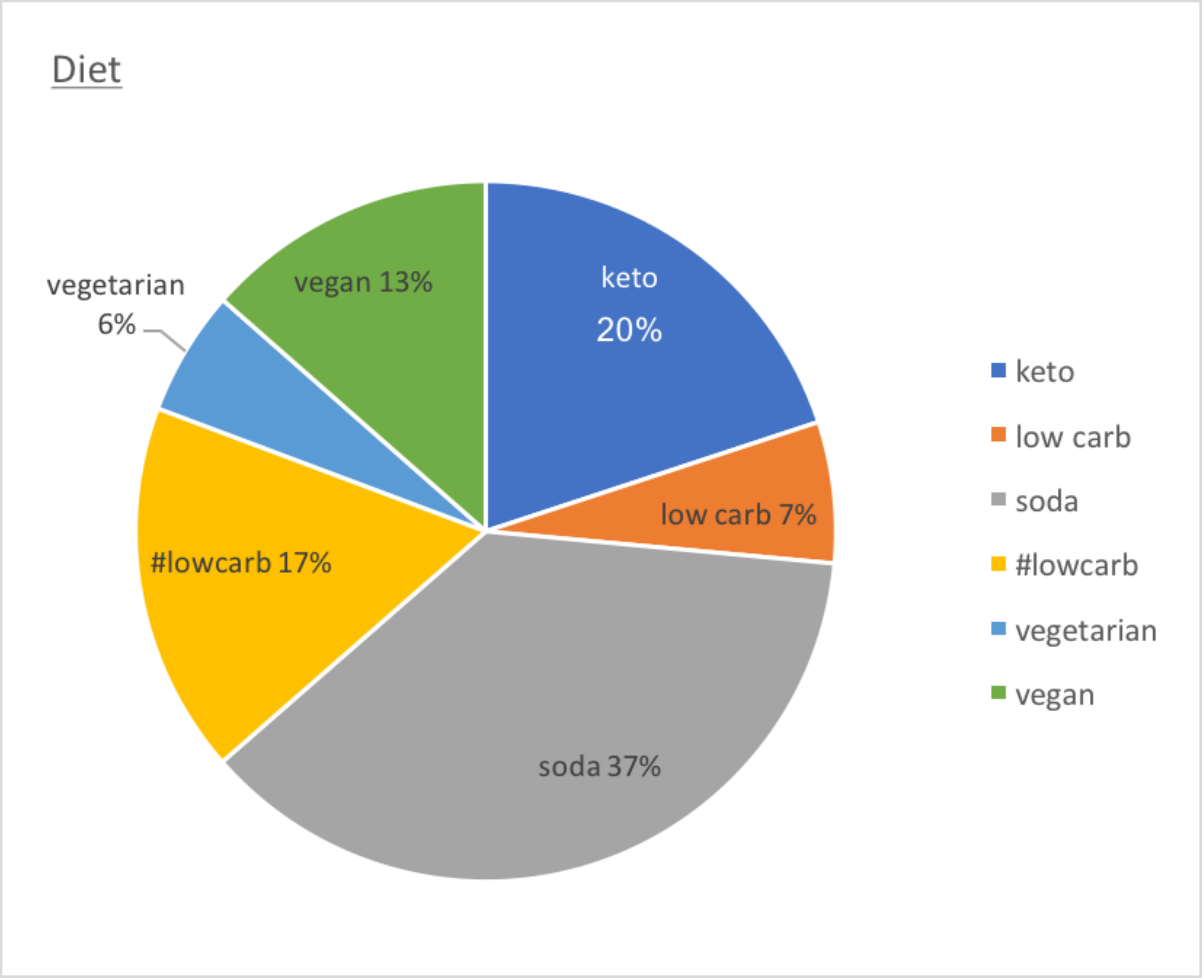


Figure 6: Diets associated with diabetes

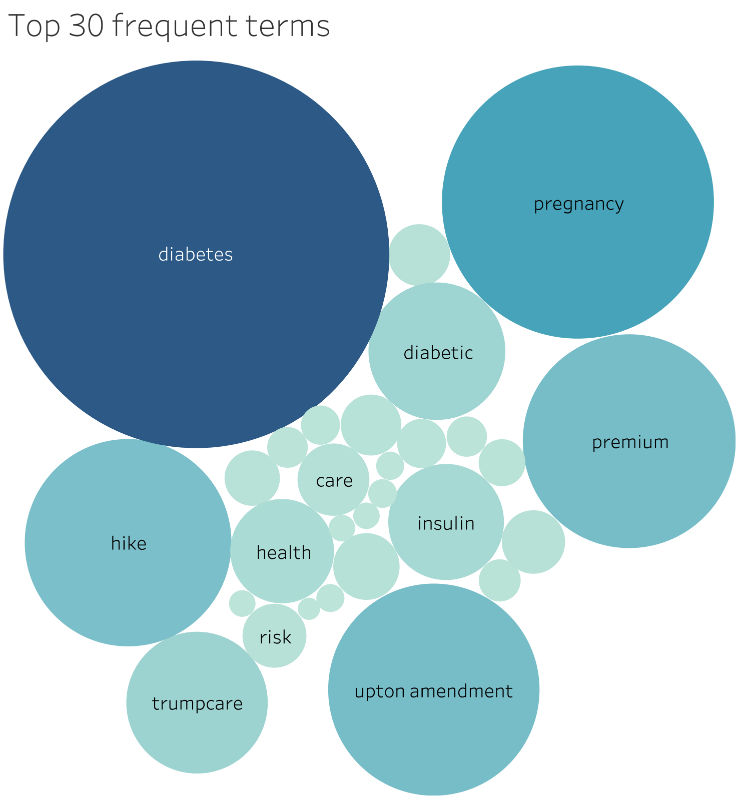


Figure 7: Frequent terms associated with diabetes

Term	Count	Term	Count
diabetes	84176	blood sugar	2076
pregnancy	41794	healthcare	1723
premium	25600	doctor	1354
upton amendment	25216	obesity	1239
hike	24098	type2	994
trumpcare	11273	natural	913
diabetic	10582	#preexistingconditions	860
insulin	7556	glucose	653
health	6085	food	465
care	2922	eat	446
insurance	2497	pancreas	440
risk	2299	soda	398
type1	2259	heal	398
patient	2149	gestation	394

Table 1: Frequent terms associated with diabetes

Geographical modelling

Figure 8 shows the normalised tweets state wise, Figure 9 shows the states that are most aware about diabetes

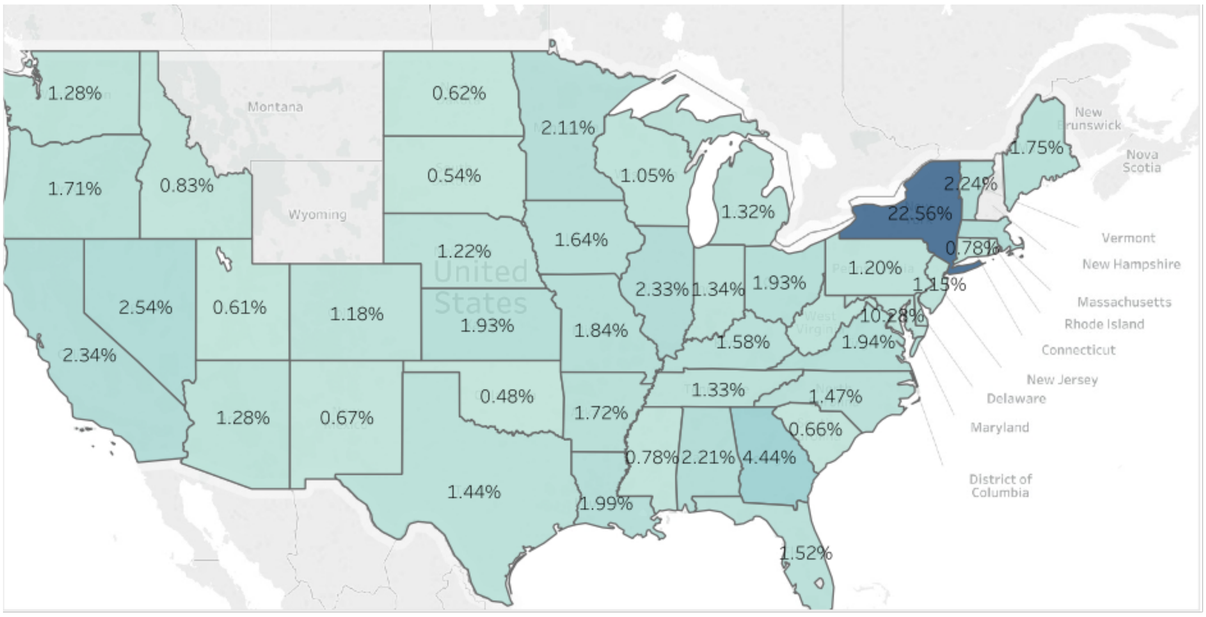


Figure 8: Normalised diabetes awareness state wise

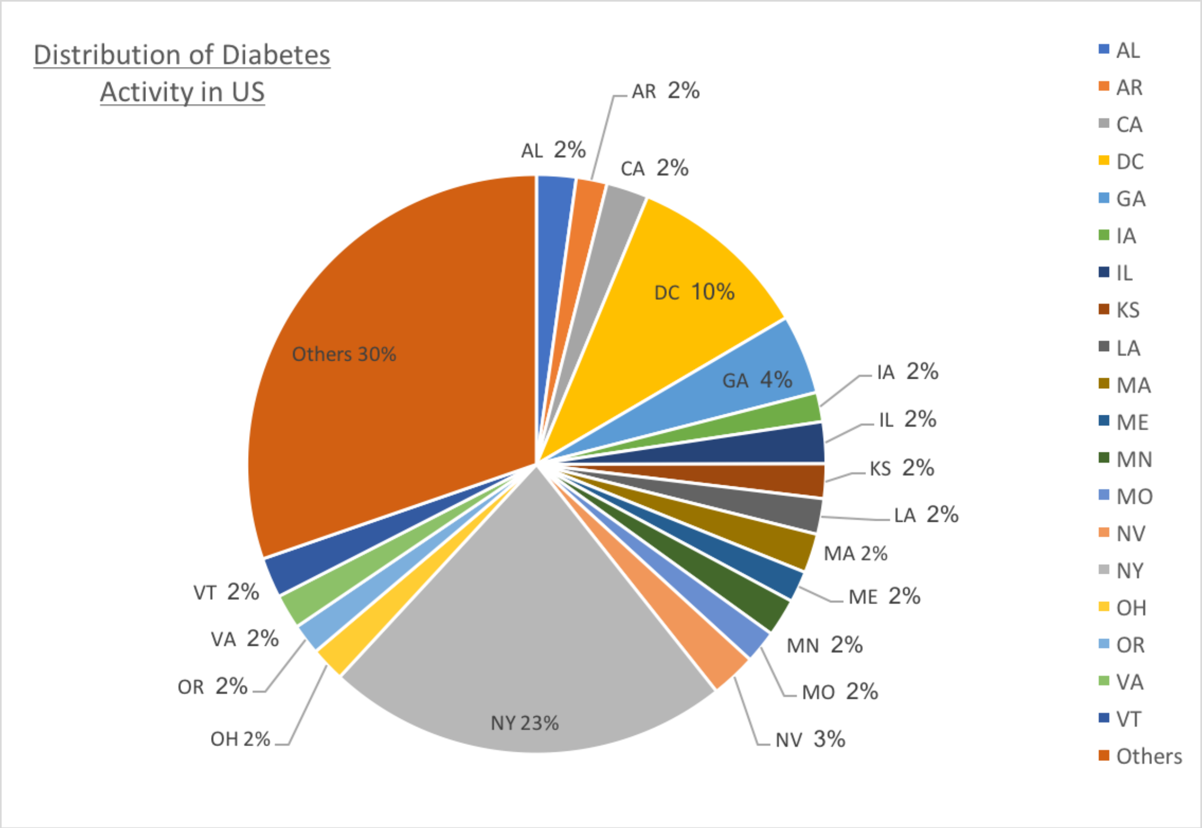


Figure 9: A handful of states cover over 40% of tweets

Sentiment Analysis

We ran the a small number of tweets through a sentiment analysis API (refer Table 2). The accuracy of the model was 66%

		Classified by API		
		Positive	Negetive	Neutral
Classified by us	Positive	35	4	7
	Negative	5	48	13
	Neutral	13	15	16

Table 2: Confusion Matrix for sentiment analysis

Temporal Analysis

Figure 10 shows us the number of tweets regarding diabetes over a period of 1 month. Figure 11 also shows us some influential trending topic on May 3 and 4 which explains the jump in diabetes tweets on that day. This can be attributed to a new bill that was passed in the senate. 'PreExisting' conditions were no longer covered in the bill or rather people with preexisting conditions could be charged a high 'premium'. This bill was known as 'upton amendment'. Figure 12 shows some popular tweets from that day. For example Steve Rattner's tweet has a whooping 30000 retweets

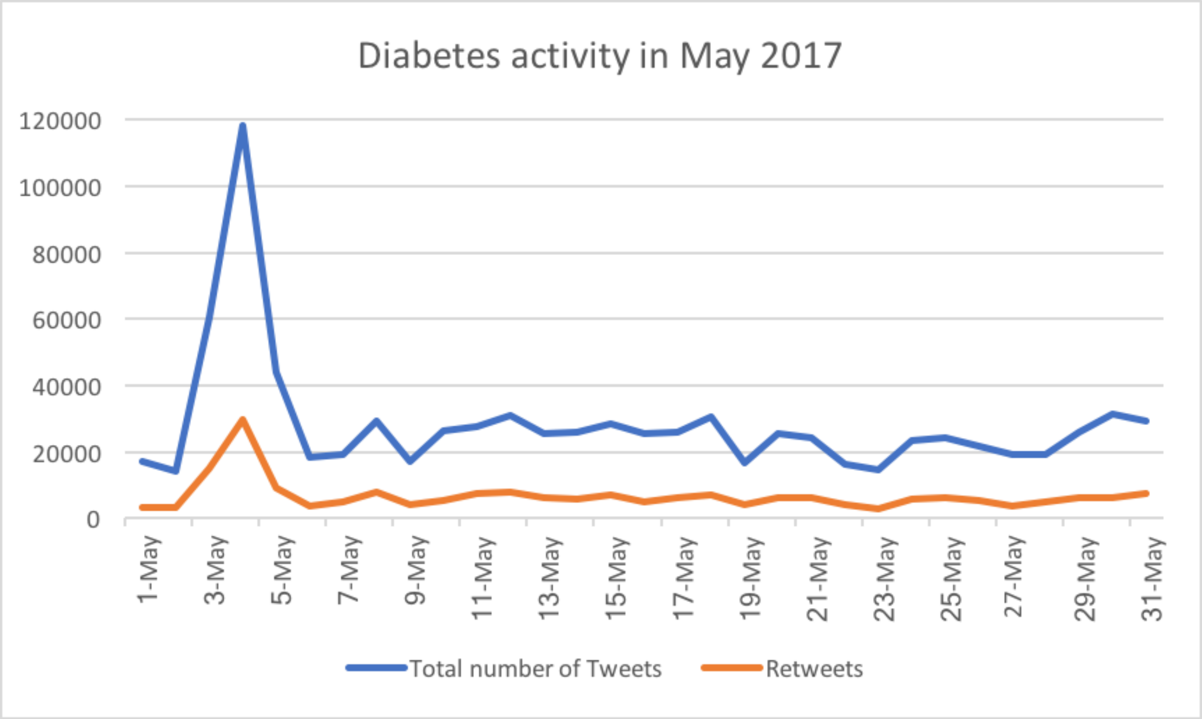


Figure 10: Distributions of tweets over a month

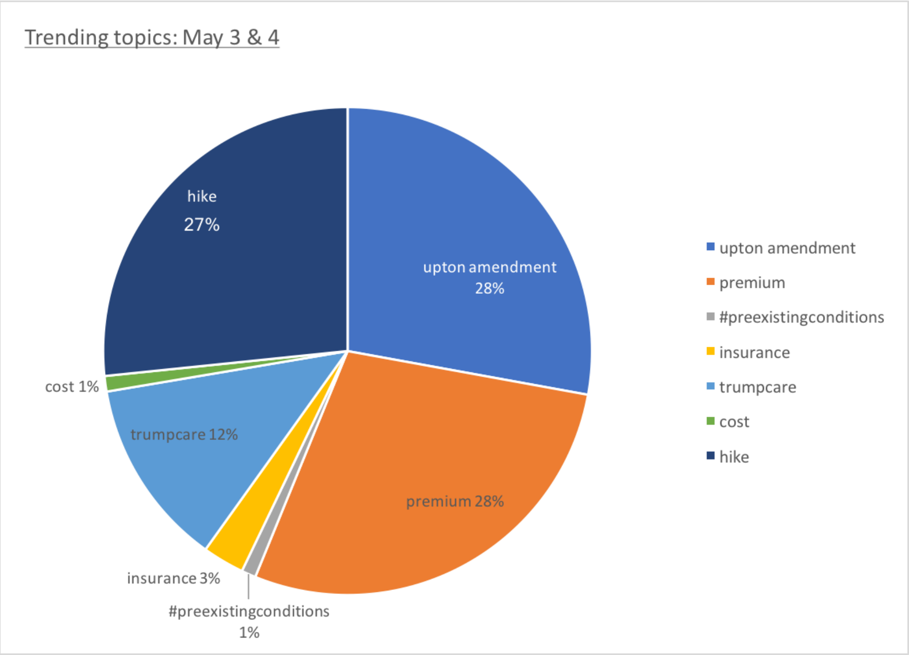
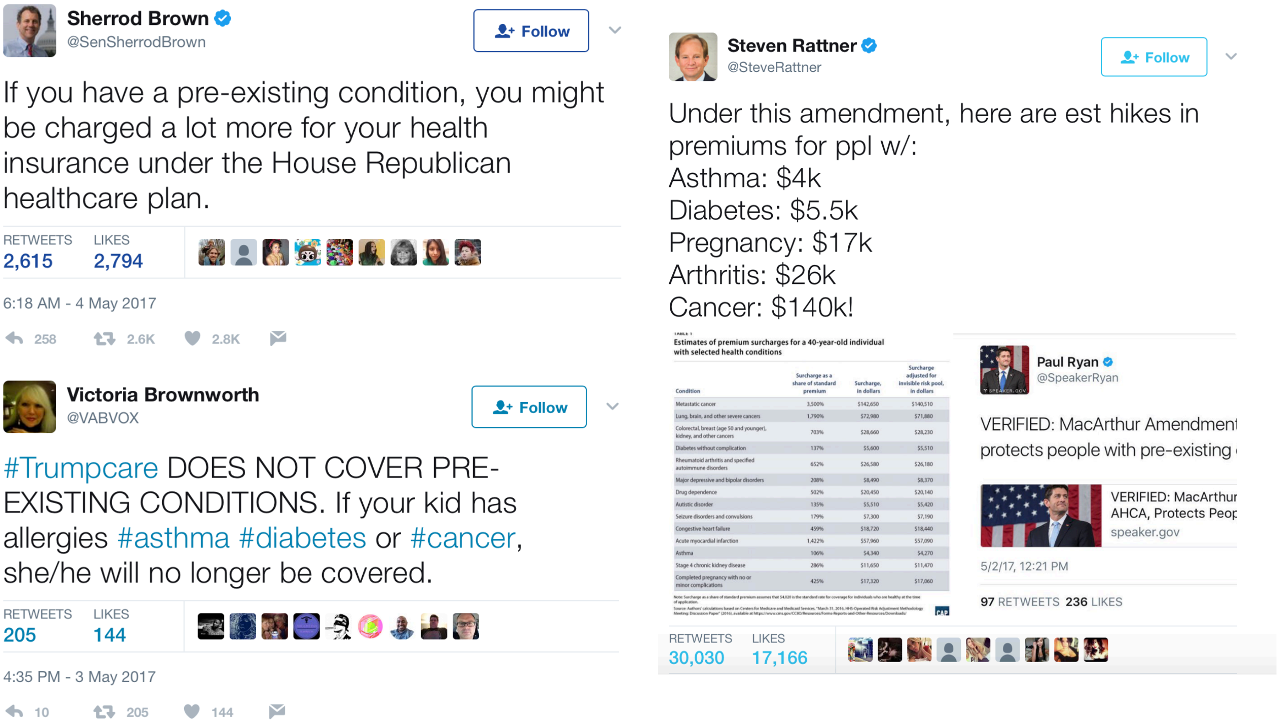


Figure 11: Trending topic on May 3rd and 4th and a spike in diabetes tweets



Sherrod Brown @SenSherrodBrown Follow

If you have a pre-existing condition, you might be charged a lot more for your health insurance under the House Republican healthcare plan.

RETWEETS 2,615 LIKES 2,794

6:18 AM - 4 May 2017

258 2.6K 2.8K

Victoria Brownworth @VABVOX Follow

#Trumpcare DOES NOT COVER PRE-EXISTING CONDITIONS. If your kid has allergies #asthma #diabetes or #cancer, she/he will no longer be covered.

RETWEETS 205 LIKES 144

4:35 PM - 3 May 2017

10 205 144

Steven Rattner @SteveRattner Follow

Under this amendment, here are est hikes in premiums for ppl w/:
Asthma: $4k
Diabetes: $5.5k
Pregnancy: $17k
Arthritis: $26k
Cancer: $140k!

TABLE 1
Estimates of premium surcharges for a 40-year-old individual with selected health conditions

Condition	Surcharge as a share of standard premium	Surcharge, in dollars	Surcharge adjusted for invisible risk pool, in dollars
Metastatic cancer	3,500%	$142,650	$140,510
Lung, brain, and other severe cancers	1,790%	$72,980	$71,880
Colorectal, breast (age 50 and younger), kidney, and other cancers	703%	$28,660	$28,230
Diabetes without complication	137%	$5,600	$5,510
Rheumatoid arthritis and specified autoimmune disorders	652%	$26,580	$26,180
Major depressive and bipolar disorders	208%	$8,490	$8,370
Drug dependence	502%	$20,450	$20,140
Autistic disorder	135%	$5,510	$5,420
Seizure disorders and convulsions	179%	$7,300	$7,190
Congestive heart failure	459%	$18,720	$18,440
Acute myocardial infarction	1,422%	$57,960	$57,090
Asthma	106%	$4,340	$4,270
Stage 4 chronic kidney disease	286%	$11,650	$11,470
Completed pregnancy with no or minor complications	425%	$17,320	$17,060

Note: Surcharge as a share of standard premium assumes that $4,020 is the standard rate for coverage for individuals who are healthy at the time of application.

Source: Authors' calculations based on Centers for Medicare and Medicaid Services, "March 31, 2016, HHS-Operated Risk Adjustment Methodology Meeting: Discussion Paper" (2016), available at https://www.cms.gov/CCIIO/Resources/Forms-Reports-and-Other-Resources/Downloads/

CAP

Paul Ryan @SpeakerRyan

VERIFIED: MacArthur Amendment protects people with pre-existing

VERIFIED: MacArthur AHCA, Protects Peop
speaker.gov

5/2/17, 12:21 PM

97 RETWEETS 236 LIKES

RETWEETS 30,030 LIKES 17,166

Figure 12: Some Popular tweets from May 3rd and 4th

GITHUB LINK TO CODE:

https://github.com/Jlakshmi235/smmSpr2017

WEBSITE:

BAG OF WORDS processing website. (from a text file)

Please download the repo, README file contains instructions on how to run it on local server

https://github.com/pencilsharpernerman/Word-Frequency-website

Conclusion

We have designed and implemented a system that tracks real-time diabetes-related activity on Twitter. Using the Twitter streaming API, we have collected around 850,000 tweets between May 1 and May 31. the quality of tweets collected were tuned by varying the keywords used in data collection. We also collected data from specific users (organization) whose concentration centers around diabetes.

The tweets were then preprocessed and only the text, timestamps and location data were retained from the raw tweets collected. Using this information, we have created geographical, textual and temporal models and analyzed the sentiments of the tweets.

All the results obtained are visualized as graphs.

This project is scalable and can be extended to track activity regarding other diseases.