Supplementary Online Materials

S1. List of keywords used with the Twitter API to collect diabetes-related tweets

English keywords		
glucose	insulin	blood glucose
#glucose	#insulin	#bloodglucose
insulin pump	diabetes	t1d
#insulinpump	#diabetes	#t1d
type 1	t2d	type 2
#type1	#t2d	#type2
#bloodsugar	#dsma	#type2diabetes
#doc	#bgnow	#wearenotwaiting
#insulin4all	dblog	diyps
hba1c	#dblog	#diyps
#hba1c	#cgm	#freestylelibre
diabetic	#gbdoc	freestyle libre
#diabetic	#gdm	finger prick
gestational diabetes	continuous glucose monitoring	#fingerprick
#gestational	#continuousglucosemonitoring	#changingdiabetes
#thisisdiabetes	#lifewithdiabetes	#diabetesadvocate
#stopdiabetes	#diabadass	#diabetesawareness
#diabeticproblems	#justdiabeticthings	#t1dlookslikeme
#diaversary	#diabetestest	#t2dlookslikeme
pwd	#duckfiabetes	#GBDoc
#pwd	#kissmyassdiabetes	

List of english keywords for the tweet extraction engine

S2. Details on data processing

Data representation

Natural language processing (NLP) methods were applied to represent tweets. We trained a FastText algorithm,[1] an extension of the Word2Vec model[2,3] using subword information, to obtain the vector representation of words, as it has been shown to be more efficient to extract meaningful semantic relationships when compared to generic pre-trained word embeddings.[4] Each tweet was then modelled as an average of its word vector representations. Cosine similarity, a widely used metric in text analysis, was used as distance measurement to decide if two tweets were similar or not.[5,6]

Data preprocessing

To reduce noise and bias in our dataset we applied several preprocessing steps. The first step removed all *retweets* and *duplicates* (tweets that are very similar in their syntax), which are often posted by chatbots, to obtain a database with unique tweets. To identify *duplicates,* cosine similarity was applied between the tweets of the same user. If the cosine similarity between two tweets was higher than 0.98, tweets were considered duplicates and only one was kept. This step led to a subset of 3.3 million unique tweets on which several additional preprocessing steps were applied (see Supplement S3 for the entire pipeline of the workflow).

Personal content classifier / Jokes classifier

Beguerisse-Diaz et al. have shown that diabetes-related tweets can be grouped into several clusters such as health information, news, social interaction and commercial.[7] In this study we aim to analyze only tweets with personal content, where feelings and emotions could be shared by people with or talking about diabetes. For this reason we further excluded *institutional* tweets (health information, news, commercial tweets) using a supervised machine learning approach. In this work, Support Vector Machine Classifiers (SVM)[8,9] have been developed to classify a tweet either as *institutional* or *personal*, in contrast to Johnsen et al. who identified personal content based on personal pronouns like 'I', 'me', 'us', etc.[10] We trained a first SVM to detect personal users and then a second SVM to detect tweets with personal content of personal users, as a personal user could also post institutional content such as advertisements. A total of 2275 and 1884 randomly chosen tweets were manually labelled by two authors (AH, GF) to train the first and second classifiers, respectively. Their performances have been assessed by cross validation.[11] The overall accuracy for detecting personal users was 0.89 and personal content of personal users 0.92.

Jokes around diabetes are common on Twitter and had to be excluded. Therefore another SVM was trained on 998 manually labelled tweets, by AH, GF. This classifier was optimised on its recall, to capture as many jokes as possible and reached a recall of 0.84.

Gender and Type of diabetes classifier

We also trained classifiers to predict gender (male, female, unknown) and type of diabetes (type 1, type 2, unknown) from information available in the dataset. Two authors (AA, GF) manually labelled 1,897 random tweets, based on the tweet's text, user description and the user name for the gender classifier and based on the tweet's text and user description for the type of diabetes classifier. The word vector representations of the first

name of the user name are a strong indicator for the prediction of gender. For this reason we added 3,614 further tweets with first names matched to the most popular baby names of the US Social Security Administration[12] to our training data to increase performance of our gender classifier. A SVM was trained and reached an accuracy of 86% (Precision to predict men: 92%; Precision to predict women: 93%) for the gender classifier and an accuracy of 74% (Precision to predict type 1 diabetes: 74%; Precision to predict type 2 diabetes: 72%).

Geolocation

The present study aimed at analyzing tweets from the USA. However, identifying the home locations of users on Twitter is a challenging task given the low number of posts with precise location information (geotags) and the need to parse user-defined location information using a gazetteer. According to the Twitter Developer page, only 1-2% of tweets are geotagged.[13] To predict the most refined geolocation possible, we developed a geolocation engine taking the geotag information from the place name field if existing, otherwise the user-defined text from the location field in Twitter user profiles, similar to the approach of Shah et al.;[14] additionally, if neither geotag nor the user location field were available we tried to extract the location information from the user's profile description. These descriptions (for instance *'Living with Type 1 diabetes for 32+ years, wife and three kids in Baltimore, MD USA.')* can provide valuable location information. The geolocation engine used the open-source software Apache Lucene[15] enabling full text indexing and searching capability in combination with Apache Spark[16] to process the large amount of data in an efficient way. The location information of tweets was matched with a freely available dictionary from the GeoNames geographical database,[17] containing over 25 million geographical names worldwide. To fine-tune the algorithm and avoid wrong classifications of user descriptions, an additional binary SVM was trained on 10.000 samples to detect if a word is a location or not (Accuracy: 91%). The binary output is then transformed into probabilities using an improved version of Platt's scaling;[18] probabilities are given as weights to the geolocation engine to improve predictions of the locations.

We found that only 6% (40,931/731,323) of the filtered tweets contained geo-coordinates shared from Twitter. After using our geolocation engine, we could infer worldwide geo-coordinates for 63% (463,623/731,323) of the tweets of which 31% (226,345/731,323) were found to be in the US and we identified 23% (167,743/731,323) in the US that also had information about a city.

The performance of our geolocation engine to infer the city information from the tweet based on the user location field was evaluated on 236 manual labelled samples, by AH and FO, with a precision of 85% , recall of 96%, F1-Score of 90%; and the improved version of the geolocation engine using the location classifier for the user description field was evaluated on 214 manual labelled samples, by AH and FO, and yielded a precision of 81%, recall of 68% and F1-score of 74%.

S3. Primary, secondary and tertiary emotions, as defined by Parrott et al.[19]

Primary emotion	Secondary emotion	Tertiary emotion
Joy	Cheerfulness	Amusement · Bliss · Gaiety · Glee · Jolliness · Joviality · Joy · Delight · Enjoyment · Gladness · Happiness · Jubilation · Elation · Satisfaction · Ecstasy · Euphoria
	Zest	Enthusiasm · Zeal · Excitement · Thrill · Exhilaration
	Contentment	Pleasure
	Pride	Triumph
	Optimism	Eagerness · Hope
	Enthrallment	Enthrallment · Rapture
	Relief	Relief
Love	Affection	Adoration · Fondness · Liking · Attraction · Caring · Tenderness · Compassion · Sentimentality
	Lust/Sexual desire	Desire · Passion · Infatuation
	Longing	Longing
Surprise	Surprise	Amazement · Astonishment
Sadness	Suffering	Agony · Anguish · Hurt
	Sadness	Depression · Despair · Gloom · Glumness · Unhappiness · Grief · Sorrow · Woe · Misery · Melancholy
	Disappointment	Dismay · Displeasure
	Shame	Guilt · Regret · Remorse
	Neglect	Alienation · Defeatism · Dejection · Embarrassment · Homesickness · Humiliation · Insecurity · Insult · Isolation · Loneliness · Rejection
	Sympathy	Pity · Mono no aware · Sympathy
Anger	Irritability	Aggravation · Agitation · Annoyance · Grouchy · Grumpy · Crosspatch
	Exasperation	Frustration
	Rage	Anger · Outrage · Fury · Wrath · Hostility · Ferocity · Bitterness · Hatred · Scorn · Spite · Vengefulness · Dislike · Resentment
	Disgust	Revulsion · Contempt · Loathing
	Envy	Jealousy
	Torment	Torment
Fear	Horror	Alarm · Shock · Fear · Fright · Horror · Terror · Panic · Hysteria · Mortification
	Nervousness	Anxiety · Suspense · Uneasiness · Apprehension (fear) · Worry · Distress · Dread

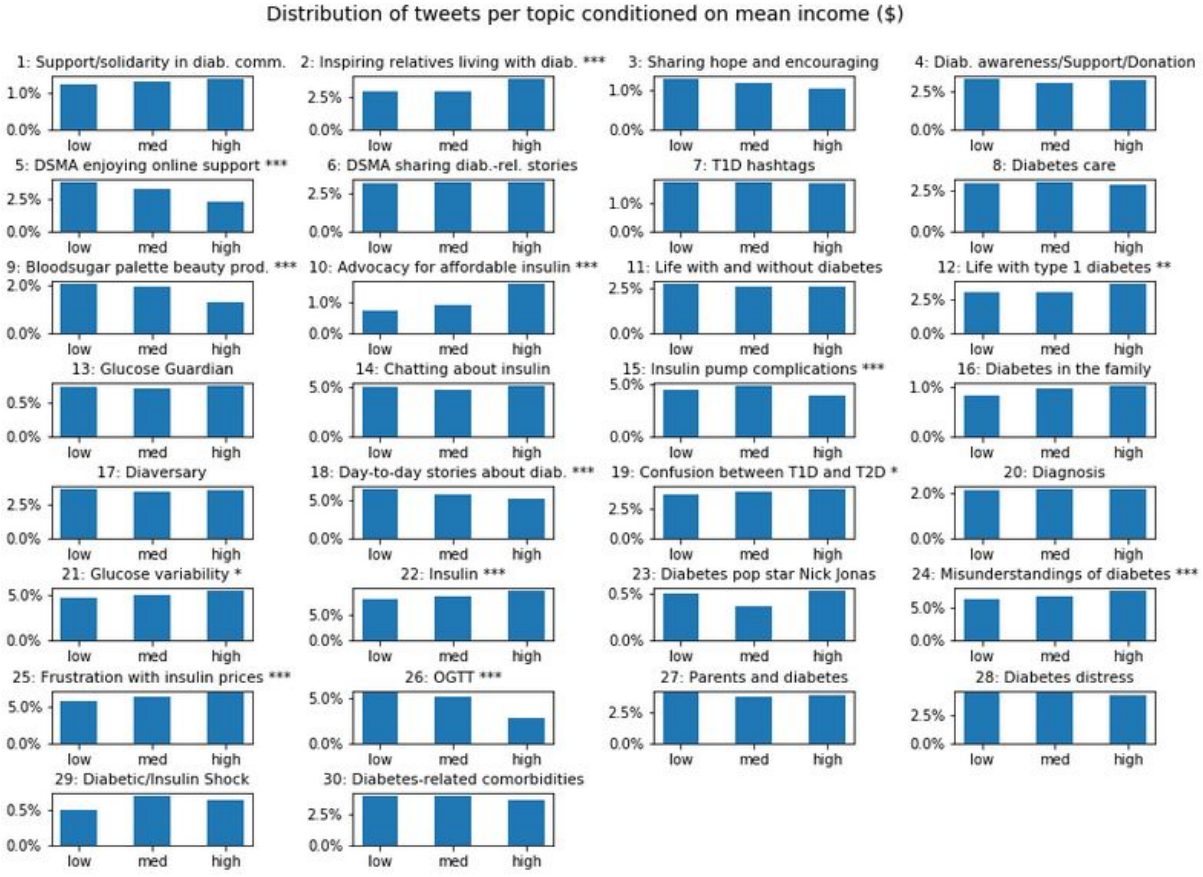
The colors in the table correspond to the distribution of primary emotions in Table 2.
For further details on the emotional words and emojis/emoticons used, please refer to https://github.com/WDDS/Tweet-Diabetes-Classification/tree/master/preprocess

S4. Top emotional keywords and emojis/emoticons by topics of interest of people living with diabetes

No.	Topic label	top emotional words	top emojis / emoticon
1	Support / solidarity in diabetes community	happy, love, hope, good, awesome, glad, celebration, sunshine	💙, ❤️, :), 😂, 😍, 😭, 🙃, 💜
2	Inspiring relatives living with diabetes	love, amazing, awesome, excited, good, happy, glad, hope	💙, ❤️, :), 😂, 😊, 😍, 💕, 💜
3	Sharing hope and encouraging	hope, good, feel, love, promise, happy, glad, understand	❤️, :(, 😔, 😊, 😂, ❤️, 😢, ❤️
4	Diabetes awareness / Support / Donation	love, good, hope, happy, feel, loved, attention, understand	💙, ❤️, :), 😂, 😃, 😭, 🙃, 😁
5	DSMA enjoying online support	good, glad, love, hope, feel, awesome, happy, excited	:), ❤️, 😂, dx, :(, 💙, ;), 😊
6	DSMA sharing diabetes-related stories	good, bad, feel, love, glad, happy, awesome, hope	:), 😂, ❤️, 😭, 🤗, 😔, 🙃, 😢
7	T1D hashtags	love, happy, good, awesome, excited, hope, amazing, hate	💙, ❤️, 😂, :), 😍, 😁, 💜, 🤣
8	Diabetes care	care, good, love, feel, hope, bad, glad, sick	😂, ❤️, 😁, 😭, 🙃, 😔, 💜, 😩
9	Bloodsugar palette (beauty products)	excited, love, happy, amazing, sad, loving, bad, feel	❤️, 😍, 😭, 💙, ♥, 💕, 💋, 😩
10	Advocacy for affordable insulin	good, love, hope, amazing, awesome, glad, attention, rally	💙, 💜, 😢, :), 😭, :3, 😱, ❤️
11	Life with and without diabetes	love, good, feel, hope, bad, understand, surprised, happy	😂, ❤️, 😭, dx, 😁, :), 🙃, 😢
12	Life with type 1 diabetes	love, good, feel, bad, happy, hate, excited, hope	💙, ❤️, 😂, 😭, 😁, :), 🙃, 😩
13	Glucose Guardian	love, good, happy, care, bad, poor, hope, feel	😂, 😭, 😩, ❤️, 🤣, 😔, 😁, 🙃
14	Chatting about insulin	love, good, feel, hope, bad, happy, funny, hate	😂, ❤️, :), 😭, :(, 🙃, dx, 😁
15	Insulin pump complications	love, hope, good, feel, bad, excited, happy, hate	😂, 🙃, 😭, 😁, :), 😩, ❤️, 😑
16	Diabetes in the family	good, love, hope, bad, feel, care, suffer, understand	❤️, 😂, 💙, 😊, 😁, 😞, 😭, 😢
17	Diaversary	good, happy, feel, love, bad, care, sick, hope	😂, ❤️, dx, :), 😁, 🙃, 😊, 💙
18	Day-to-day stories about diabetes	good, feel, bad, happy, sick, love, hate, excited	😂, 😭, 😁, 🙃, ❤️, 😩, :3, :)
19	Confusion between T1D and T2D	good, love, feel, bad, understand, hope, sick, hate	💙, 😭, ❤️, 😂, 😁, 🙃, :(, :)
20	Diagnosis	sick, love, good, feel, happy, care, bad, hope	dx, ❤️, 😭, 💙, 😂, 😢, :(, 😡
21	Insulin	good, feel, bad, understand, love, hope, hate, sick	😂, 😁, 😭, 🙃, ❤️, :), :(, 😑
22	Glucose variability	good, bad, feel, love, hate, sick, happy, understand	😂, 😭, 🙃, ❤️, 😁, 😩, :), ♥
23	Diabetes pop star Nick Jonas	cried, crying, love, cry, crush, feel, good, sad	😭, 😂, :(, 💜, 😔, :/, 🤣, 😱
24	Misunderstandings of diabetes	understand, love, bad, good, feel, hate, sick, sad	😂, 😕, 😭, ❤️, 😡, 🙃, 😩, 😢
25	Frustration with insulin prices	good, care, outrageous, sad, love, hope, bad, feel	😕, 😡, 😭, 😂, 🙃, 😢, :(, 😑
26	OGTT	bad, good, hate, feel, nervous, dreading, excited, hope	😭, 😩, 😁, 🤢, 🙃, 😂, 😔, 😷
27	Parents and diabetes	bad, good, sick, love, sad, poor, care, feel	😂, 😭, ❤️, 😢, 😔, :(, 😁, 😞
28	Diabetes Distress	feel, hate, sick, bad, sad, worried, funny, glad	😂, 😭, :(, 😁, 😢, :), ❤️, 😞
29	Diabetic / Insulin shock	shock, love, worried, good, bad, horrible, sweetness, trauma	😳, 😂, 😩, 🙃, 😔, 😢, 😊, 😱
30	Diabetes-related comorbidities	pain, depression, anxiety, bad, feel, suffer, good, damage	❤️, 🙃, 😭, 😂, 😁, :(, 😢, 😑

For each topic the most frequent emotional words and emojis/emoticons are listed

S5: Associations between the mean household income at the city level and the topics of interest of people living with diabetes



Plot of the frequency of tweets per category of mean household income at the city level (low, medium, high) for each topic. A p-value derived from a Chi2 test was calculated.
* p-value < 0.05; ** p-value < 0.01; *** p-value < 0.001

Supplementary - References

1. Bojanowski P, Grave E, Joulin A, et al. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 2017; 5: 135–146.

2. Tomas Mikolov, Kai Chen, Greg Corrado, et al. Efficient Estimation of Word Representations in Vector Space.

3. Tomas Mikolov, Ilya Sutskever, Kai Chen, et al. Distributed representations of words and phrases and the ir compositionality. *In Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing System s* 2013; 3111–3119.

4. Khatua A, Khatua A, Cambria E. A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks. *Information Processing & Management* 2019; 56: 247–257.

5. Tan P-N, Steinbach M, Karpatne A, et al. *Introduction to Data Mining*. Pearson, 2019.

6. Strehl, Alexander, Joydeep Ghosh, and Raymond Mooney. Impact of similarity measures on web-page clustering. 2000, pp. 58–64.

7. Beguerisse-Díaz M, McLennan AK, Garduño-Hernández G, et al. The 'who' and 'what' of #diabetes on Twitter. *Digit Health* 2017; 3: 2055207616688841.

8. Joachims T. Text categorization with Support Vector Machines: Learning with many relevant features. *Machine Learning: ECML-98* 1998; 137–142.

9. Sewalk KC, Tuli G, Hswen Y, et al. Using Twitter to Examine Web-Based Patient Experience Sentiments in the United States: Longitudinal Study. *J Med Internet Res* 2018; 20: e10043.

10. Johnsen J-AK, Eggesvik TB, Rørvik TH, et al. Differences in Emotional and Pain-Related Language in Tweets About Dentists and Medical Doctors: Text Analysis of Twitter Content. *JMIR Public Health and Surveillance* 2019; 5: e10432.

11. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. New York, NY: Springer New York, 2009.

12. US Social Security Administration Babynames, https://www.ssa.gov/oact/babynames/limits.html (accessed 12 August 2019).

13. Twitter Developer - Tweet geospatial metadata, https://developer.twitter.com/en/docs/tutorials/tweet-geo-metadata.

14. Shah Z, Martin P, Coiera E, et al. Modeling Spatiotemporal Factors Associated With Sentiment on Twitter: Synthesis and Suggestions for Improving the Identification of Localized Deviations. *J Med Internet Res* 2019; 21: e12881.

15. Apache Software Foundation. Apache Lucene, https://lucene.apache.org/ (accessed 20 August 2019).

16. Matei Zaharia, Apache Software Foundation, UC Berkeley AMPLab, Databricks. Apache Spark. *Apache Spark*, https://spark.apache.org/ (accessed 20 August 2019).

17. Geonames Database, https://www.geonames.org/ (accessed 19 July 2018).

18. Lin H-T, Lin C-J, Weng RC. A note on Platt's probabilistic outputs for support vector machines. *Machine Learning* 2007; 68: 267–276.

19. Gerrod Parrott W. *Emotions in Social Psychology: Essential Readings*. Psychology Press, 2001.