BMJ Open Diabetes Research & Care

Global diabetes burden: analysis of regional differences to improve diabetes care

Charline Bour,[1,2] Adrian Ahne,[3] Gloria Aguayo ,[1] Aurélie Fischer,[1] David Marcic,[4] Philippe Kayser,[4] Guy Fagherazzi [1]

To cite: Bour C, Ahne A, Aguayo G, *et al*. Global diabetes burden: analysis of regional differences to improve diabetes care. *BMJ Open Diab Res Care* 2022;**10**:e003040. doi:10.1136/bmjdrc-2022-003040

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjdrc-2022-003040).

Received 21 July 2022
Accepted 14 October 2022

Check for updates

[1]Department of Precision Health, Deep Digital Phenotyping Research Unit, Luxembourg Institute of Health, Strassen, Luxembourg
[2]Faculty of Science, Technology and Medicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg
[3]Center for Research in Epidemiology and Population Health (CESP), INSERM, Villejuif (Paris), Île-de-France, France
[4]Department of Precision Health, Data Integration and Analysis Unit, Luxembourg Institute of Health, Strassen, Luxembourg

Correspondence to
Dr Guy Fagherazzi;
Guy.Fagherazzi@lih.lu

ABSTRACT

Introduction The current evaluation processes of the burden of diabetes are incomplete and subject to bias. This study aimed to identify regional differences in the diabetes burden on a universal level from the perspective of people with diabetes.

Research design and methods We developed a worldwide online diabetes observatory based on 34 million diabetes-related tweets from 172 countries covering 41 languages, spanning from 2017 to 2021. After translating all tweets to English, we used machine learning algorithms to remove institutional tweets and jokes, geolocate users, identify topics of interest and quantify associated sentiments and emotions across the seven World Bank regions.

Results We identified four topics of interest for people with diabetes (PWD) in the Middle East and North Africa and another 18 topics in North America. Topics related to glycemic control and food are shared among six regions of the world. These topics were mainly associated with sadness (35% and 39% on average compared with levels of sadness in other topics). We also revealed several region-specific concerns (eg, insulin pricing in North America or the burden of daily diabetes management in Europe and Central Asia).

Conclusions The needs and concerns of PWD vary significantly worldwide, and the burden of diabetes is perceived differently. Our results will support better integration of these regional differences into diabetes programs to improve patient-centric diabetes research and care, focused on the most relevant concerns to enhance personalized medicine and self-management of PWD.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Twitter data can be a useful resource to monitor key concerns of people with diabetes, complementary to what can be achieved with questionnaires in clinical studies.

WHAT THIS STUDY ADDS

⇒ This study included a worldwide analysis of a dataset of 34 millions of tweets from 172 countries to detect the most important topics of interest of people with diabetes and to study their differences across the seven World Bank regions.

⇒ We have identified universal topics of concern. The concerns related to glycemic control and food are common to seven and six regions of the world, respectively.

⇒ Other topics were found to be more important in some specific regions, such as insulin pricing in North America or the burden of daily diabetes management in Europe and Central Asia.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Our results can support the development of tailored diabetes programs at the regional level to focus on the most important concerns and thus to enhance personalized medicine and self-management of people with diabetes.

INTRODUCTION

The term ‘burden of disease’ describes the overall consequences (loss of health, social aspects, costs to society, death) caused by diseases, injuries and risk factors worldwide and is often measured using quality-adjusted life years (QALYs) or disability-adjusted life years (DALYs).[1–3] However, QALYs and DALYs prevent us from understanding the drivers of the diabetes burden, such as the role of diabetes distress or the quality of care. Diabetes distress defines the emotional distress linked to living with diabetes and day-to-day management but also worrying about complications.[4] It has been shown that one in four people with type 1 diabetes and one in five people with type 2 diabetes have high levels of diabetes distress.[5] Emotional distress is associated with diabetes self-management and glycemic control issues.[6]

Conceiving patient-centered instruments helped measure the quality of care for PWD. Many of these have additional subscales ortheir evaluation aspects overlap.[6] These gaps in the assessment methods of the quality of care for PWD need to be identified. The most important factors must be prioritized and become objectives to address. As priorities for a person with diabetes in the USA may differ vastly between a PWD in Western

Europe, the Middle East or South Asia, determining the regional objectives is necessary to improve the lives of PWD. It is crucial to understand the regional differences in how the diabetes burden is perceived to integrate them into future diabetes programs. These could then address the most relevant local factors of diabetes burden.

One international source of data that captures the viewpoint of people with diabetes is Twitter. With more than 130 million users in 2019, it proved to be compatible with health research in various ways, but mainly to collect a considerable volume of data for public health surveillance, early event detection, outbreak prediction and analysis of a population's sentiments and emotions.[7–12] Sentiment analysis aims to recognize polarity in texts (positivity, negativity or neutrality), while emotion analysis determines the emotional state of an individual (anger, fear). Several diabetes communities have developed on Twitter, where users can share their experiences, ask for advice or chat. They can be found with relevant hashtags (#dsma: Diabetes Social Media Advocacy, #gbdoc: UK Diabetes Online Community). It is thus possible to access large quantities of diabetes-related data from individuals and communities of PWD on Twitter. Social media data enables a better understanding of the principal daily concerns and associated emotions related to diabetes, diabetes management, diabetes distress or diabetes burden.[13] More broadly, social media data may provide insights into how concerns differ between countries. As Twitter is embraced globally by numerous people and does not rely on predefined questions like evaluation scales, collecting and analyzing tweets can be considered an innovative and complementary way to understand PWD's feelings and concerns about their diabetes. Ahne *et al*[14] previously showed that such analysis could be efficient in identifying primary concerns in the USA.

Because precision health starts by contextualizing the needs of the patients, we have tested the hypothesis that it is feasible to use a reproducible approach to analyze online data to better understand the determinants of diabetes burden and to identify regional differences that will serve to design more patient-centered diabetes programs in the future.[10]

RESEARCH DESIGN AND METHODS

Data collection

Tweets are public by default and can be collected using the Twitter Application Programming Interface, which provides access to 1% of all Twitter data in real time based on keywords. To collect diabetes-related tweets, we defined a list of 272 diabetes-related keywords such as *diabetes, insulin* and *blood glucose* in 30 different languages (online supplemental appendix 1). Overall, the collection includes 34 million tweets published between May 2017 and April 2021. The data collected for this study only includes publicly posted tweets.

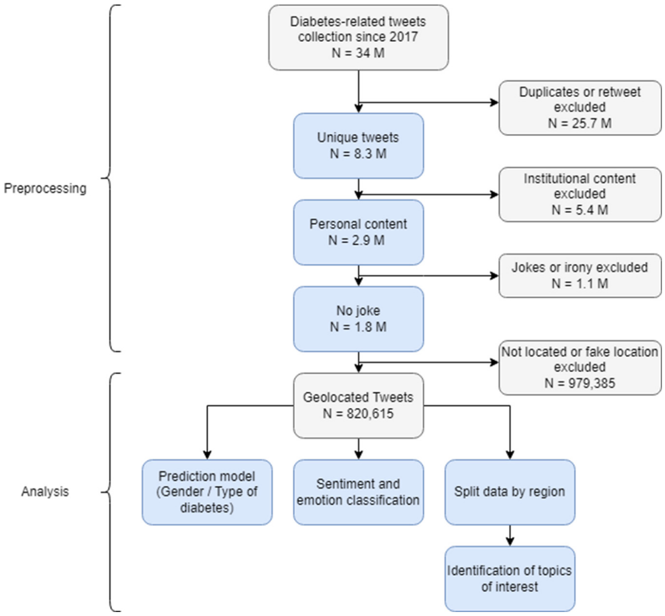
Preprocessing
Analysis
Diabetes-related tweets collection since 2017 N = 34 M
Duplicates or retweet excluded N = 25.7 M
Unique tweets N = 8.3 M
Institutional content excluded N = 5.4 M
Personal content N = 2.9 M
Jokes or irony excluded N = 1.1 M
No joke N = 1.8 M
Not located or fake location excluded N = 979,385
Geolocated Tweets N = 820,615
Prediction model (Gender / Type of diabetes)
Sentiment and emotion classification
Split data by region
Identification of topics of interest

Figure 1 Workflow showing the data preprocessing and analysis. Blue boxes correspond to steps where machine learning methods apply.

Preprocessing

The first step consisted of deleting duplicates and retweets to keep unique tweets and quote retweets (a retweet with an added comment). Second, non-English tweets were translated into English. Third, two classifiers were applied to keep only tweets with personal, non-joke or non-ironic content from users sharing diabetes-related information about themselves or relatives. The workflow can be seen in figure 1.

Geolocation

A tweet object provides meta-data, including information about the user account and location. The users provide their geographical area via an entry in their public profile. The precision in their description may vary. After applying the process described in online supplemental appendix 2, tweets were separated into the following seven regions: North America, East Asia and Pacific, Europe and Central Asia, Latin America and the Caribbean, the Middle East and North Africa, South Asia, and Sub-Saharan Africa. These regions comply with the 'World Bank Country and Lending Groups' classification from The World Bank Group.[15]

Sentiment analysis

We used Valence Aware Dictionary for Sentiment Reasoning to assess whether there was a positive or negative sentiment within a tweet.[16] The primary metric used for the sentiment analysis was the compound score (polarity), a unidimensional and normalized measure of sentiment between −1 and +1.

Topic extraction

We applied a k-means algorithm to the tweets in each region and gave each cluster a label according to the 20 closest tweets to the topic center and the most frequent words (top words) in the cluster.[17 18]

Emotion analysis

To determine the predominant emotion in each tweet, a classifier was developed based on texts focusing on four emotions: fear, anger, joy and sadness.[19] We applied this classifier to all tweets to predict the probability of a tweet belonging to each of the four emotions.

Every algorithm used for this study is available on Github: https://github.com/Chbour/Global_diabetes_burden. More details about the methodology can be found in online supplemental appendix 2.

Role of the funding source

The content of this publication is solely the author's responsibility and does not necessarily represent the official views of the funders.

RESULTS

Spatial distribution of diabetes-related tweets

After preprocessing, we included 820 615 geolocated tweets in this study. Tweets were distributed as follows: 568 020 from North America (n=69.2%, three countries included), 176 124 from Europe and Central Asia (n=21.5%, 49 countries included), 31 426 from East Asia and Pacific (n=3.8%, 27 countries included), 20 465 from Sub-Saharan Africa (n=2.5%, 36 countries included), 15 935 from South Asia (n=1.9%, eight countries included), 4554 from Latin America and the Caribbean (n=0.6%, 29 countries included) and 4091 from the Middle East and North Africa (n=0.5%, 20 countries included). Figure 2 displays the distribution of tweets in each region.

Topics of interest

Among all tweets, 269 323 (32.8%) were predicted as posted by men, 311 343 (37.9%) by women, and 239 949 (29.2%) from unknown sex; 254 564 (31%) were from people with type 1 diabetes, 94 948 (11.6%) from type 2 diabetes and 471 203 (57.4%) from people where diabetes type was impossible to predict. Females were over-represented in East Asia and Pacific, Europe and Central Asia, Middle East and North Africa, and North America. Men were over-represented in Latin America and the Caribbean, and South Asia. In all regions, tweets identified as type 1 diabetes-related were predominant.

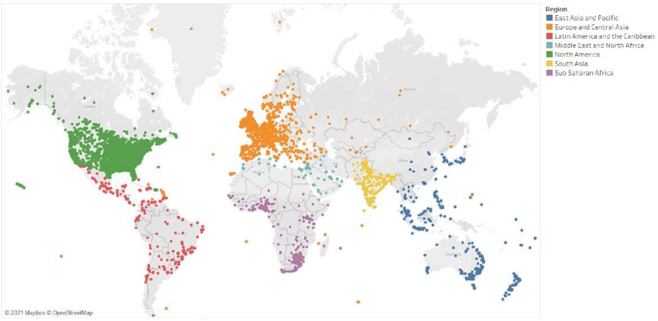
Region
🟦 East Asia and Pacific
🟧 Europe and Central Asia
🟥 Latin America and the Caribbean
🟦 Middle East and North Africa
🟩 North America
🟨 South Asia
🟪 Sub Saharan Africa
© 2021 Mapbox © OpenStreetMap

Figure 2 Map showing the distribution of diabetes-related tweets according to the region (n=820 615).

We identified four topics of interest for the people with diabetes from the Middle East and North Africa, 6 for South Asia, 8 of interest for East Asia and Pacific, 7 for Latin America and the Caribbean, 10 for Europe and Central Asia, 14 for Sub-Saharan Africa and 18 for North America. They are further described below for each region and in online supplemental appendix 3. 'Glycemic Control' was a topic found in all regions. Six out of seven showed a common interest such as 'Family and relatives' and 'Food', whereas 'Insulin' matched for five regions. Four regions had common topics related to 'Comorbidities'. The significance of comparing percentages among emotions in topics in each region was determined using a Student's t-test. All p values shown are two tailed.

Overall, South Asia had the most positive diabetes-related tweets and was associated with a higher polarity score, while Latin America and the Caribbean had the most negative ones and were associated with a lower score (table 1). On the 820 615 included tweets, 356 683 were identified as positive (n=43.5%), and 308 811 were identified as negative (n=37.6%). South Asia and Europe and Central Asia had a higher proportion of positive tweets (47.6% and 46%, respectively). Latin America and the Caribbean, and North America had a higher proportion of negative tweets (38.2% and 38.5%, respectively). As shown in table 1, the South Asia region was associated with a higher average polarity score, while Latin America and the Caribbean were associated with a lower score. The averaged sentiment scores were slightly positive and between 0.01887 (Latin America and the Caribbean) and 0.10376 (South Asia). Most regions had a positive score (greater than 0.05). In contrast, Latin America and the Caribbean, and North America had a neutral score (between −0.05 and 0.05) as these regions had a higher proportion of tweets with negative sentiment scores.

East Asia and Pacific

On average, topics referring to users sharing support and advice such as 'Type 1 diabetes communities' (48% compared with 31.4% on average in all other topics, p<0.001) and 'Glycemic control' (39% compared with 31.6% on average in all other topics) were associated with higher rates of joy (p<0.001) but also with higher rates of fear (respectively 16.9% and 14.6% compared with 12.1% and 12.3% on average in all other topics, p<0.001) due to frequent fears about the future. 'Insulin affordability' was associated with a higher rate of anger (28% compared with 16.6% on average in all other topics, p<0.001) because of users reacting to the huge insulin pricing gap between the USA and East Asia and Pacific.[20] 'Diabetes-related complications and family history' was associated with a higher probability of sadness (45.8% compared with 38.1% on average in all other topics, p<0.001).

Table 1 Average sentiment score and distribution sentiment scores

Region	Mean sentiment score	Number of tweets with negative, neutral and positive sentiment scores
East Asia and Pacific	0.06961	Negative: 11 189 (n=35.6%). Neutral: 5860 (n=18.6%). Positive: 14 377 (n=45.7%).
Europe and Central Asia	0.07209	Negative: 63 205 (n=35.9%). Neutral: 31 902 (n=18.1%). Positive: 81 017 (n=46%).
Latin America and the Caribbean	0.01887	Negative: 1741 (n=38.2%). Neutral: 938 (n=20.6%). Positive: 1875 (n=41.2%).
Middle East and North Africa	0.07022	Negative: 1431 (n=35%). Neutral: 804 (n=19.6%). Positive: 1856 (n=45.4%).
North America	0.02792	Negative: 218 717 (n=38.5%). Neutral: 108 246 (n=19.1%). Positive: 241 057 (n=42.4%).
South Asia	0.10376	Negative: 5198 (n=32.6%). Neutral: 3149 (n=19.8%). Positive: 7588 (n=47.6%).
Sub-Saharan Africa	0.05002	Negative: 7330 (n=35.8%). Neutral: 4182 (n=20.4%). Positive: 8953 (n=43.7%).

A sentiment score is considered negative, when lower or equal to −0.05, positive when greater than or equal to 0.05 and considered neutral when strictly between −0.05 and 0.05.[16]

Europe and Central Asia

The two topics dealing with insulin ('Insulin access' and 'Insulin and insulin supplies') were associated with a higher probability of anger (respectively 28.6% and 26.2% compared with 15.87% and 16.3% on average in all other topics, p<0.001). Topics discussing relatives' life with diabetes and complications ('Diabetes-related complications and family history' and 'Life changes since diagnosis') were associated with sadness (respectively 45.6% and 43% compared with 35.6% and 35.9% on average in all other topics, p<0.001). Topics 'Daily management of diabetes' and 'Type 1 diabetes communities' were mostly associated with joy (respectively 43.7% and 50.4% compared with 32% and 33% on average in all other topics, p<0.001).

Latin America and the Caribbean

Similar to Europe and Central Asia, the topic 'Insulin issues' was associated with a higher probability of anger (28.7% compared with 15.6% on average in all other topics, p<0.001). Topics in which users shared love and advice ('Love and support' and 'Glycemic control') were associated with a higher probability of joy (respectively 46.02% and 37.9% compared with 29% and 29.1% on average in all other topics, p<0.001). Finally, topics dealing with relatives' health complications and life with diabetes ('Complications and comorbidities' and 'Experiences from relatives living with diabetes') were associated with a higher probability of sadness (respectively 47.9% and 47.8% compared with 42.7% and 40.4% on average in all other topics, p<0.001).

Middle East and North Africa

Topic 'Insulin and insulin supplies' was associated with a higher probability of anger (28.8% compared with 15.6% on average in all other topics, p<0.001). In this topic, users were reacting to the difficulty of insulin and insulin supplies self-management. However, sadness was the main identified emotion in all topics (39% on average).

North America

The five topics dealing with insulin pricing and affordability ('Inability to afford insulin', 'Consequences of insulin unaffordability', 'Insulin prices increase', 'Insulin pricing including insurance' and 'Costs implied by diabetes management') were associated with a higher probability of anger (between 20.1% and 32.7% compared with 17.9% to 18.8% on average in all other topics, p<0.001). Most topics were associated with a higher probability of sadness (41% on average) except 'Type 1 diabetes communities', 'Glucose tests' and 'Sharing daily life' were associated with a higher probability of joy (respectively 46.1%, 48.7%, and 42.5% compared with 29.8%, 29.6% and 28.9% on average in all other topics, p<0.001).

South Asia

The highest average of anger was associated with the topic 'Insulin use' (25.4% compared with 13.9% on average in all other topics, p<0.001). 'Food habits' was associated with joy (39.01% compared with 30.2% on average in all other topics, p<0.001), while all other topics were mainly dominated by high rates of sadness (more than 40%).

Sub-Saharan Africa

The topic 'Insulin' was associated with anger (22.5% compared with 15.3% on average in all other topics, p<0.001) because of users' angry reactions to diabetes misunderstanding and struggles to get insulin. The topic dealing with 'Glucose guardian' was dominated by joy (39.3% compared with 28.9% on average in all other topics, p<0.001) as users were thanking others for their help or shared excellent glucose levels. In comparison, all other topics were dominated by sadness (between 37% and 46.02%).

Details about the average probabilities of sentiment distribution are available in online supplemental appendix 3.

CONCLUSION

In this study, we used worldwide social media data to better assess the global diabetes burden, from the perspective of PWD, and to study regional differences, which will serve to design more patient-centered diabetes programs. Social media data provide direct access to individual points of view and experiences of PWD, which can improve our understanding of how diabetes impacts their daily lives.

We have shown that some concerns are universal and shared by different online communities of PWD, while others are region-specific (eg, North America, which has five insulin-related topics). We found that matters related to food, glycemic control, family and relatives, insulin and comorbidities were shared by at least four of the seven regions. Tweets in which users shared their concerns and experiences about their relatives' diabetes, family health history and comorbidities were associated with higher rates of sadness (47.2% of all related clusters and regions combined compared with 38.7% on average). On the contrary, most joyful tweets referred to users sharing advice, motivation and peer-supporting and encouraging each other (37.7% of all related clusters and all regions combined compared with 31.1% on average). We also observed that 5 out of the 18 topics of interest in North America were related to insulin pricing, unaffordability and the consequences of such pricing on health (on physical and mental health). Overall, these tweets correspond to 18.95% (n=101 019) of all tweets originating from the USA (n=532 981).[21] Additionally, these topics were associated with higher rates of anger (28.04% compared with 19.2% on average in the USA and 19.1% in North America). Meanwhile, users from Europe and Asia and other regions (Europe and Central Asia, East Asia and Pacific) were sympathetic to patients from the USA, sharing their disgust and misunderstanding of the insulin pricing gap between their region and the USA. These results from North America are consistent with the previous work from Ahne *et al*,[14] who showed that insulin pricing is a central concern among PWD on Twitter in the USA.

Presumably, no previous study relied on such an extensive international database of posts from PWD to describe the diabetes burden. Our approach is more inclusive than those relying on questionnaires, such as patient-reported outcome measures or patient-reported experience measures scales with predefined items. We monitored key diabetes-related concerns of PWD and quantified the associated emotions in different communities around the world. We have observed an elevated global burden of diabetes, with regional specificities that need to be taken into account more diligently.[22] Diabetes-related distress is present in every diabetes community and is sometimes under-researched, such as in Sub-Saharan Africa, and social media can help overcome these concerns.[23] Özcan *et al*[24] studied people with type 2 diabetes from different ethnicities in the Netherlands and showed that ethnicity is independently associated with high diabetes distress. However, Gariepy *et al*[25] showed that diabetes distress in people with type 2 diabetes potentially varies according to some geographical and sociodemographic factors (such as social and physical order or cultural and social environment), which reinforces our hypothesis to compare diabetes burden determinants in different regions of the world. Besides, patients' state of mind heavily influences their self-management habits. Richman *et al* showed that positive emotions were associated with overall better health status, whereas Coccaro *et al* suggested that diabetes distress is associated with negative emotions and the regulation of emotions.[26 27] Thus, as recommended by Kalra *et al*[28], tackling patients' intellectual and emotional needs would be one solution to overcome the psychological barrier to adherence and self-care. Our findings corroborate earlier research, indicating that diabetes burden is a common issue discussed on social media in all different regions of the world and at different levels of severity. These findings also suggest that diabetes self-management is one of the biggest concerns, as PWD from the seven World Bank Regions shared concerns regarding glycemic control and food. Moreover, concerns at the regional level were identified, such as insulin pricing in North America or the fear of complications and comorbidities in Latin America and the Caribbean. This discovery highlights the need to develop new global methodologies to tackle universal concerns regarding self-care and focus on more specific ones at a regional or country level to improve PWD experiences and deal with their outcomes.

This study has several limitations. First, the list of the diabetes-related keywords we used to collect the tweets may have been incomplete. This list has been created by translating an original list of English keywords, and we may have missed specific local diabetes-related keywords and associated issues in some countries. Second, some language-specific subtleties may have gone astray, as translating non-English tweets to English may obscure the original meaning. Third, although this study essays the diabetes burden on a global level, we did not manage to recover data from every country. However, this is the most comprehensive analysis on an international scale to date. Fourth, a bias in the geolocation analysis might exist, as the location is self-reported by users. We manually excluded areas that appeared to be fake. Some tweets have been localized as coming from China where Twitter is blocked. Twitter is still accessed by a lot of Chinese people who are, for instance, using a VPN. This may explain why some users localize themselves in China.[29] Furthermore, the geographical coordinates provided by a tweet's metadata were identified as being, by default, in the center of the country. As a result, the distribution map of the tweets shows geographical markers that are not necessarily in populated areas. Fifth, the precision of the different classifiers we used was not perfect. An additional limitation is that our results are based on subjective statements from people using social media and do not represent all PWD. Finally, due to the prevalence of sarcasm and irony on social media and the fact that we searched to define key emotions in every tweet, we cannot ensure that all emotions were correctly identified, despite our efforts to remove jokes and irony.

In this work, we demonstrated that the global needs and concerns of PWD varied vastly based on region and that the diabetes burden was perceived differently, despite some shared concerns. Our results suggest a necessity to improve the integration of these regional and global factors into future diabetes programs to enhance patient-centric diabetes research and care from the perspective of people with diabetes. This will contribute to improving the personalization of diabetes care and self-management.

Contributors GF, as the guarantor, accepts full responsibility for the work and the conduct of the study, had access to the data, and controlled the decision to publish. The authors' contributions were as follows: GF designed the research; CB and GF conducted the research; AA and CB collected the data; CB, GF, and AA labeled the data; CB and GF analyzed data; CB and GF interpreted the data; CB and GF drafted the article; GF, AF, DM, PK, GA and AA revised the manuscript critically.

Funding This work was supported by the MSDAvenir Foundation (World Diabetes DIstress Study) and the Luxembourg Institute of Health.

Map disclaimer The inclusion of any map (including the depiction of any boundaries therein), or of any geographic or locational reference, does not imply the expression of any opinion whatsoever on the part of BMJ concerning the legal status of any country, territory, jurisdiction or area or of its authorities. Any such expression remains solely that of the relevant source and is not endorsed by BMJ. Maps are provided without any warranty of any kind, either express or implied.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. According to the Twitter API, tweets cannot be shared but tweets' IDs can be provided on request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs

Gloria Aguayo http://orcid.org/0000-0002-5625-1664
Guy Fagherazzi http://orcid.org/0000-0001-5033-5966

REFERENCES

1 Hessel F. *Burden of disease. encyclopedia of public health*. Springer, Dordrecht, 2008: 94–6.
2 Organization WH, Others. *Burden of disease: what is it and why is it important for safer food. who int*, 2019.
3 Measures of disease burden (event-based and time-based) and population attributable risks including identification of comparison groups appropriate to public health, 2010. Available: https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/measures-disease-burden [Accessed 20 Oct 2021].
4 Kiriella DA, Islam S, Oridota O, *et al*. Unraveling the concepts of distress, burnout, and depression in type 1 diabetes: a scoping review. *EClinicalMedicine* 2021;40:101118.
5 BridgetChapple. What is diabetes distress and burnout? In: Diabetes UK [Internet]. Available: https://www.diabetes.org.uk/guide-to-diabetes/emotions/diabetes-burnout [Accessed 17 Feb 2022].
6 Baek RN, Tanenbaum ML, Gonzalez JS. Diabetes burden and diabetes distress: the buffering effect of social support. *Ann Behav Med* 2014;48:145–55.
7 Mohamed Ridhwan K, Hargreaves CA. Leveraging Twitter data to understand public Sentiment for the COVID-19 outbreak in Singapore. *International Journal of Information Management Data Insights* 2021;1:100021.
8 Q1 2019 earnings report. Twitter. Available: https://s22.q4cdn.com/826641620/files/doc_financials/2019/q1/Q1-2019-Slide-Presentation.pdf
9 Yeung AWK, Kletecka-Pulker M, Eibensteiner F, *et al*. Implications of Twitter in health-related research: a landscape analysis of the scientific literature. *Front Public Health* 2021;9:654481.
10 Bour C, Ahne A, Schmitz S, *et al*. The use of social media for health research purposes: Scoping review. *J Med Internet Res* 2021;23:e25736.
11 Jordan S, Hovet S, Fung I, *et al*. Using Twitter for public health surveillance from monitoring and prediction to public response. *Data* 2018;4:6.
12 Weng J, Lee BS. Event detection in Twitter. *Proceedings of the Fifth International Conference on Weblogs and Social Media*; July 17-21, 2011, Barcelona, Catalonia, Spain, 2011.
13 Kreider KE. Diabetes distress or major depressive disorder? A practical approach to diagnosing and treating psychological comorbidities of diabetes. *Diabetes Ther* 2017;8:1–7.
14 Ahne A, Orchard F, Tannier X, *et al*. Insulin pricing and other major diabetes-related concerns in the USA: a study of 46 407 tweets between 2017 and 2019. *BMJ Open Diabetes Res Care* 2020;8:e001190.
15 World bank country and lending groups – world bank data help desk. Available: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups [Accessed 11 Oct 2021].
16 Hutto C, Gilbert E. Vader: a parsimonious rule-based model for Sentiment analysis of social media text. *ICWSM* 2014;8:216–25.
17 MacQueen JB. *Some methods for classification and analysis of multivariate observations*, 1966.
18 Steinhaus H O. Sur La division des Corps matériels en parties. *Bull Acad Polon Sci* 1956;1:801.
19 Jack RE, Garrod OGB, Schyns PG. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Curr Biol* 2014;24:187–92.
20 Mulcahy AW, Schwam D, Edenfield N. *Comparing insulin prices in the United States to other countries*. Rand Corporation, 2020.
21 Willner S, Whittemore R, Keene D. "Life or death": Experiences of insulin insecurity among adults with type 1 diabetes in the United States. *SSM Popul Health* 2020;11:100624.

22 Lin X, Xu Y, Pan X, *et al*. Global, regional, and national burden and trend of diabetes in 195 countries and territories: an analysis from 1990 to 2025. *Sci Rep* 2020;10:1–11.

23 Zimmermann M, Bunn C, Namadingo H, *et al*. Experiences of type 2 diabetes in sub-Saharan Africa: a scoping review. *Glob Health Res Policy* 2018;3:1–13.

24 Özcan B, Rutters F, Snoek FJ, *et al*. High diabetes distress among ethnic minorities is not explained by metabolic, cardiovascular, or lifestyle factors: findings from the Dutch diabetes pearl cohort. *Diabetes Care* 2018;41:1854–61.

25 Gariepy G, Smith KJ, Schmitz N. Diabetes distress and neighborhood characteristics in people with type 2 diabetes. *J Psychosom Res* 2013;75:147–52.

26 Coccaro EF, Lazarus S, Joseph J, *et al*. Emotional regulation and diabetes distress in adults with type 1 and type 2 diabetes. *Diabetes Care* 2021;44:20–5.

27 Richman LS, Kubzansky L, Maselko J, *et al*. Positive emotion and health: going beyond the negative. *Health Psychol* 2005;24:422–9.

28 Kalra S, Jena BN, Yeravdekar R. Emotional and psychological needs of people with diabetes. *Indian J Endocrinol Metab* 2018;22:696.

29 Bamman D, O'Connor B, Smith N. Censorship and deletion practices in Chinese social media. *First Monday* 2012. doi:10.5210/fm.v17i3.3943. [Epub ahead of print: 19 Sep 2022].

Language	**Keywords**	**Countriescovered**
Afrikaans	insulien, #insulien, suikersiekte, #suikersiekte, diabeet, #diabeet, Bloedglukose, #blodglucose	South Africa, Namibia
Amharic	ኢንሱሊን, #ኢንሱሊን, የስኳር ህመም, #የስኳር, የስኳር ህመም, #የሱካር በሽታ, የደም ግሉኮስ, #የግሉኮስ	Ethiopia
Arabic	لأنسولين, #الأنسولين, داء السكري, #داءالسكري, مريض بالسكر, #مريض بالسكر, جلوكوز الدم, يكافح السكري, مرض السكر النوع 1, مرض السكر النوع 2	Algeria, Bahrain, Chad, The Comoros, Djibouti, Egypt, Eritrea, Iraq, Jordan, Kuwait, Lebanon, Libya, Mauritania, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Somalia, Sudan, Syria, Tanzania, Tunisia, United Arab Emirates,Yemen
Chinese	胰岛素, #胰岛素, 糖尿病, #胰岛素, 糖尿病, #糖尿病患者, 血糖, #血糖, #2型糖尿病, #1型糖尿病	China, Singapore, Taiwan (Republic of China)
Danish	diabetiske,#diabetiske	Denmark
Dutch	#suikerziekte, bloed glucose, #bloedglucose, #diabetestype1, #diabetestype2	Aruba, Belgium, Curacao, The Netherlands, Sint Maarten, Suriname
English	insulin, #insulin, diabetes, #diabetes, diabetic, #diabetic, #diabeticproblems, blood glucose, #bloodglucose, blood sugar, #bloodsugar, #diabeticstruggles, #lifewithdiabetes, #type1diabetes, #type2diabetes, #insulin4all, #thisisdiabetes, #stopdiabetes, #fingerprick	Antigua and Barbuda, Australia, The Bahamas, Barbados, Belize, Canada, Dominica, Grenada, Guyana, Ireland, Jamaica, Malta, New Zealand, St Kitts and Nevis, St Lucia, St Vincent and the Grenadines, Trinidad and Tobago, United Kingdom, United States of America
Filipino	Diyabetis, #diyabetis, #dyabetiko, Dugo glucose, #asukalsadugo, problema sa diabetes	Philippines

French	insuline, #insuline, Diabète, #Diabète, diabétique, #diabétique, #problèmedediabétique, glucose sanguin, #glucosesanguin, #maviedediabétique, #diabètetype1, #diabète de type 2, #diabètedetype1, #diabetetype2, #mondiabete	France, Canada, Belgium, Switzerland, Congo-Kinshasa, Congo-Brazzaville, Côte d'Ivoire, Madagascar, Cameroon, Burkina Faso, Niger, Mali, Senegal, Haiti, Benin
German	Diabetiker, #Diabetiker, #diabetikerproblem, Blutzucker, #blutzucker, #meindiabetes, #Diabetikerleben, #diabetestyp1, #Typ1Diabetes, #diabetestyp2, #Typ2Diabetes	Germany, Belgium, Austria, Switzerland, Luxembourg, Liechtenstein
Greek	ινσουλίνη, #ινσουλίνη, Διαβήτης, #Διαβήτης, διαβητικός, #διαβητικός, γλυκόζη αίματος, #γλυκόζηςστοαίμα, ζωή με διαβήτη, #διαβήτηςτύπου1, #διαβήτης τύπου 2	Greece, Cyprus
Hausa	ciwon diabet, #ciwonsukari, jini glucose, #jiniglucose	Nigeria, Niger, Cameroon, Chad, Sudan
Hindi	इंसुलिन, #इंसुलिन, मधुमेह, #मधुमेह, मधुमेह की समस्या, रक्त द्राक्ष - शर्करा, मधुमेह के संघर्ष, के साथ जीवन मधुमेह, टाइप 1 मधुमेह, टाइप 2 मधुमेह, madhumeh	India
Indonesian	gula darah, #guladarah, perjuangan diabetes, hidup dengan diabetes, diabetes tipe 1, #diabetestipe1, diabetes tipe 2, #diabetestipe2	Indonesia
Italian	glucosio nel sangue, #glucosionelsangue, diabete di tipo 1, #diabeteditipo1, diabete di tipo 2, #diabeteditipo2	Italy, San Marino, Switzerland, Vatican City
Japanese	インスリン, #インシュリン, 糖尿病, #糖尿病, 糖尿病, #糖尿病の, 糖尿病の問題, 血糖, #血糖, 1型糖尿病, #1型糖尿病, 2型糖尿病, #2型糖尿病, #私の糖尿病	Japan

Korean	인슐린, #인슐린당뇨병, #당뇨병, 당뇨병 환자, #당뇨병 환자, 혈당, #혈당, 당뇨병 투쟁, 제 1 형 당뇨병, 제 2 형 당뇨병	Korea
Malay	pesakit kencing manis, #pesakitkencingmanis, glukosa darah, #glukosadarah, diabetes jenis 1, diabetes jenis 2	Malaysia, Brunei, Indonesia, Singapore, The Philippines, Thailand
Norwegian	blodsukker, #blodsukker, #mindiabetes	Norway
Polish	cukrzyca, #cukrzyca, cukrzycowy, #cukrzycowy, glukoza we krwi, #glukozawekrwi, #cukrzycatypu1, #cukrzycatypu2, #mojacukrzyca, życie z cukrzycą	Poland
Portuguese	#problemasdiabéticos, glicose no sangue, #glicosenosangue, #vidacomdiabetes	Brazil, Mozambique, Angola, Portugal, Guinea-Bissau, East Timor, Equatorial Guinea, Cape Verde, São Tomé and Príncipe
Romanian	insulină, #insulinăDiabet, #Diabet, glucoza din sange, #glucozadinsange, diabet de tip 1, diabet de tip 2	Romania, Republic of Moldova
Russian	инсулин, #инсулин, сахарный диабет, #сахарный диабет, диабетом, #диабетический, содержание глюкозы в крови, #жизньсдиабетом, диабет 1 типа, #диабет1типа, диабет 2 типа, #диабет2типа	Russia, Belarus, Kyrgyzstan, Kazakhstan
Spanish	insulina, #insulina, diabético, #diabético, #problemas de la diabetes, glucosa en sangre, #glucosaenlasangre, #Diabetestipo1, #Diabetestipo2, #detenerladiabetes, #estoesdiabetes	Argentina, Bolivia, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Equatorial Guinea, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Spain, Uruguay
Swahili	Insulini, #insulini, Kisukari, #kisukari, Glucose ya damu, kisukari type 1, kisukari type 2	Tanzania, Kenya, Uganda, The Democratic Republic of Congo, the Comoros Islands
Swedish	blodsocker, #blodsocker, #typ1diabetes, #typ2diabetes	Sweden

Thai	กลูโคส, #กลูโคส, อินซูลิน, #อินซูลิน, โรคเบาหวาน, #โรคเบาหวาน, การต่อสู้กับโรคเบาหวาน, ระดับน้ำตาลในเลือด, #ระดับน้ำตาลในเลือด, การต่อสู้กับโรคเบาหวาน, โรคเบาหวานประเภท 1, โรคเบาหวานประเภท 2, โรคเบาหวานของฉัน	Thailand, Vietnam, Laos
Turkish	ensülin, #ensülin, şeker hastalığı, #şeker hastalığı, şeker hastası, #şekerhastası, kan şekeri, #kan şekeri, tip 1 diyabet, #tip1diyabet, tip 2 diyabet, #tip2diyabet	Albania, Azerbaijan, Bosnia and Herzegovina, Bulgaria, Greece, Northern Cyprus, Kosovo, the Republic of Macedonia, Moldova, Montenegro, Romania, Russia, Serbia, Syria, Turkey, Turkmenistan,Uzbekistan
Urdu	انسولین, #انسولین, ذیابیطس, #ذیابیطس, خون میں گلوکوز, #خون میں گلوکوز, ٹائپ 1 ذیابیطس, ٹائپ 2 ذیابیطس,	Pakistan, India, Afghanistan, Saudi Arabia
Vietnamese	Bệnh tiểu đường, #Bệnhtiểuđường, Bệnh tiểu đường., #mắcbệnhđáiđường, đường huyết, #đường huyết, bệnh tiểu đường loại 1, bệnh tiểu đường loại 2	Vietnam

Details on data processing

We followed most of the same processes described by *Ahne et al.* in their Supplementary online material 2.[1]

Data collection

We accessed the API Standard v1.1 by applying for a Twitter Developer account.[2] We connected to Twitter's API using the Python library Tweepy and streamed the keywords to collect tweets together with users' metadata (attributes provided on user's profile, such as user screen name, user location, user description.).[3]

Data representation

We used FastText implementation in the Gensim package to process each word into a vector to extract meaningful semantic relationships.[4] The average of each tweet's word vector representations was then used to model it and similarities in their semantics were analyzed.

Geolocation process

We aimed to keep only geolocated tweets for this study. To do so, we first looked at users' locations, provided by users themselves in their public profile, and removed all tweets without such location. We then grouped all locations and manually detected which ones were fake in order to exclude it (e.g. "the Internet", "in Hell"). We also replaced all contractions to full words (eg. "CA, USA" to "California, United States of America") to facilitate geolocation by the Python package geograpy4.[5] We then applied the package to the different grouped locations in order to access the latitude and the longitude of the location associated with each tweet. If the determined place was a country, latitude and longitude were identified in the center of the country. We evaluated the overall precision of this geolocation step as 85%.

Tweets translation

In order to apply unique classifiers to all tweets, we translated all tweets originally not written in English to English using the Python package deep-translator, a free and unlimited python API for Google Translate.[6]

Personal content classifier / Jokes classifier

We developed two classifiers that aimed to filter out institutional tweets and jokes in order to keep only tweets with personal content and not jokes or irony about diabetes. A tweet was considered as personal if the user expressed his feelings and own experiences, dealing with his own diabetes or a relative one. A tweet was considered as a joke/irony if diabetes was used as an insult or with a sugar-related joke. To train these two classifiers, three authors (AA, CB, GF) manually labelled 1648 randomly chosen tweets for the personal classifier and 1398 for the jokes one. We used *Bidirectional Encoder Representations from Transformers* (BERT), a machine learning technique for natural language processing pre-training developed by Google.[7] More precisely, we applied *BERTweet*, a pre-trained model for English Tweets.[8] The overall accuracy of the personal content classifier was 88% and the accuracy from the jokes classifier was 94%.

Gender and Type of diabetes classifier

Following the scripts by *Ahne et al.*, we trained classifiers to predict gender (male, female, unknown) and type of diabetes (type 1, type 2, unknown) from each user.[1] Three authors (AA, CB, GF) manually labelled 1670 tweets from different regions of the world to better match our dataset. A SVM was trained and reached an accuracy of 86% for the gender classifier and 74% for the type of diabetes classifier.

Topic extraction

All tweets are represented via their word vector representations. Then, a K-means algorithm was applied to the tweets in each region. To define the optimal number of clusters *k*, the silhouette score average for *k* between 4 and 24 was calculated and silhouette analysis applied.[9]

Emotion classifier

Initially, two authors (CB, GF) labelled 1000 randomly chosen tweets according to the main emotion in the tweet text. In order to increase the accuracy of our classifier, we combined our dataset with online labelled datasets of emotions which led to the extension of 47,000 tweets from Github and Kaggle.[10–12] We then trained a Calibrated Linear Support Vector classifier to predict the probability of a tweet belonging to each of the four emotions.[13] We applied this classifier to all tweets to predict the probability of a tweet belonging to each of the four emotions.

This led to the creation of a dataset of 47,926 texts labelled as one of the following emotions: joy, anger, fear or sadness. We based our classifier on the following scripts: https://thecleverprogrammer.com/2021/02/19/text-emotions-detection-with-machine-learning/. We calibrated the classifier in order to predict the probability of a tweet to belong to each of the 4 classes.[14] The calibrated SVM was trained and reached an accuracy of 80.56% (precision to predict joy only: 88%, precision to predict fear only: 94%, precision to predict anger only: 93%, precision to predict sadness only: 89%).

Details

Python (V.3.6) with the packages scikit-learn (machine learning algorithms and data preprocessing methods), gensim (text processing, word representation), and statsmodels (module for the estimation of many different statistical models) were exploited for the analysis.[15–17] Tableau (2020.4) and OpenStreetMap 2021 were used to visualize the data.

Additional references

1. Ahne A, Orchard F, Tannier X, Perchoux C, Balkau B, Pagoto S, et al. Insulin pricing and other major diabetes-related concerns in the USA: a study of 46 407 tweets between 2017 and 2019. BMJ Open Diabetes Res Care [Internet]. 2020 Jun;8(1). Available from: http://dx.doi.org/10.1136/bmjdrc-2020-001190

2. Website [Internet]. Available from: “Data Dictionary: Standard v1.1.” n.d. Twitter Developer Platform. https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/user.

3. Roesslein J. Tweepy: Twitter for python. URL: https://github com/tweepy/tweepy. 2020;484.

4. fastText [Internet]. [cited 2021 Oct 12]. Available from: https://fasttext.cc/index.html

5. ThomasShih. GitHub - ThomasShih/geograpy4: Extract countries, regions and cities from a URL [Internet]. [cited 2021 Oct 13]. Available from: https://github.com/ThomasShih/geograpy4

6. Welcome to deep_translator’s documentation! — deep_translator documentation [Internet]. [cited 2021 Oct 12]. Available from: https://deep-translator.readthedocs.io/en/latest/

7. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Internet]. 2018 [cited 2021 Oct 12]. Available from: http://arxiv.org/abs/1810.04805

8. Nguyen DQ, Vu T, Nguyen AT. BERTweet: A pre-trained language model for English Tweets [Internet]. 2020 [cited 2021 Oct 12]. Available from: http://arxiv.org/abs/2005.10200

9. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis [Internet]. Vol. 20, Journal of Computational and Applied Mathematics. 1987. p. 53–65. Available from: http://dx.doi.org/10.1016/0377-0427(87)90125-7

10. Praveen. Emotions dataset for NLP [Internet]. [cited 2021 Oct 13]. Available from: https://www.kaggle.com/praveengovi/emotions-dataset-for-nlp

11. Jcharis. GitHub - Jcharis/end2end-nlp-project: End 2 End NLP Project with Python [Internet]. [cited 2021 Oct 13]. Available from: https://github.com/Jcharis/end2end-nlp-project

12. Kharwal A. Text Emotions Detection with Machine Learning [Internet]. 2021 [cited 2021 Oct 13]. Available from: https://thecleverprogrammer.com/2021/02/19/text-emotions-detection-with-machine-learning/

13. Tang Y. Deep Learning using Linear Support Vector Machines [Internet]. arXiv [cs.LG]. 2013. Available from: http://arxiv.org/abs/1306.0239

14. sklearn.calibration.CalibratedClassifierCV [Internet]. [cited 2021 Oct 13]. Available from: https://scikit-learn/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html

15. Rehurek R, Sojka P. Gensim--python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic. 2011;3(2).

16. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.

17. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python [Internet]. Proceedings of the 9th Python in Science Conference. 2010. Available from: http://dx.doi.org/10.25080/majora-92bf1922-011