



Mining the sociome for Health Informatics: Analysis of therapeutic lifestyle adherence of diabetic patients in Twitter

Gael Pérez-Rodríguez^{a,b,c}, Martín Pérez-Pérez^{a,b,c}, Florentino Fdez-Riverola^{a,b,c},
Anália Lourenço^{a,b,c,d,*}

^a Department of Computer Science, University of Vigo, ESEI, Campus As Lagoas, 32004 Ourense, Spain

^b The Biomedical Research Centre (CINBIO), Campus Universitario Lagoas-Marcosende, 36310 Vigo, Spain

^c SING Research Group, Galicia Sur Health Research Institute (ISS Galicia Sur), SERGAS-UVIGO, Spain

^d Centre of Biological Engineering (CEB), University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal



ARTICLE INFO

Article history:

Received 7 November 2019

Received in revised form 30 March 2020

Accepted 18 April 2020

Available online 23 April 2020

Keywords:

Sociome

Community detection

Topic modelling

Knowledge graphs

Diabetes

Twitter

ABSTRACT

In recent years, the number of active users in social media has grown exponentially. Despite the thematic diversity of the messages, social media have become an important vehicle to disseminate health information as well as to gather insights about patients' experiences and emotional intelligence. Therefore, the present work proposes a new methodology of analysis to identify and interpret the behaviour, perceptions and appreciations of patients and close relatives towards a health condition through their social interactions. At the core of this methodology are techniques of natural language processing and machine learning as well as the reconstruction of knowledge graphs, and further graph mining. The case study is the diabetes community, and more specifically, the patients communicating about type 1 diabetes (T1D) and type 2 diabetes (T2D). The results produced in this study show the effectiveness of the proposed method to discover useful and non-trivial knowledge about patient perceptions of disease. Such knowledge may be used in the context of Health Informatics to promote healthy lifestyles in more efficient ways as well as to improve communication with the patients.

© 2020 Elsevier B.V. All rights reserved.

1. Introduction

In recent years, the number of active users in social media has grown tremendously. It is estimated that the amount of social media users will increase from 2.65 billion in 2018 to 3.1 billion in 2021 [1]. Within this scenario, Twitter is one of nowadays' most popular social networks, with more than 330 million registered users producing public messages (i.e. tweets) daily [2].

Although the content of the tweets is quite diverse, several studies denote the importance of these contents for public and consumer health informatics [3–5]. Social media sites are easily accessible, broad audience vehicles to disseminate health information as well as to gather insights about patients' experiences and emotional intelligence. Self-management has become a critical approach for improving health outcomes among these

patients and, in particular, a majority of adolescents report the use of (and even prefer to communicate via) social networking sites.

The present work proposes a methodology of analysis combining natural language processing (NLP), machine learning (ML) and graph mining to identify and interpret the behaviour, perceptions and appreciations of patients and close relatives towards a health condition through their social interactions. The proof of concept is the social interplay played by the diabetes communities on Twitter, namely, the users communicating about type 1 diabetes (T1D) and type 2 diabetes (T2D). It has been shown that conflicting beliefs about diabetes contribute to greater distress [6–8], whereas, positive relations and shared coping between individuals living with the disease can alleviate distress [9,10]. Hence, social media may offer new opportunities in diabetes management, notably the development of new dynamic behavioural engagement models and enhancements to existing health promotion models [11]. Such approaches should focus on the self-management skills of the patients, wellness knowledge, education, psychosocial assessments, satisfaction, and rate of adherence with treatment (namely to diet-related recommendations) [12].

So, the main contribution of this work lays in developing the capacity to promote healthy lifestyles and to communicate with the patients more efficiently, two main objectives of the public

* Correspondence to: ESEI: Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004 Ourense, Spain.

E-mail addresses: gaeperez@uvigo.es (G. Pérez-Rodríguez), martiperez@uvigo.es (M. Pérez-Pérez), riverola@uvigo.es (F. Fdez-Riverola), analialourenco@uvigo.es (A. Lourenço).

URLs: <http://sing-group.org/> (G. Pérez-Rodríguez), <http://sing-group.org/> (M. Pérez-Pérez), <http://sing-group.org/riverola> (F. Fdez-Riverola), <http://sing-group.org/> (A. Lourenço).

and consumer health informatics. To the best of our knowledge, previous works on diabetes *sociome* did not address the analysis at such a holistic and multidimensional scale.

The remainder of this paper is organised as follows. Section 2 provides a literature survey on social mining, whereas diabetes conditions are introduced in Section 3. Section 4 presents the proposed methodology and Section 5 discusses the results obtained for the diabetes case study. Finally, Section 6 summarises the work in terms of aims and results and describes new lines of research.

2. Related work

The study of the contents generated in social media platforms, such as Twitter, Facebook and Reddit, is an active topic of research [13]. Among those studies, Twitter contents, especially health-related posts, has received increasing research attention in recent years [14]. In particular, there has been a proliferation of works on social mining about chronic diseases, such as diabetes or cancer [15]. For example, the study of Beguerisse-Díaz et al. [16] aimed to discover the general themes of diabetes conversations in Twitter (e.g. news and commercials), without considering groups of users (e.g. patients or public agencies). The authors combined techniques from anthropology, network science and information retrieval to discover the most influential and contributing people in the diabetes communities. Another work proposed a multi-component semantic and linguistic framework to collect, process and analyse topics of interest about the interrelations between diabetes, diet, exercise and obesity [17]. The authors applied the Latent Dirichlet Allocation technique to discover the topics of discussion for each of these health issues and then calculated the correlations among them. Lastly, the work of Gabarron et al. [18] analysed the emotions (i.e. positive and negative sentiments) of T1D and T2D conversations for one-week period. This analysis showed that tweets related to T2D tend to have a more negative connotation than those related to T1D.

The present work was also inspired by works addressing the social interplay of other health-related issues. For example, Bello et al. [19] explored a new methodology to detect, track and categorise the discussion of vaccination communities. This work explored different community detection algorithms to analyse the different vaccination groups based on their retweets. Other works explored different graph methodologies for the reconstruction of semantic networks that provide insights into the contents and interplay of social conversations. For example, the study of Kim et al. [20] identified core keywords and sub-topic groups of studies in domestic and geriatric frailty syndrome by analysing and clustering keywords based on the co-occurrence frequency and centrality analysis. The work of Thag et al. [21] performed an unsupervised analysis to explore how the word frequencies and their co-occurrence can be applied for assessing the frames used in social media discussions. The authors focused their work in the emerging infectious diseases in the United States and based their study in the creation, interpretation, and quantification of semantic networks. Also, Perez et al. [22] implemented a supervised analysis to reconstruct semantic networks based on the co-occurrence of the health discussed topics to discover how patients discuss, feel, and react to symptoms, changes in habits, and medication based on the frequency of relations.

Taking all such previous studies into consideration, the present work developed a complementary *sociome* analysis method, focused on the point of view of health management. In this sense, this work explores the categorisation of users as organisations and individuals, and more specifically, of individuals as patients. Moreover, this work proposes a formal method to combine NLP and graph mining, through the application of domain ontologies, to explore patterns in patient's communications. To the best

of authors' knowledge, no previous work addressed the characterisation of users by role (i.e. organisations, individuals and patients) and the analysis of the emotions sensed by the health communities in T1D and T2D. In particular, the reconstruction of semantic knowledge graphs, using standardised vocabulary, enables the study of opinions on diabetes-related trends and patterns of conversation.

3. The case of diabetes mellitus

Diabetes mellitus (or diabetes) is a chronic, lifetime condition that affects the ability of the body to use the energy found in food. This disease affects more than 425 million people worldwide, and the number is expected to reach 693 million by the year 2045 [23]. Also, the directed (e.g. medical care) and undirected (e.g. loss of productivity) economic costs associated with this condition are incrementing every year. Due to this, the cost of producing proper treatments and medical support for the general population is increasing and causing that a wide part of patients and their families cannot afford them. This represents a barrier to appropriate care in vulnerable populations increasing, even more, adverse outcomes and costs [24,25].

There are multiple types of diabetes, some of which are more common than others [26]. The most prevalent form of diabetes worldwide is T2D, whereas T1D is more common in children and gestational diabetes may occur during pregnancy. Other, newly "discovered" types are still under research, like type 3 diabetes, which is related to Alzheimer's disease [27]. All of these conditions have in common the malfunctioning of the hormone insulin that enables the degradation of ingested carbohydrates into glucose and ensures that the body takes in the glucose and uses it for energy. The inability of the body to process glucose leads to high levels of blood glucose and thus, to the damage of the blood vessels of main organs, such as the kidneys, the heart, the eyes, and the nervous system. Conditions differ in terms of epidemiology and thus, treatment.

T1D is also called insulin-dependent diabetes and was once called juvenile-onset diabetes because it often begins in childhood. This type of diabetes affects 7%–12% of individuals diagnosed with diabetes [28]. This is an autoimmune condition, i.e. the body attacking its pancreas with antibodies, which makes it unable to produce insulin. This type of diabetes may be caused by a genetic predisposition, but it can also be the result of faulty beta cells in the pancreas that normally produce insulin. Most of the associated medical risks stem from damage to the blood vessels in the eyes (i.e. diabetic retinopathy), nerves (i.e. diabetic neuropathy), and kidneys (i.e. diabetic nephropathy). Even more serious, it may lead to an increased risk of heart disease and stroke. Treatment for T1D involves taking insulin, which needs to be injected subcutaneously. Individuals with T1D must change their lifestyle significantly, including frequent testing of blood sugar levels, careful meal planning, daily exercise, and taking insulin and other medications as needed [29].

T2D accounts for more than 451 million people and other 352 million people are at risk of developing it. Although this disease is worldwide, the Western Pacific Region has the highest prevalence with 168.4 million people (i.e. 37% of the total global diabetes population) [23]. T2D used to be called adult-onset diabetes, but the ever-increasing number of obese and overweight children has significantly raised the development of T2D in teenagers. T2D is often a milder form of diabetes than T1D, but T2D can still cause major health complications, particularly in the blood vessels that nourish the kidneys, nerves, and eyes. Likewise, it also increases the risk of heart disease and stroke. With T2D, either the pancreas usually produces some insulin, but the amount produced is not enough or the cells are resistant to it. Insulin resistance, or lack

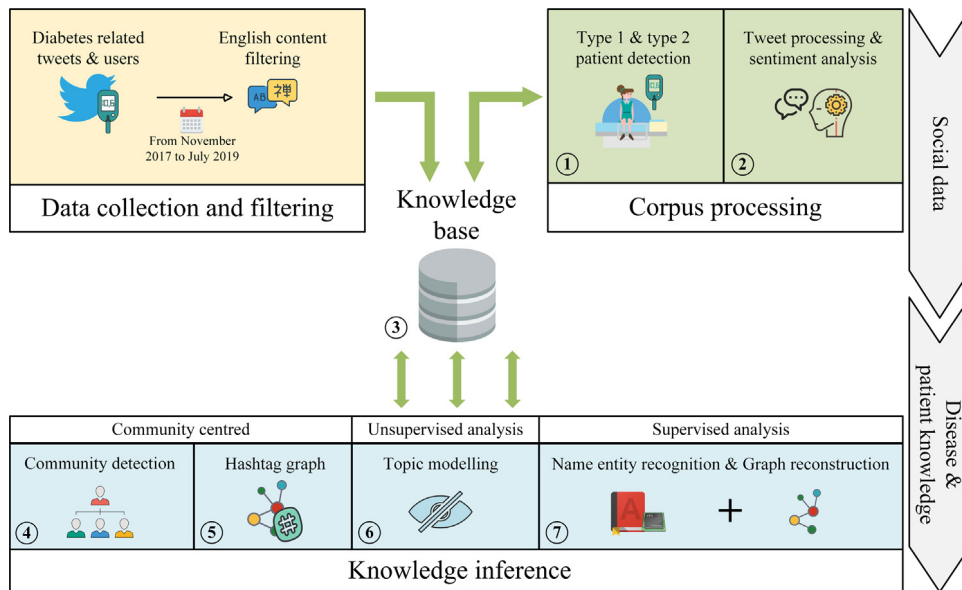


Fig. 1. The workflow implemented to retrieve and extract, process and analyse diabetes-related tweets.

of sensitivity to insulin, happens primarily in fat, liver, and muscle cells. Therefore, obese people (i.e. those with more than 20% over their ideal body weight for their height) are at high risk of developing this condition. With insulin resistance, the pancreas has to work overly hard to produce more insulin and, even then, it is hard to keep the sugar level normal. Therefore, T2D control is often focused on weight management, both in terms of nutrition and exercise. Unfortunately, this condition tends to progress, and diabetes medications are often needed. Periodic A1C testing may be advised to see how well diet, exercise, and medications are working to control blood sugar and prevent organ damage [30].

4. Materials and methods

The methodology developed here to retrieve, process and study the emotions, perceptions and appreciations of patients and close relatives, consists of three fundamental steps, i.e. (i) data collection and filtering; (ii) corpus processing; and (iii) knowledge inference (Fig. 1).

In the first step, the Twitter4J, a Java library for the Twitter API, was applied to retrieve diabetes-related tweets in English [31]. Specifically, the collection of posts dates from November 22nd, 2017 to July 1st, 2019, obtaining a total of 1.3 million unique tweets from 546,739 distinct users.

In the second step, an in-house developed algorithm was applied to identify patient accounts (see details in Section 4.1). Then, the tweets were processed (e.g. tokenization, stop words removal, lemmatization, etc.) and labelled according to its sentiment (see details in Section 4.2).

In the third step, knowledge inference is performed based on the reconstruction of knowledge graphs and unsupervised learning (see details in Section 4.3). User communities were detected through an unsupervised approach (see details in Section 4.4). Graph mining supported the exploration of hashtags used in T1D and T2D patient conversations (see details in Section 4.5) while topic modelling (see details in Section 4.6) and semantic entity recognition (see details in Section 4.7) enabled a more detailed characterisation of the contents of these conversations.

4.1. Identification of patient accounts

The algorithm devised to categorise the user accounts is two-fold, i.e. individual accounts are differentiated from organisation

accounts and then, the accounts of patients (and their relatives) are identified. The algorithm takes into consideration not suspended accounts and considers that an account belongs to an individual if one of the following criteria is met: (i) the user account name is recognised as a personal name based on a lookup over a worldwide name dictionary [32]; (ii) a face of a person is recognised in the user profile image by a convolutional neural network model [33]; (iii) Twitter recognises the user as a contributor or a translator; (iv) the account description is written in the first person, i.e. pronouns and their variant forms (i.e. possessive and reflexive); (v) the description has emojis or emoticons, (i.e. the assumption is that organisation profiles are described more formally); and (vi) the user account name or the description contains an English honorifics (e.g. Ms. or Mr.). The application of these criteria led to the labelling of 393,891 user accounts (72% of total users) as “Individual” and 152,848 (28% of total users) as “Other”. The F-score metric was at the basis of the evaluation of the accuracy of the algorithm [34]. Therefore, a random sample of 1000 users was labelled manually and then compared with the automatic labelling, resulting in an F-score of 0.84. The identification of patient-related accounts, i.e. people that often share first or closely related experiences about the disease, focused on the accounts labelled as “Individual”. This part of the algorithm is based on the matching of characteristic phrasing. In particular, it was devised to identify people who explicitly express suffering from the disease (i.e. “I have diabetes”) or relatives who talk about the condition in a third-person way (i.e. “My son has diabetes”). To do so, the matching rules try to find a pronoun and/or a list of verbs (e.g. “I have”, “I was diagnosed”, “He suffer”, etc.) related with a mention of the disease and/or its variants (i.e. “diabetes”, “t1d”, “dt2”, etc.). The application of these heuristics led to the labelling of 35,112 users as “Patient” (9% of Individual accounts) and 358,779 users as “Other” (91% of Individual accounts). Once again, to check the effectiveness of the algorithm, a random sample of 1000 tweets was labelled manually and then compared with the automatic labelling. The algorithm scored an F-score of 0.81.

For convenience in terms of further analysis, the corpus of patient tweets was split into two, i.e. a corpus containing patient-related tweets about T1D and another one containing patient-related tweets about T2D. Such splitting was based on previous user identification and the recognition of named conditions in

tweet contents. If the tweet mentioned both conditions, then it was included in both corpora. A total of 27,128 tweets were related to T1D patients (16.3% of patient tweets) and 15,112 tweets were related to T2D patients (9.1% of patient tweets).

4.2. Process social information

Special characters that did not provide useful information were removed (e.g. '&', '(', ')', '*', '+', '<', or '>'). Likewise, tweets were stripped of user mentions (represented with '@'), hashtags (represented with '#'), Uniform Resource Locators (URLs) and emojis. All these operations were carried out using the Twitter-Text library [35]. Besides, three or more consecutive and identical characters were removed from word tokens (e.g. 'haaaappppy' to 'haappy').

Spelling error correction was achieved using the Hunspell dictionary, i.e. a collection of specific medical terms [36]. The correction was done automatically by selecting the suggested word with the highest similarity to the original (incorrect) term. The similarity was calculated using the Normalized Levenshtein algorithm [37].

Although Twitter has increased the maximum length of tweets from 140 to 280 characters, the use of abbreviations is still very common in conversations. Therefore, a custom dictionary of abbreviations was constructed to perform an expansion of abbreviations and shorthand terms (e.g. GDM to gestational diabetes mellitus).

Additional processing was performed to prepare the tweets for topic modelling and named entity recognition (NER) tasks, namely: (i) tokenization (i.e. to split a set of text up into words, phrases or other meaningful elements); (ii) English and domain-specific stop words removal (i.e. too frequent, not content-bearing tokens); (iii) part of speech (POS) tagging (i.e. to identify the lexical category of each token); (iv) number removal; (v) small tokens removal (i.e. less than two characters); (vi) extra whitespaces removal; (vii) convert tokens to lowercase; and, (viii) lemmatization (i.e. to obtain the lexeme form of the tokens). Besides single word tokens (unigrams), bigrams and trigrams, i.e. contiguous of 2 or 3 sequences of tokens, were also considered. All the previous tasks were implemented using the Stanford CoreNLP pipeline [38].

Finally, the analysis of sentiments was performed using the Valence Aware Dictionary and sEntiment Reasoner (VADER). This is a lexicon and rule-based tool that is specifically adjusted to the detection of sentiments in social posts [39].

4.3. Reconstruction of knowledge graphs

The reconstruction of knowledge graphs obtained through the processing of large volumes of text has proven to be a very powerful tool to obtain new knowledge and discover non-trivial information patterns [40–42]. Thus, several techniques of knowledge graph reconstruction and analysis were introduced in the proposed methodology. The goal was to represent the information extracted accurately and be able to perform a comprehensive analysis of the exchanged contents and social interplay. Therefore, graph-based metrics were calculated to evaluate the relevance and interrelation of the discussed topics, and to explore the motivations behind user relationships and how the information flows. To this end, undirected knowledge graphs were applied to represent the co-occurrence of terms and the co-occurrence of hashtags, whereas the user interactions were depicted in a directed and weighted knowledge graph. The following equations define the rationale behind these knowledge graphs in a formal manner.

A corpus of tweets, C , given a specific domain, D , is represented as a collection of tweets:

$$C = \{T_0, T_1, \dots, T_{N-1}\} \quad (1)$$

where T_i is the i th tweet in corpus C and N represents the total number of tweets.

A domain, D , is represented by a set of meaningful concepts:

$$D = \{c_0, c_1, \dots, c_{N'-1}\} \quad (2)$$

where c_i is the i th concept associated with domain D and N' represents the total number of identified concepts in the tweet.

A concept, c , for the domain, D , is represented by a set of terms (n-grams) of similar meaning:

$$c = \{t_0, t_1, \dots, t_{J-1}\} \quad (3)$$

where t_i is the i th term associated with concept c and J represents the total number of associated terms.

Similarly, a hashtag, h , is represented by a set of user-selected terms (n-grams):

$$h = \{t_0, t_1, \dots, t_{J'-1}\} \quad (4)$$

where t_i is the i th term associated with hashtag h and J' represents the total number of hashtags in the tweet.

So, a tweet message, T , can be represented by the set of domain concepts, c , being mentioned:

$$T = \{c_0, c_1, \dots, c_{M-1}\} \quad (5)$$

where c_i is the i th concept associated with domain D and M represents the number of domain concepts mentioned. Conversely, the same tweet, T , can be also represented by the set of included hashtags, h :

$$T = \{h_0, h_1, \dots, h_{M'-1}\} \quad (6)$$

where h_i is the i th hashtag used in the message and M' represents the number of hashtags mentioned.

The idea of co-occurrence establishes that two concepts, or two hashtags, are more related to each other if they are used together more frequently. In this line, the co-occurrence between two concepts, c_i and c_j , is defined formally as:

$$co - occurrence(c_i, c_j) = \sum_{k=0}^n f(k, C) \quad (7)$$

where $f(k, C)$ is defined as:

$$f(k, C) = \begin{cases} 1, & \text{if } [c_i \in T_k \wedge c_j \in T_k] \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

The value of the *co-occurrence* (c_i, c_j) ranges between 0 and the size of the corpus, C . A value greater than zero represents the existence of an edge, e , between two vertexes v_i and v_j in the graph:

$$e_{c_i, c_j} = \begin{cases} \exists, & \text{if } co - occurrence(c_i, c_j) = 0 \\ \exists, & \text{if } co - occurrence(c_i, c_j) \geq 0 \end{cases} \quad (9)$$

Regarding the reconstruction of the user knowledge graph, there is a collection of users, U , defined as:

$$U = \{u_0, u_1, \dots, u_{S-1}\} \quad (10)$$

where u_i stands for the i th user and S represents the total number of users contributing to the collection of tweets.

A retweet, RT_{u_i, u_j} , between two specific users, u_i and u_j , is represented as an edge, e , between the vertexes v_i and v_j that represent those users, and it is defined as:

$$e_{u_i, u_j} = \begin{cases} \exists, & \text{if } RT_{u_i, u_j} = 0 \\ \exists, & \text{if } RT_{u_i, u_j} \geq 0 \end{cases} \quad (11)$$

In general, knowledge graphs are described in terms of the number of vertexes and edges and analysed using several well-known centrality metrics, namely degree centrality, betweenness centrality, clustering coefficient, and eigenvector centrality [43]. The degree centrality for undirected graphs measures the total amount of links with the other vertexes, and it is defined as:

$$D_c(v_i) = d_i \quad (12)$$

where d_i is the number of adjacent edges for a given vertex, v_i .

Similarly, in the case of directed graphs, the degree centrality measures the sum of the out-degree, $deg^+(v)$, and the in-degree, $deg^-(v)$, and it is defined as:

$$D_c(v_i) = \sum_{v \in V} deg^+(v) + \sum_{v \in V} deg^-(v) \quad (13)$$

where V stands for the set of the vertexes of the graph and v_i is the considered vertex.

The betweenness centrality measures the mediation role of the vertexes, and it is defined as:

$$B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (14)$$

where V stands for the set of the vertexes of the graph, σ_{st} represents the total number of shortest paths from vertex s to vertex t and $\sigma_{st}(v)$ is the number of those paths that pass through v .

The clustering coefficient measures how much a vertex is grouped (or interconnected) with its neighbours. The local clustering coefficient used for undirected graphs, G_i , for a vertex v_i is defined as:

$$G_i = \frac{2 \left| \{e_{ij}: v_i, v_j \in N_i, e_{ij} \in E\} \right|}{k_n(k_n - 1)} \quad (15)$$

where e_{ij} represents the edge that connects the vertex v_i with the vertex v_j , E is the set of edges that connects the set of the vertexes V , k_n is the number of neighbours of a vertex and N_i is the neighbourhood of a vertex, defined as:

$$N_i = \{v_j: e_{ij} \in E \vee e_{ji} \in E\} \quad (16)$$

where e_{ij} is the edge that connects the vertex v_i with the vertex v_j and E is the set of the edges that connects the set of the vertexes V .

Finally, the Eigenvector centrality measures the influence of a vertex on a graph while considering the importance of its neighbours, being defined as:

$$c_e(v_i) = \frac{1}{\lambda} \sum_{j=1}^k A_{j,i} c_e(v_j) \quad (17)$$

where A is the adjacency matrix of the graph, $C_e(v_j)$ denotes the eigenvector centrality of the vertex v_j , and λ represents the largest eigenvalue associated with the eigenvector of the adjacency matrix, A .

4.4. Community detection

User community detection was performed using the Community Clustering algorithm (GLay), an implementation of the Girvan–Newman fast greedy algorithm [44,45]. The quality of the partitions was assessed using the modularity metric [46]. The basic assumption is that a random graph is not expected to have a cluster structure. Therefore, the possible existence of clusters is revealed by the comparison between the actual density of edges

in a subgraph of the graph under analysis and the density that would be expected in a random subgraph:

$$modularity = \sum_i (e_{ii} - a_i^2) \quad (18)$$

where e_{ii} stands for the probability that an edge is in module i and a_i^2 represents the probability that a random edge would fall into module i . That is, it measures the fraction of the edges in the graph that connect vertexes of the same type (i.e. within community edges) minus the expected value of the same metric in a graph with the same community divisions but random connections between the vertexes. If the number of within-community edges is not greater than the random value then the *modularity* = 0. Values approaching *modularity* = 1, which is the maximum possible value, indicate the graph has a strong community structure.

The analysis of eigenvector centrality is complementary to the previous analysis and helps depict social influence, namely the most relevant users within each community [43]. The basic assumption is that having more important followers usually provides a higher relevance degree, i.e. generalises the concept of degree centrality by incorporating the importance of the neighbours.

Finally, the most active users in the community were identified based on a minimum threshold of 10 retweets and ignoring self-retweets.

4.5. Co-attention graph semantic analysis

A hashtag is a term or a phrase, prefixed with the symbol “#”, which summarises an explicit topic or idea. Having its origin in the social network Twitter, hashtags are mainly used to integrate information, i.e. as a shortcut to quickly find content on the same topic (e.g. highlight important topics or events).

The diabetes corpus contained 6690 unique hashtags in the T1D and T2D patient tweets, notably 42% of T1D and 28% of T2D tweets contained at least one or more hashtags. The knowledge graph representing the co-occurrence of hashtags was constructed with a double purpose: (i) showing the variety of keywords or topics in public conversations about diabetes; and (ii) gaining a better understanding of the overall satisfaction and subjectivity expressed by the people (i.e. average tweet sentiments) about these keywords or topics.

The hashtag connectedness was measured by the vertex degree, measuring the total amount of edges with other vertexes (i.e. the higher degree, the more central is the vertex), and, the clustering coefficient was used to measure the tendency of the vertex to form local interconnected groups [47]. The analysis of the relation between the degree and the local clustering coefficients offers a different perspective of the topology of complex graphs and allows the evaluation of the different roles of the terms [48]. Highly connected vertexes (i.e. high degree) with low clustering coefficient are associated with inter-modular hubs, whereas highly connected and highly clustered vertexes are associated with intra-modular hubs [49,50]. Inter-module hubs bridge different modules or clusters (i.e. terms that link the context of two topics). On the other hand, intra-module hubs have high connectivity to the members in a module (i.e. terms that are highly interconnected and conform strongly clustered groups).

4.6. Unsupervised semantic analysis

The Gensim implementation of the Latent Dirichlet allocation (LDA) method was applied to discover latent topics in the tweets [51]. LDA is a three-level hierarchical Bayesian model, in which each document is modelled as a finite mixture over

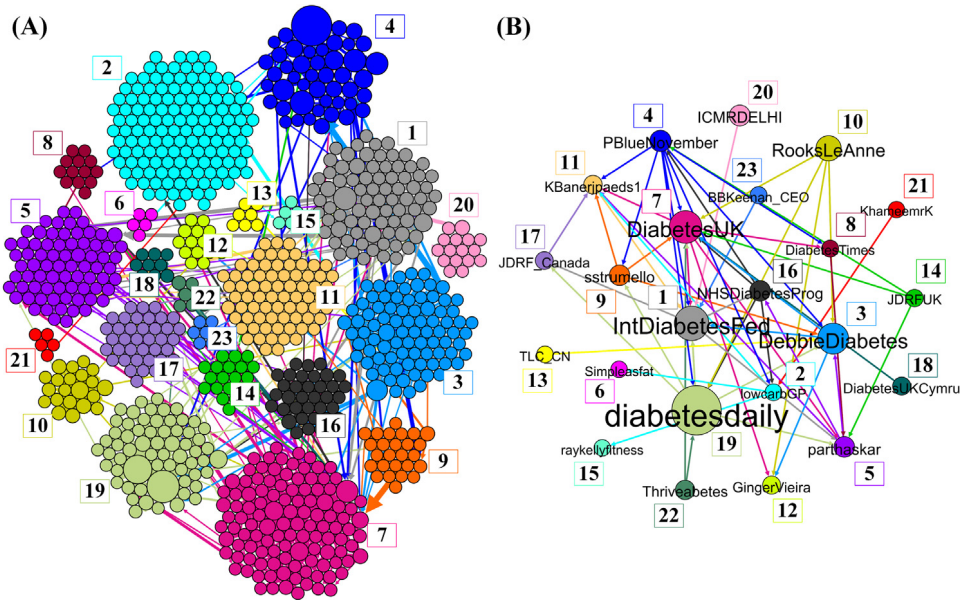


Fig. 2. (A) Graph showing the detected communities based on their Eigenvector Centrality metric but only considering the users inside each community (i.e. intra-community relations). (B) Graph showing the most influential users considering the users in the whole diabetes community (intra- and inter-community relations). The vertex colour represents the community whilst the number is the associated identifier, the vertex size is based on the eigenvector metric, whilst the edge colour stands for the colour of the community that made the retweet and the edge size stands for the number of retweets between a pair of users.

an underlying set of topics. In turn, each topic is modelled as an infinite mixture over an underlying set of topic probabilities. Therefore, these probabilities provide an explicit representation of the document. LDA models have two hyperparameters α and β , used to tune the document-topic distribution and the topic-word distribution, respectively. Therefore, a low value of α places more weight on having documents with a few but dominant topics. Similarly, a low value of β places more weight on having each topic composed of only a few dominant grams (i.e. uni-, bi- and tri-grams).

The LDAvis method supported the interpretation of the obtained topic models [52]. Specifically, each topic is described based on the saliency and relevance of the grams [52,53]. The saliency of a gram, w , is defined as:

$$\text{saliency}(w) = P(w) \times \text{distinctiveness}(w) \quad (19)$$

where $\text{distinctiveness}(w)$ is defined as:

$$\text{distinctiveness}(w) = \sum_T P(k|w) \log \left(\frac{P(k|w)}{P(k)} \right) \quad (20)$$

where k is a latent topic, $P(k|w)$ represents the conditional probability for a given gram, w , $P(k)$ stands for the marginal probability and the $\text{distinctiveness}(w)$ is defined as a Kullback–Leibler divergence [54] between $P(k|w)$ and $P(k)$. The relevance of a gram w to a topic k was given by a weight parameter λ , where $0 \leq \lambda \leq 1$, as:

$$\text{relevance}(w, k|\lambda) = \lambda \times \log(\phi_{kw}) + (1 - \lambda) \log \left(\frac{\phi_{kw}}{p(w)} \right) \quad (21)$$

where λ determines the weight given to the probability of gram w under topic k relative to its lift (i.e. how prevalent a term is across all topics). That is, $\lambda = 1$ ranks words in decreasing order of their topic-specific probability, and $\lambda = 0$ ranks words based only on their lift.

Finally, to improve the topic coherence of the model only nouns and proper names were considered in this analysis [55]. Moreover, a grid search approximation was applied to find the optimal hyperparameters of the LDA models. The objective was to maximise both model perplexity and the topic coherence

score [56]. Perplexity is a way to calculate the likelihood and it is defined as the reciprocal geometric mean of the token likelihoods in the test corpus given de model. So, the perplexity is defined as:

$$p(\vec{w}|M) = \exp - \frac{\sum_{m=1}^M \log p(\vec{w}_m|M)}{\sum_{m=1}^M N_m} \quad (22)$$

where N_m represents the length of the text, M is the trained model, and \vec{w}_m represents the word vector in document m . Lower values of perplexity indicate lower misrepresentation of the words of the test documents by the trained topics.

The topic coherence computes the sum of pairwise scores on the grams w_1, \dots, w_n used to describe the topic, usually the top n grams by frequency $p(w|k)$. This measure can be seen as the sum of all edges on the complete graph and it is defined as:

$$\text{coherence} = \sum_{i < j} \text{score}_{UCI}(w_i, w_j) \quad (23)$$

The score_{UCI} uses a pairwise score function, the Pointwise Mutual Information (PMI), and it is defined as:

$$\text{score}_{UCI}(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \quad (24)$$

where $p(w)$ represents the probability of the gram, w_i , occurring in a random document, whereas $p(w_i, w_j)$ gives the probability both grams w_i and w_j co-occurring in a random document.

4.7. Domain-specific graph reconstruction

The following domain-related ontologies and dictionaries were applied to recognise and extract semantic concepts from the tweets: the Diabetes Mellitus Treatment Ontology (DMTO) [57], the Disease Ontology (DO) [58], the FoodOn ontology [59], the Uber Anatomy Ontology (UBERON) [60] and the lexicon of Drug-Bank [61]. Overall, a lexicon of 89,919 entries supported the entity recognition task. Concepts were semantically grouped into the categories “Disease”, “Food & Nutrition”, “Anatomy”, “Drug & Chemical compounds”, “Symptoms” and “Physical activity”.

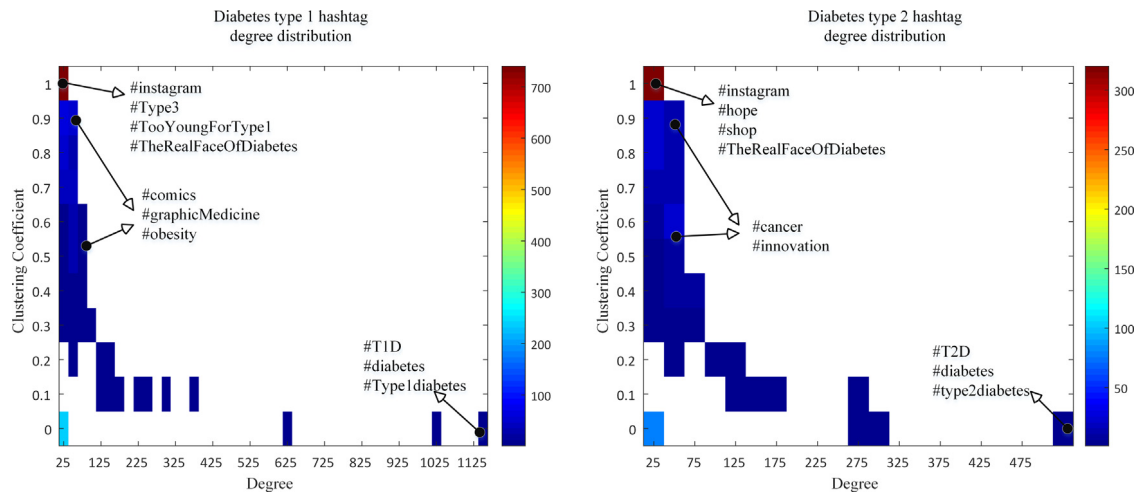


Fig. 3. Two-dimensional distribution of T1D and T2D hashtags. The most frequently used hashtags fall into the lower right region of the plots whereas the more clustered hashtags fall into the upper left region of the plots.

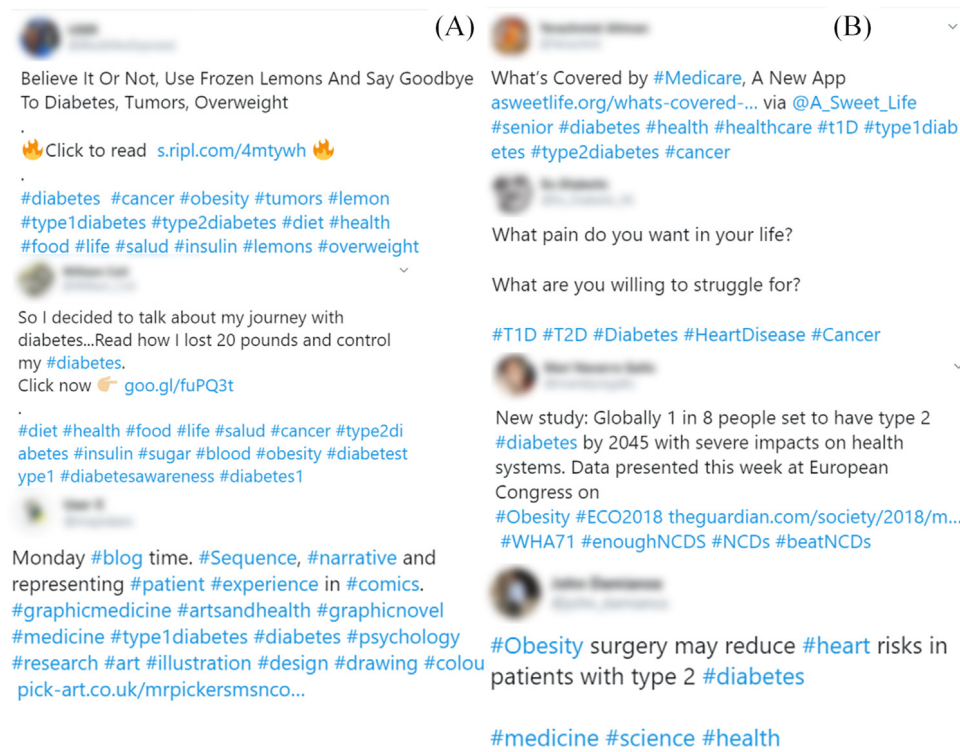


Fig. 4. (A) Example of excessive usage of hashtags in conversation to generate a flood of messages about miracle remedies. (B) Example of the usage of one identical hashtag to join different themes.

To recognise these concepts along with their semantic categories, an in-house NER workflow was applied. This NER entailed dictionary lookup, as well as pattern and rule-based recognition. To improve the efficiency of the NER, an inverted recognition technique was used [62]. The idea behind this implementation is to use the terms in the text as patterns to be matched against the lexicon. This approximation suits optimal for the type of texts analysed in this study due to their short length compared to the size of the lexicon. Moreover, recognition preference was given to the longest possible n-grams (e.g. systemic lupus erythematosus instead of only lupus) and concepts that may be associated with

more than one semantic category were ignored. Additionally, the recogniser accepted perfect matches as well as lexical variations of the terms (i.e. lemmatised entries and abbreviations).

The co-occurrence of the semantic concepts recognised in tweets was quantified by the coefficient of association for binary variables [63], ϕ , defined as:

$$\phi_{c_i c_j} = \frac{A_{c_i n c_j} A_{c_i' n c_j'} - A_{c_i' n c_j} A_{c_i n c_j'}}{\sqrt{A_{c_i} A_{c_i'} A_{c_j} A_{c_j'}}} \quad (25)$$

where A_{c_i} represents the number of tweets containing the concept c_i , $A_{c_i'}$ stands for the number of tweets not containing the

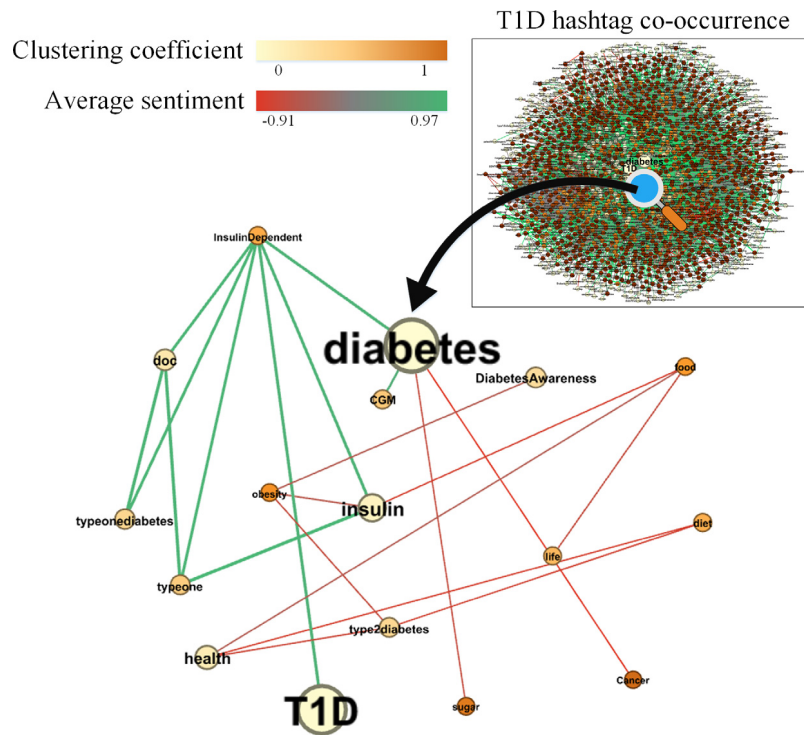


Fig. 5. The graph of T1D hashtag occurrences. The vertex colour denotes the clustering coefficient of the hashtag (i.e. the darker the vertex, the higher the coefficient), the vertex size represents the degree of the hashtag, the edge colour depicts the average sentiment of the including tweets (i.e. the redder the edge, the more negative the feeling) and the edge size stands for the total users.

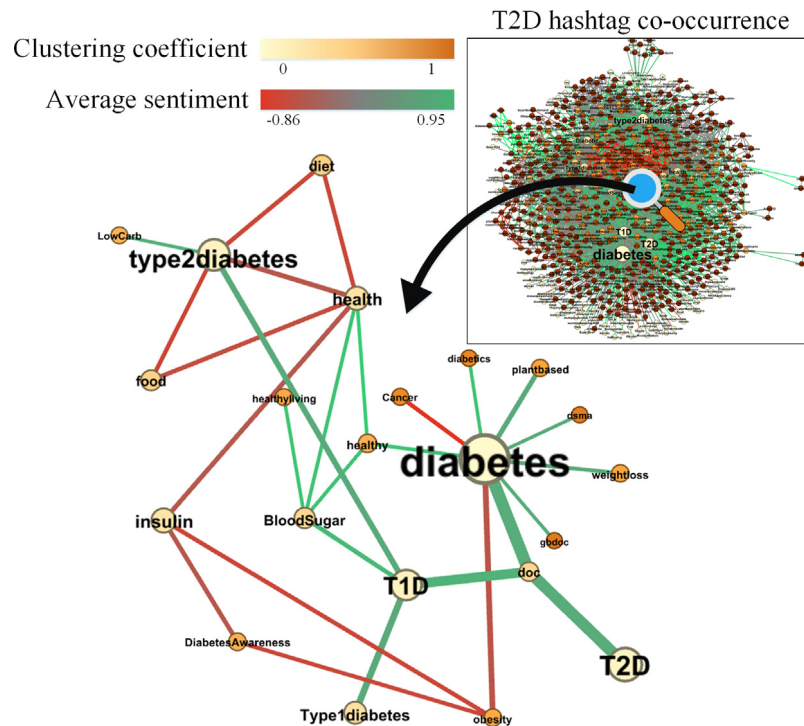


Fig. 6. The graph of T2D hashtag occurrences. The vertex colour denotes the clustering coefficient (i.e. the darker the vertex, the higher the coefficient), the vertex size represents the degree, the edge colour is based on the average sentiment of the including tweets (i.e. the redder the edge, the more negative the feeling) and the edge size stands for the total users.

term c_i , $A_{c_i \cap c_j}$ indicates the number of tweets containing both terms c_i and c_j , $A_{c_i \cap c_j'}$ indicates the number of tweets not containing both terms c_i and c_j , and $A_{c_i \cap c_j}$ represents the number of tweets containing the term c_i but not term c_j . In this context, the

ϕ coefficient ranges between -1 to $+1$ representing the extent to which tweets tend to discuss one topic but not the other, none of the topics or both topics together.

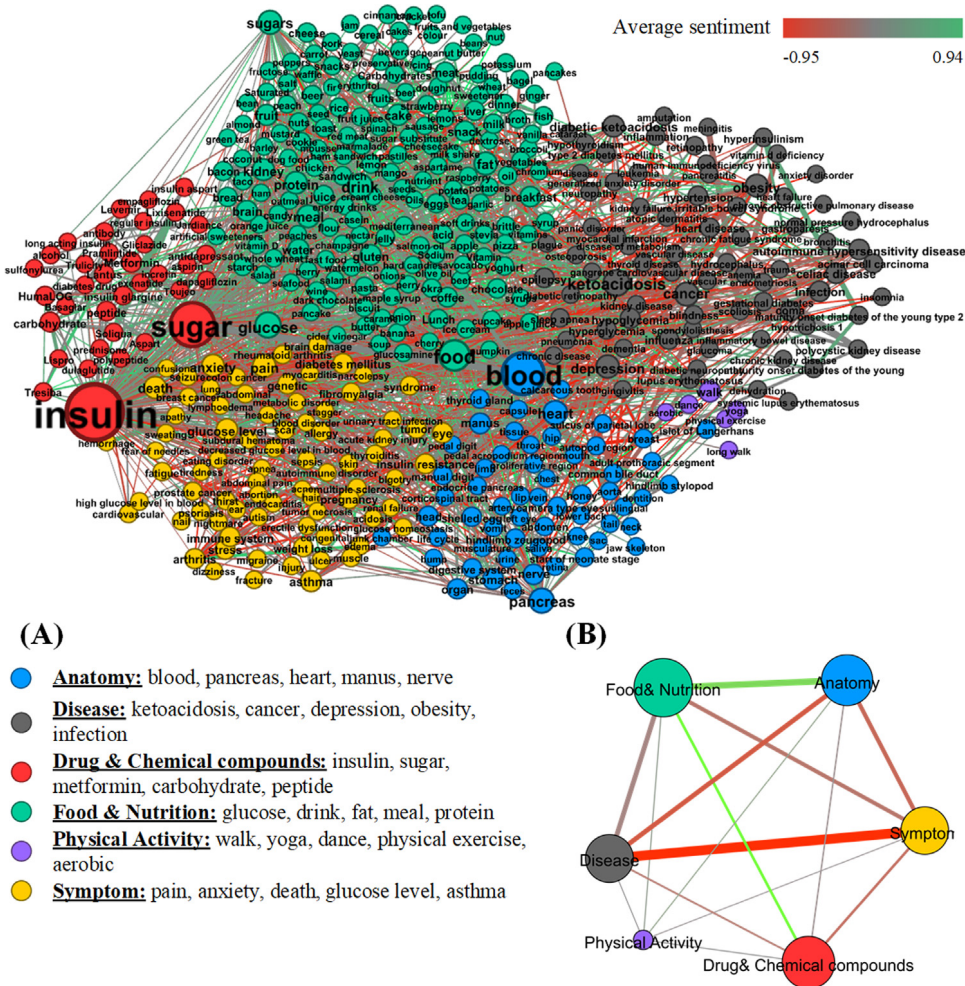


Fig. 7. Graph depicting the occurrence of semantic concepts in T1D patient tweets. Rendering is based on the Circle Pack layout. The vertex colour denotes the semantic category, the vertex size represents the degree, the edge colour is based on the sentiment of the majority of the associated tweets and the edge size stands for the total interactions between the terms. (A) Concepts with the highest degree by semantic category. (B) Network obtained by collapsing each semantic category in a single node with edges weighted by the sum of the coefficient of association among distinct categories and coloured by the sum of the sentiments among different categories.

5. Results and discussion

5.1. Identifying and analysing diabetes communities

The general characterisation of the communities talking about diabetes (independently of the type of user) was deemed relevant to gain a better understanding about the users contributing to the domain and, implicitly, to look into the social positioning of patients. That is, identifying those with whom patients relate the most one can understand their main motivations to participate in this social community. This graph contained 899 vertexes (i.e. unique users) and 1265 edges (i.e. retweets). The edges were weighted based on the number of retweets between a source and a target user.

The GLayer algorithm identified 23 distinct communities with a modularity value of 0.83. The average size of each community was 39 users with a maximum size of 138 users and a minimum size of 5 users. The importance of the detected communities and the influence of the different users was evaluated based on the values of degree, the eigenvector and the betweenness characterising the connectedness of the communities. Table 1 shows the 5 communities with most retweets and the centrality metrics of the most influential user in each community. Supplementary Material 1 contains a full description of the communities.

One can observe that the communities with most retweets included a considerable amount of users and interactions among them. The most influential users inside each community were very relevant (i.e. high degree centrality) and possessed the control of the information flow (i.e. high betweenness centrality). These centralities are corroborated when exploring the profiles of these users, namely: @grumpy_pumper is a well-known blogger and sufferer of T1D, @shashiiyengar is an influencer focused on the use of ketogenic diets in T2D, @kidfears99 is a well-known campaigner and a sufferer of T1D and, @diabetesdaily and @Int-DiabetesFer are well-recognised organisations related to disease control.

Fig. 2 illustrates the 23 communities and the interactions among them. Particularly, the size of the vertexes in Fig. 2A is based on the calculation of the eigenvector centrality metric while considering only the users within each community (e.g. the community 1 only considers the relationships among the vertexes of this community). Conversely, Fig. 2B represents the most influential users of each community taking into account all the diabetes community (e.g. the community 1 considers the relationships among the vertexes inside the community 1 and with the rest of the communities if exists). Considering both perspectives, it is possible to observe that the most influential user and the one who controls the flow of information inside

(A)

Semantic categories	Anatomy	Disease	Drug & Chemical compounds	Food & Nutrition	Physical Activity	Symptom
Anatomy	0.0572	0.0670	0.0438	0.0564	0.0277	0.0579
Disease	0.0670	0.1032	0.0489	0.0643	0.0072	0.0987
Drug & Chemical compounds	0.0438	0.0489	0.1158	0.0230	0.0619	0.0583
Food & Nutrition	0.0564	0.0643	0.0230	0.0910	0.0258	0.0658
Physical Activity	0.0277	0.0072	0.0619	0.0258	0.0000	0.0008
Symptom	0.0579	0.0987	0.0583	0.0658	0.0008	0.0901

(B)

Total interactions \ Positive Interactions							TOTAL
	Anatomy	Disease	Drug & Chemical compounds	Food & Nutrition	Physical Activity	Symptom	
Anatomy	98 128	81 99	34 80	133 154	9 11	81 105	436 577
Disease	81 99	148 152	42 80	84 98	1 1	121 129	477 559
Drug & Chemical compounds	34 80	42 80	84 114	143 208	3 7	53 92	359 581
Food & Nutrition	133 154	84 98	143 208	558 570	8 12	65 73	991 1115
Physical Activity	9 11	1 1	3 7	8 12	0 0	1 3	22 34
Symptom	81 105	121 129	53 92	65 73	1 3	112 114	433 516

Fig. 8. Adjacency heatmap matrix of the T1D knowledge graph. (A) Depicts the average coefficient of association among the different semantic categories. (B) Depicts the number of interactions with a positive coefficient of association (upper section of the cell) and the total number of interactions (bottom section of the cell) among the different semantic categories.

each community is not necessarily the same. For example, for community 7, the most influential user inside the community was @grumpy_pimper whereas @DiabetesUK was the most influential user of community 7 in the other communities. That is, information exchange within the community is more likely to be focused on specific topics, whereas inter-community information flows are driven by known organisations and individuals. Therefore, this behaviour denotes that specific expertise may be hidden for someone who does not belong to the community.

In terms of the most influential users in the diabetes community, it was noticed a wide variety of roles (Fig. 2B). Half of these user accounts (12 out of 23 users) correspond to top health organisations related to the disease, such as @IntDiabetesFed, @PBlueNovember, @DiabetesUK or @diabetesdaily. There were also health and industry professionals (7 out of 23 users), such as doctors, nutritionists and sports “experts” (i.e. @lowcarbGP, @parthaskar or @RooksLeAnne) and, in small number, bloggers and influencers (4 out of 23 users), like @DebbieDiabetes or @sstrumello. This variety of influential users is in concordance with the assumption that laymen people (including patients and relatives) often turn to knowledgeable users for help.

5.2. Diabetes topical co-attention via hashtags

Lexical diversity is an interesting concept in the area of interpersonal communications because it provides a quantitative measure of the diversity of an individual’s or group’s vocabulary. The analysis of the hashtags used in the tweets (and their combination) is useful to know the topics that the participants of the community have decided to emphasise. In this sense, the topological analysis of T1D and T2D hashtag graphs showed that the degree (i.e. the number of co-occurrences of each concept), is, generally, inversely proportional to the clustering coefficient

(i.e. quantifies how much a concept is grouped or interconnected with its neighbours), that is when the degree is higher, the clustering coefficient is lower, and vice versa (Fig. 3).

Besides, the behaviour of both graphs was assortative, i.e. higher degree hashtags were adjacent to one another (forming a few, large clusters) whereas low degree hashtags were adjacent to other low degree hashtags (forming many small clusters). This behaviour suggests that the degree determines the clustering coefficient, i.e. there are a set of popular hashtags (located at the bottom-right side of the plots) used in combination with other popular hashtags (e.g. #T1D and #diabetes or #T2D and #insulin), whereas more specific hashtags (located at the top-left side of the plots) are usually combined among themselves (e.g. #Instagram and #shop or #TooYoungForType1 and #TheRealFaceOfDiabetes).

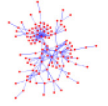
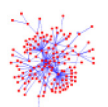
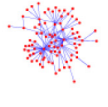
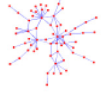
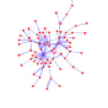
Noteworthy, hashtags with an intermediate value of degree and clustering coefficient (located in the middle part of the plots) were used to connect highly interconnected hashtags. This tendency is mainly due to: (i) the indiscriminate use of hashtags in conversations, usually done by spammers; and (ii) the use of hashtags as wildcards to join different popular topics. Fig. 4A shows two examples of the excessive usage of hashtags in a conversation to generate a flood of messages about miracle remedies. Likewise, Fig. 4B shows examples of users using the same hashtag (i.e. #cancer or #obesity) in combination with other of distinct natures (i.e. #Medicare, #HeartDisease, #NCDs or #Science) to post messages about different themes.

The analysis of the co-occurrence of T1D hashtags and T2D hashtags combined with the sentiment associated with the including tweets gave additional insight into the information flows. Figs. 5 and 6 represent the most mentioned hashtags pairs in patients’ messages with strong positive or negative sentiment about T1D and T2D, respectively.

Some top used hashtags are common for the two types of diabetes (e.g. “#sugar”, “#cancer”, “#diet” and “#food”), but some

Table 1

The 5 user communities with more retweets. The communities are sorted by the number of retweets (descendant order). The centrality metrics (degree, eigenvector and betweenness) relate to the most influential user of the community (i.e. top user).

Cluster	ID	N. Vertex	N. Edges	N. Retweets	Top User	Degree	Eigen.	Betwe.	Most frequently retweet
	7	114	158	3,087	Grumpy_pumper	366	0.47	2,648	It's #WorldDiabetesDay! Today, millions of people come together to raise awareness of diabetes. This year we're celebrating your support network. Your #DiabetesFamily. Maybe that's a friend, or a neighbour.
									Your mum, or best mate. Who's in yours? Share and celebrate them today: https://t.co/HMb6HDftb
	2	138	192	2,999	shashiyengar	224	0.45	1,405	Please RT: KETO 2y results on 350 ppl w/ T2 diabetes: 53.5% reversed diabetes. Repeat: diabetics NO LONGER have diabetes Attn! @AmDiabetesAssn ^{POI} Chair David Herrick @EthicOneLLC ^{POI} & CEO Tracey Brown @Type2CEO, When will yr guidelines be updated? Link: https://t.co/QU4U4h2XME https://t.co/cKxulP46pDP
	3	85	121	2,005	Kidfears99	695	0.56	3,648	Got my keyboard keys changing color *when my blood sugar goes up!* @daskeyboard @NightsoutProj #WeAreNotWaiting #diabetes https://t.co/DSBDeO7RE
	19	62	75	1,953	diabetesdaily	450	0.70	1,205	Colorado just became the first state to cap #insulin co-pays at \$100 per month. This is awesome! Huge thanks to our Advocacy team, #Diabetes Advocates, Governor Jared Polis, Representative Dylan Roberts, and Senators Kevin Priola and Kerry Donovan! https://t.co/plyNYdZRhm
	1	79	100	1,862	IntDiabetesFed	905	0.57	4,038	Today is World Diabetes Day! Did you know that it is celebrated on November 14 because it marks the birthday of Sir Frederick Banting, who co-discovered insulin in 1921? Almost 100 years since the discovery, insulin remains out of reach for many #WDD2018 https://t.co/seJvEFp9xR https://t.co/lxAnoe9W9d

interesting differences are also noted. For example, some of the most important topics of the T1D community were “#insulinDependent” and “#cgm”, whereas “#weightloss” and “#lowcarb” were at the top of the T2D conversations. T1D patients seemed to be more interested in daily issues (reflecting a positive sentiment), such as their dependence on insulin pumps and continuous glucose monitors (CGM). On the other hand, T2D patients were more positively interested in discussing overweight and diets, such as the low-carb diet. In terms of hashtags associated with a negative sentiment, both communities mentioned “#cancer”, “#food” and “#sugar”, and in T2D their combination accentuated the negative incidence of the hashtag #diet.

Finally, it is interesting to analyse the negative incidence of the popular hashtag “#DiabetesAwareness” (i.e. an international

campaign that draws attention to the multiple disease complications) and its usage in combination with other hashtags. In this sense, the most related hashtag used by the T1D patients was “#obesity” whereas in the case of T2D patients it was “#insulin”.

5.3. Topics in patient information flows

Due to the diverse vocabulary used in social media conversations, an unsupervised approach to topic modelling may help to identify the topic distribution in general conversations prior to capturing specific terminology. In this sense, out of all the tweets related to T1D and T2D (27,128 and the 15,112 tweets respectively), the implemented LDA models identified four optimal

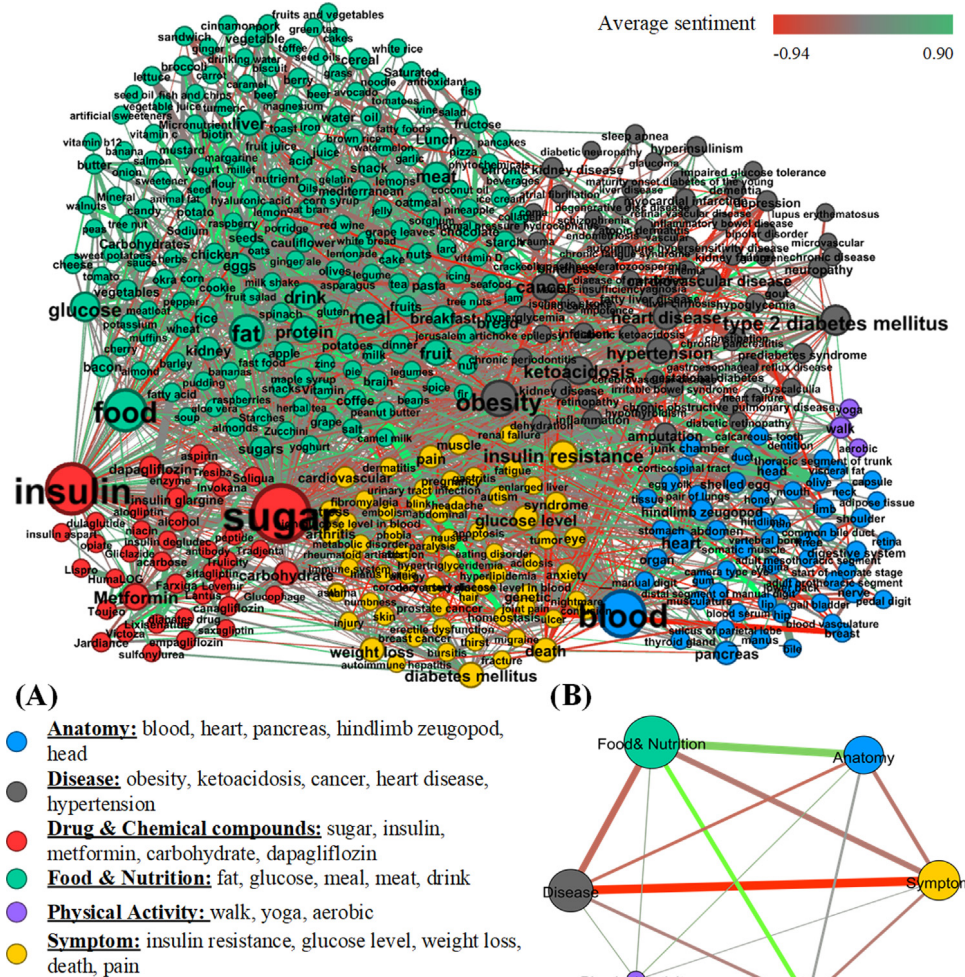


Fig. 9. Term co-occurrence graph for T2D patient tweets using the Circle Pack layout. The vertex colour denotes the semantic categories, the vertex size represents the degree, the edge colour is based on the sentiment of the majority of the associated tweets and the edge size stands for the total interactions between the terms. (A) Concepts with the highest degree by semantic category. (B) Network obtained by collapsing each semantic category in a single node with edges weighted by the sum of the coefficient of association among distinct categories and coloured by the sum of the sentiments among different categories.

topics. Every one of the topics was manually labelled based on its most relevant terms. For example, in the case of the T1D topics, the “Topic 1” was labelled “Pregnancy”, because the terms that most contributed to the topic were “Pregnancy”, “Gestational”, “Weight”, “Foetus” and “Control”. Tables 2 and 3 describe the topics discovered for T1D and T2D in terms of their five most weighted terms and an exemplifying tweet.

The results obtained for both types of diabetes are quite interesting. The patients of T1D were most concerned about the control of their blood sugar during the pregnancy, due to the change of the hormones, and about the possibility of harming the baby [64]. Another topic under discussion was the deregulation of insulin prices. There were a high number of users urging the pharmaceutical companies, and even the president of their country, to regulate the prices and make them affordable to every person in need [65]. Another minor topic of interest was events, tips and hashtags about diabetes, which gathered miscellaneous tweets about patient experiences, events announced by diabetes organisations, and humour posts related to the disease.

The patients of T2D focused their conversations on the importance of doing exercise to lower the blood sugar level (and increment insulin sensitivity) and control their weight [66]. The need to maintain a healthy, balanced diet was another important topic of discussion, in particular, the importance of eating foods

with low carbohydrates (i.e. recommendation of keto diets) or with a high content of antioxidants and vitamins (i.e. ingestion of superfoods) [67]. To a lesser extent, T2D patients also discussed the usage of certain drugs and supplements to help lower the blood sugar level or to cover any deficiency affecting the production of insulin. For example, the consumption of vitamin D to improve the function of pancreatic cells [68].

Finally, a topic related to foods was identified in both T1D and T2D corpora. T1D conversations focused on the discussion of the im/possibility of ingesting foods that may affect the sugar levels (e.g. chocolates, fruits or cakes) due to the incapacity to produce insulin naturally. In turn, the patients of T2D were concerned about foods that may increase the levels of blood pressure and cause hypertension, another disease closely related to this type of diabetes [69].

5.4. Medical domain knowledge in patient communication

Besides gaining a deeper understanding of the conversations, the recognition of semantic concepts related to diabetes and further study of their interrelation in patient conversations can help health-related stakeholders in the development of better (more focused) information strategies to educate the public and prevent the spread of misinformation.

Table 2
Topics discovered for the T1D corpus.

Topic ID	Topic label	Top words	% of words	Example tweet
1	Pregnancy	Pregnancy, gestational, weight, foetus, control	34.6	Children of mothers who had gestational diabetes during pregnancy could be at increased risk of type 1 diabetes themselves, according to @mcgillu researchers https://t.co/f5mUHManbR
2	Foods & sugar levels	Sugar, blood, fruit, chocolate, cake	27.1	My son is a type 1 diabetic and I recently purchased the protein mix as well as the banana paleo mix because he LOVES pancakes but they are always a problem for his blood sugar levels. I cook pancakes twice a week... https://t.co/0D9Z4dblyB
3	Medication prices	Prices, insulin, fixing, presidents, deregulation	26.1	@realDonaldTrump How about helping us with 1200 percent price increase of insulin since 1996. You trash all the other presidents, well help all our type 1 diabetic children as we go broke paying in average 275 for a vial of insulin. #insulin4all
4	Events, Tips & Hashtags	Diabetesproblems, program, diabetesawareness, diabadass, gojuicego	12.3	So the #problem with #diabetes comes when the #insulin is not utilised by our muscles and fat tissues to take in the glucose... Conclusion- #checkyourdiet! #diabetesawareness #diabetesresearcher #diabeteslife #diabetesdiet #healthydiet #research #stayup...

Table 3
Topics discovered for the T2D corpus.

Topic ID	Topic label	Top words	% of tokens	Example tweet
1	Foods & Tension	Food, hypertension, diet, pressure, obesity	37.9	A vegetarian diet can reduce the risk of death from cardiovascular disease by 40% and hypertension by 34%. It can reduce the risk of metabolic syndrome and type 2 diabetes, and help in weight management.
2	Exercises & Glucose regulation	Glucose, exercise, walk, burn, gym	24	I just finished my 5 mile run & walk this morning. My daily exercise & proper diet helps me keep my blood sugar level in the normal range. I have been a type 2 diabetic now for 14 yrs & I am still doing well. Remember to take care of your body the temple of God. @phylisbowyer https://t.co/ZaxBCujRdv
3	Diet & Remedies	Beta cell, superfood, ketoacidosis, keto, gestational	22.3	@DrSarahHallberg @cnnhealth My dad is a 10yr T2D. I asked his diabetes dr if a ketogenic diet will help. The dr said no because he had been diabetic too long and his pancreas was no longer able to create beta cells so dietetic intervention would not work for him.
4	Drugs & Supplements	Afrezza, fibre, vitamin, supplement, reduce	15.8	Are you on diabetes medications & not at your A1C goal? Adding Afrezza inhaled insulin may help. Ask your healthcare provider at your next appt and visit https://t.co/qoxvnmZyBE to learn more. #diabetes #insulin #t2d #t1d #inhaledinsulin (See Safety Info...)

In both T1D and T2D knowledge graphs, the vertexes with the highest degree in the graph are hubs, analogously they are expectedly domain stopwords (e.g. Insulin or glucose). Fig. 7 illustrates the T1D knowledge graph that is composed of a total of 431 vertexes (i.e. unique concepts) and 1658 edges. The rendering of the graph is based on the Circle Pack layout [70] to ensure the vertexes of the same semantic category stick together. The colour of the vertexes represents the semantic category of the concept, i.e. red stands for drugs and chemical compounds, grey represents diseases, green relates to food and nutrition, blue stands for anatomy, yellow represents symptoms and purple represents physical activity. The edge size was calculated based on the coefficient of association (i.e. thicker edges represent a stronger association between the vertex), whilst colour stands for the sentiment associated, i.e. green indicates a positive sentiment, grey represents neutral sentiment and red stands for negative sentiment. Fig. 7A depicts the top five concepts per semantic category (i.e. concepts with the highest degree). Noteworthy, common domain concepts are not considered here (e.g. “food” in “Food & Nutrition”). In turn, Fig. 7B shows inter-category relationships. This network was obtained by collapsing each semantic category into a single node with edges weighted by the sum of the coefficient of association among distinct categories and coloured by the sum of the sentiments among the different categories.

“Food & Nutrition” and “Symptom” were the semantic categories with the highest number of unique concepts (i.e. 173 and 78, respectively). Likewise, “Food & Nutrition” and “Drug & chemical compounds” were the semantic categories with the highest number of occurrences in the corpus (i.e. 1115 and 581, respectively). Moreover, it was possible to detect 331 relations with a positive sentiment and 485 relations expressing negative

sentiments. The prevalence of negative messages was somewhat expected since patients usually use social platforms to vent out their emotions and frustrations regarding diseases that have a challenging treatment, such as cancer, and do not have a cure, such as diabetes [71,72].

Fig. 8 shows the adjacency matrix and discusses the association among the semantic categories of the T1D knowledge graph (Fig. 7B) from a different point of view. Notably, Fig. 8A depicts the average coefficient of association among the different semantic categories, whereas Fig. 8B depicts the number of positive associations (i.e. co-occurrences with a coefficient of association bigger than 0) and the total number of associations among each semantic category.

Ignoring the interactions among the terms of the same category (e.g. “Disease” and “Disease”), Fig. 8A shows that “Disease” and “Symptom” were the semantic categories with the highest coefficient of association between them (i.e. a value of 0.09), whereas Fig. 8B shows that “Drug & Chemical compounds” and “Food & Nutrition” were the categories with the highest number of interactions (i.e. 143 positive interactions and 208 total interactions).

The T2D knowledge graph (Fig. 9) was composed of a total of 429 vertexes and 1723 edges. “Food & Nutrition” and “Disease” were the semantic categories with the highest number of unique concepts (i.e. 192 and 73, respectively). These semantic categories also had the highest number of occurrences in the corpus (i.e. 1405 and 604, respectively). Moreover, the number of negative relations almost doubled the number of positive relations, i.e. 286 positive relations and 425 negative relations.

Fig. 10 illustrates the adjacency matrix and discusses the association among the semantic categories of the T2D knowledge

(A)

Semantic categories	Anatomy	Disease	Drug & Chemical compounds	Food & Nutrition	Physical Activity	Symptom
Anatomy	0.1040	0.0584	0.0742	0.0771	0.0451	0.0858
Disease	0.0584	0.0785	0.0590	0.0678	0.0607	0.0914
Drug & Chemical compounds	0.0742	0.0590	0.1511	0.0363	0.0893	0.0649
Food & Nutrition	0.0771	0.0678	0.0363	0.1263	0.1152	0.0773
Physical Activity	0.0451	0.0607	0.0893	0.1152	0.0000	0.0115
Symptom	0.0858	0.0914	0.0649	0.0773	0.0115	0.1249

(B)

Total interactions	Positive Interactions						TOTAL
	Anatomy	Disease	Drug & Chemical compounds	Food & Nutrition	Physical Activity	Symptom	
Anatomy	78	55	34	119	7	51	344
Disease	92	66	63	145	7	65	438
Drug & Chemical compounds	66	182	91	139	4	122	604
Food & Nutrition	63	34	63	120	145	2	415
Physical Activity	63	91	138	191	4	79	566
Symptom	119	112	145	786	3	85	1250
	145	139	191	820	5	105	1405
	7	3	2	3	0	4	19
	7	4	4	5	0	5	25
	51	111	51	85	4	78	380
	65	122	79	105	5	82	458

Fig. 10. Adjacency heatmap matrix of the T2D knowledge graph. (A) Depicts the average coefficient of association among the different semantic categories. (B) Depicts the number of interactions with a positive coefficient of association (upper section of the cell) and the total number of interactions (bottom section of the cell) among the different semantic categories.

graph (Fig. 9B). In particular, Fig. 10A shows that “Anatomy” and “Food & Nutrition” were the semantic categories with the highest coefficient of association (i.e. with a value of 0.12), whereas Fig. 10B shows that “Drug & chemical compounds” and “Food & Nutrition” were the semantic categories with the highest number of interactions (i.e. 145 positive interactions and 191 total interactions).

When comparing these co-occurrence graphs, one may observe that the 4 concepts with the highest degree (i.e. more appearances in tweets) were similar and quite broad, i.e. “insulin”, “sugar”, “blood” and “food”. However, it was possible to observe more interesting differences by looking into concepts with a moderate presence in the conversations. For example, in the T1D graph, other concepts with a moderate degree were “ketoacidosis” and “drink”, whereas in the T2D were “obesity” and “fat”. When considering the coefficient of association of “ketoacidosis” in the T1D graph, it is possible to observe that its highest coefficient of association is linked to “death”, with an obvious negative connotation. This can be explained by the fact that the T1D community promotes self-awareness about the dangers of ketogenic diets and ketoacidosis in diabetes type 1 (Fig. 11A). On the other hand, “drink” had a high coefficient of association with “water” (positive association) that, in turn, was many times mentioned in association with “insulin” and “blood”. The rationale here is that patients are becoming aware of the importance of drinking water to regulate blood sugar levels (Fig. 11B). However, although water intake is recommended by several institutions such as the American Diabetes Association [73] and the WHO [74] to contribute to the reduction of the growing prevalence of type 2 diabetes and its pre-stages, the evidence for positive effects of water in improving glycemic parameters in diabetic and non-diabetic persons is low and the results are heterogeneous with no clear result [75].

Performing an analogous analysis over the T2D knowledge graph, one can observe that the highest number of associations were among “fat”, “food” and “obesity”. Noteworthy, the highest coefficient of associations in both graphs was related to “okra” and “cancer”. Many tweets discussed the multiple benefits of “okra” for treating obesity and diabetes, whereas other messages warned about the placebo effect of this plant (Fig. 12).

On the other hand, T1D and T2D conversations talking about “cancer” and “diabetes” were usually related to food and “miraculous” remedies (Fig. 13). A high number of messages that contain both terms could be considered spam messages (i.e. identical messages posted by different accounts with similar names or messages that belong to accounts that are no longer available at present or are currently suspended). In this line, the claims supporting that the lemons (and, more specifically, the Limonene compound) kill cancer cells are well recognised fake news without any reputable scientific or medical studies evidence. However, recent studies reference that some citrus fruits contain certain compounds that may potentially have anti-cancer properties that could help ward off some types of cancer [76].

Attending to the interactions and mentions of “Drug & Chemical compounds” in both graphs, the commercial drugs most commonly mentioned by T1D patients were “Humalog”, “Lantus” and “Metformin” whereas T2D patients mentioned “Metformin”, “Dapagliflozin” and “Jardiance”. In the case of T1D patients, there were many conversations about the abusive price of “Humalog” and “Lantus” (Fig. 14).

In the case of T2D, “Dapagliflozin” was many times discussed together with “Farxiga”, “Sitagliptin” and “Saxagliptin” in the context of results of new T2D therapies [77]. In the case of “Jardiance”, most of the attention related to news talking about the development of complications due to certain drug combinations, namely “Gangrene” and “Infection” (Fig. 15).

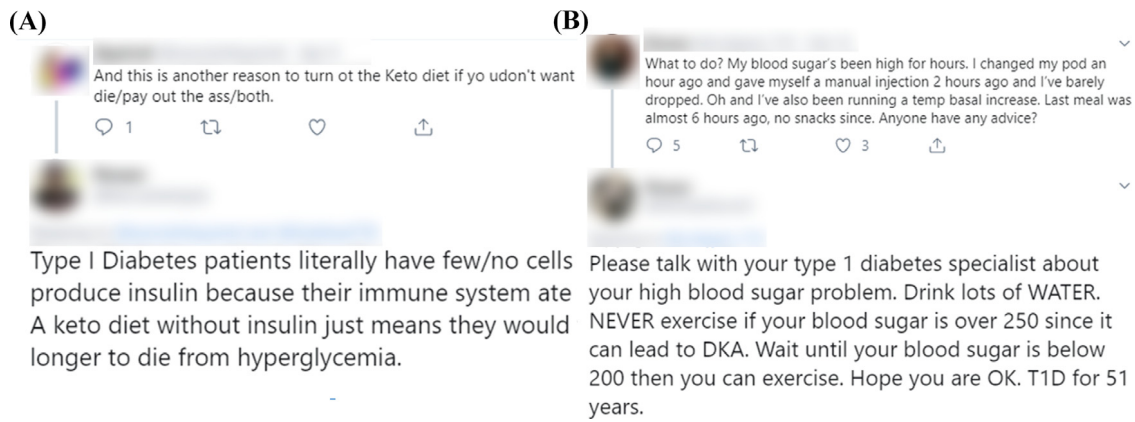


Fig. 11. Example of T1D patient tweets. (A) A conversation about the importance of the keto diet. (B) A conversation about the importance of drinking water.

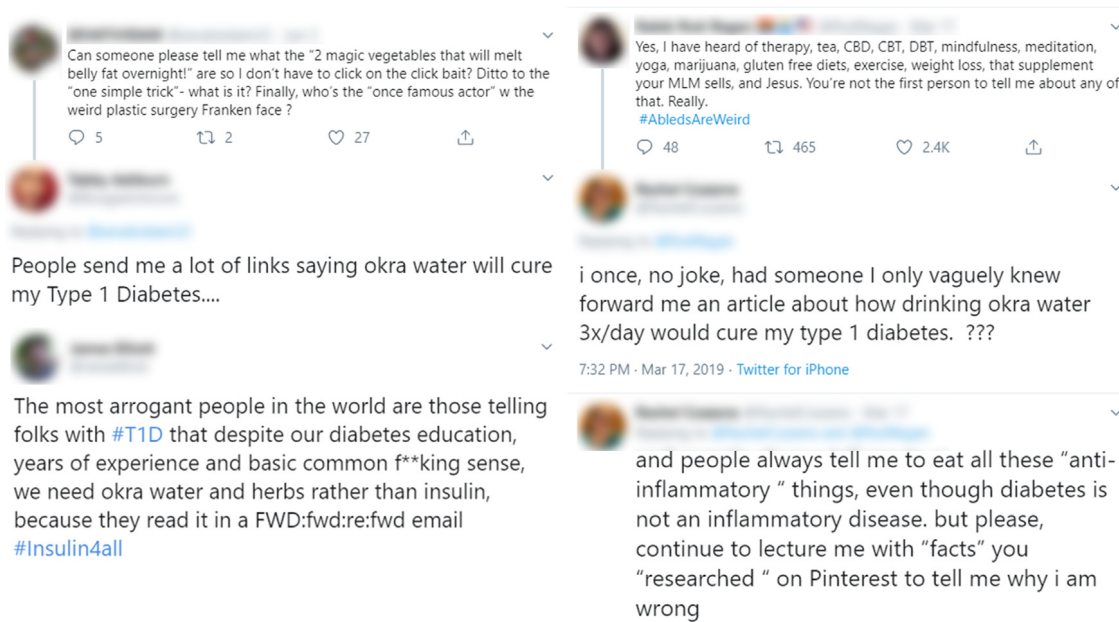


Fig. 12. Example of T1D tweets and T2D tweets related to okra.



Fig. 13. Example of T1D patient tweets and T2D patient tweets related to cancer and natural home remedies.

These posted messages are echoed of several recent studies that claim that sodium-glucose co-transporter 2-inhibitors like Jardiance or Ertugliflozin may produce gangrene and genital infection [78,79].

6. Conclusions

This work presents a new methodology focused on the study of social interactions to better understand the behaviour, perceptions and appreciations of patients and close relatives towards a

given health condition. The practical relevance of the proposed methodology is demonstrated in a study of almost two years of social conversations about diabetes on Twitter. This corpus comprised a total of 1.3 million tweets from 546,739 users.

Community detection techniques are applied to identify the most influential users in the diabetes community (namely, organisations, bloggers and influencers) and study the information exchange within the communities and among communities. Interestingly, intra-community discussions are topic-driven, whereas known organisations and individuals are at the centre

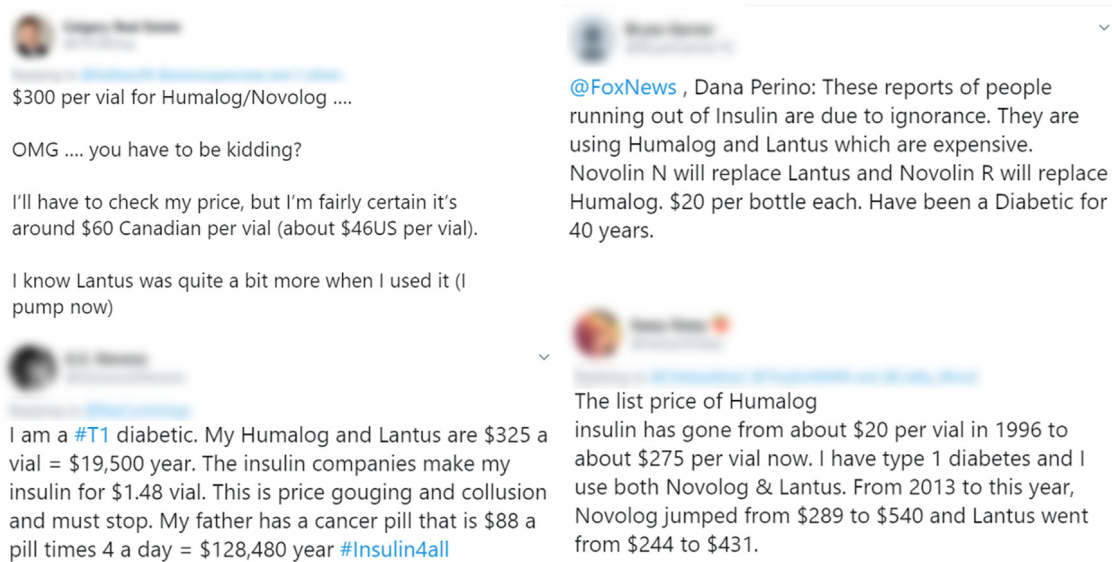


Fig. 14. Example of T1D patient tweets talking about drugs and their prices.

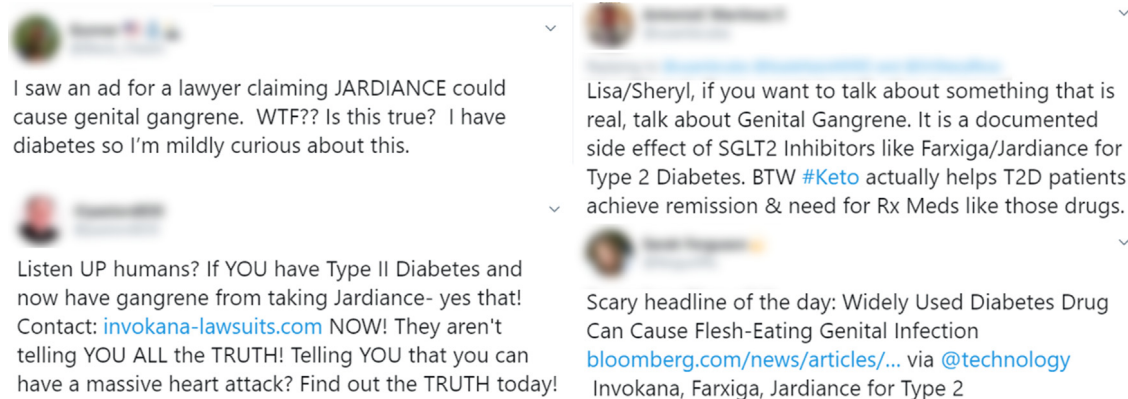


Fig. 15. Example of T2D tweets talking about new studies describing the side effects that certain treatments can cause.

of inter-community information flows. NLP and ML techniques are further used to gain a deeper understanding of the topics of conversation of patients and close relatives, namely in the context of T1D and T2D conditions. The reconstruction of knowledge graphs enables a holistic, multi-layered analysis of the acquired knowledge, looking into different levels of detail and perspective views.

This real-world case study exemplifies the broad range of non-trivial and practical knowledge that the proposed methodology can gather. This methodology can be applied by any interested stakeholder to gain insight into any other health-related topic and, most notably, it can be used to improve the capacity of stakeholders to promote healthy lifestyles and to communicate with the patients in a more efficient way, two main objectives in the fields of public and consumer health informatics.

Future work will be centred in improving the method capacities towards extending the perspectives of analysis. For instance, it is relevant to embrace other meaningful classes of users, such as industry, Health organisations and public institutions. This will pave the way to the analysis of additional and complementary information flows, namely targeting or originated by patients. Also, it will be considered the complementary application of network text analysis for classification purposes, such as the approach proposed by [80], to group users communities based on posted contents. Finally, the resources shared by the users, such as web

resources, will also be further explored, namely under the scope of current health research and development initiatives, health promotion actions, and other events that may be discussed and shared by the community.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Gael Pérez-Rodríguez: Methodology, Investigation, Software, Writing - original draft, Writing - review & editing. **Martín Pérez-Pérez:** Data curation, Visualization, Software, Writing - original draft, Writing - review & editing. **Florentino Fdez-Riverola:** Supervision, Validation, Writing - original draft, Writing - review & editing. **Anália Lourenço:** Conceptualization, Supervision, Validation, Writing - original draft, Writing - review & editing.

Acknowledgements

SING group thanks CITI (*Centro de Investigación, Transferencia e Innovación*) from the University of Vigo for hosting its IT infrastructure.

This work was partially supported by the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UID/BIO/04469/2013 unit, COMPETE 2020 (POCI-01-0145-FEDER-006684), the Xunta de Galicia (Centro singular de investigación de Galicia accreditation 2019–2022) and the European Union (European Regional Development Fund - ERDF)- Ref. ED431G2019/06, and Consellería de Educación, Universidades e Formación Profesional (Xunta de Galicia) under the scope of the strategic funding of ED431C2018/55-GRC Competitive Reference Group. The authors also acknowledge the Postdoc contract of Martín Pérez-Pérez, funded by the Xunta de Galicia.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.future.2020.04.025>. Supplementary material 1: this file contains the 23 user communities detected using the GLayer algorithm.

References

- [1] Statista, Number of social media users worldwide 2010–2021, 2019, <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/> (accessed 28 October 2019).
- [2] Statista, Twitter: number of active users 2010–2019, 2019, <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/> (accessed 28 October 2019).
- [3] M. Househ, E. Borycki, A. Kushniruk, Empowering patients through social media: The benefits and challenges, *Health Inform. J.* 20 (2014) 50–58, <http://dx.doi.org/10.1177/1460458213476969>.
- [4] S.A. Moorhead, D.E. Hazlett, L. Harrison, J.K. Carroll, A. Irwin, C. Hoving, A new dimension of health care: Systematic review of the uses, benefits, and limitations of social media for health communication, *J. Med. Internet Res.* 15 (2013) <http://dx.doi.org/10.2196/jmir.1933>.
- [5] Y. Pershad, P. Hangge, H. Albadawi, R. Oklu, Social medicine: Twitter in healthcare, *J. Clin. Med.* 7 (2018) 121, <http://dx.doi.org/10.3390/jcm7060121>.
- [6] S. Majidi, K.A. Driscoll, J.K. Raymond, Anxiety in children and adolescents with type 1 diabetes, *Curr. Diabetes Rep.* 15 (2015) 47, <http://dx.doi.org/10.1007/s11892-015-0619-0>.
- [7] R. Whittemore, R. Delvy, M.M. McCarthy, The experience of partners of adults with type 1 diabetes: an integrative review, *Curr. Diabetes Rep.* 18 (2018) 19, <http://dx.doi.org/10.1007/s11892-018-0986-4>.
- [8] E. Berry, M. Davies, M. Dempster, Managing type 2 diabetes as a couple: The influence of partners' beliefs on diabetes distress over time, *Diabetes Res. Clin. Pract.* 141 (2018) 244–255, <http://dx.doi.org/10.1016/j.diabres.2018.05.020>.
- [9] M.M. Franks, Z.S. Sahin, A.J. Seidel, C.G. Shields, S.K. Oates, C.J. Boushey, Table for two: diabetes distress and diet-related interactions of married patients with diabetes and their spouses, *Fam. Syst. Health* 30 (2012) 154–165, <http://dx.doi.org/10.1037/a0028614>.
- [10] L.J. Trump, J.R. Novak, J.R. Anderson, T.J. Mendenhall, M.D. Johnson, A.C. Scheufler, A. Wilcox, V.L. Lewis, D.C. Robbins, Evaluative coping, emotional distress, and adherence in couples with type 2 diabetes, *Fam. Syst. Health* 36 (2018) 87–96, <http://dx.doi.org/10.1037/fsh0000302>.
- [11] F.S. Malik, N. Panlasigui, J. Gritton, H. Gill, J.P. Yi-Frazier, M.A. Moreno, Adolescent perspectives on the use of social media to support type 1 diabetes management: Focus group study, *J. Med. Internet Res.* 21 (2019) e12149, <http://dx.doi.org/10.2196/12149>.
- [12] M.A. Powers, J. Bardsley, M. Cypress, P. Duker, M.M. Funnell, A.H. Fischl, M.D. Maryniuk, L. Siminerio, E. Vivian, Diabetes self-management education and support in type 2 diabetes, *Diabetes Educ.* 43 (2017) 40–53, <http://dx.doi.org/10.1177/0145721716689694>.
- [13] Social media for scientists, *Nat. Cell Biol.* 20 (2018) 1329, <http://dx.doi.org/10.1038/s41556-018-0253-6>.
- [14] L. Sinnenberg, A.M. Büttenheim, K. Padrez, C. Mancheno, L. Ungar, R.M. Merchant, Twitter as a tool for health research: A systematic review, *Am. J. Public Health* 107 (2017) 1–8, <http://dx.doi.org/10.2105/AJPH.2016.303512>.
- [15] I. De La Torre-Díez, F.J. Díaz-Pernas, M. Antón-Rodríguez, A content analysis of chronic diseases social groups on Facebook and Twitter, *Telemed. e-Health* 18 (2012) 404–408, <http://dx.doi.org/10.1089/tmj.2011.0227>.
- [16] M. Belguerisse-Díaz, A.K. McLennan, G. Garduño Hernández, M. Barahona, S.J. Uglijszek, The 'who' and 'what' of #diabetes on Twitter, *Digit. Health* 3 (2017) 205520761668884, <http://dx.doi.org/10.1177/2055207616688841>.
- [17] A. Karami, A.A. Dahl, G. Turner-McGrievy, H. Kharrazi, G. Shaw, Characterizing diabetes, diet, exercise, and obesity comments on Twitter, *Int. J. Inf. Manage.* 38 (2018) 1–6, <http://dx.doi.org/10.1016/j.jinfomgt.2017.08.002>.
- [18] E. Gabarrón, E. Dorronzoro, O. Rivera-Romero, R. Wynn, Diabetes on twitter: A sentiment analysis, *J. Diabetes Sci. Technol.* 13 (2019) 439–444, <http://dx.doi.org/10.1177/1932296818811679>.
- [19] G. Bello-Ortiz, J. Hernández-Castro, D. Camacho, Detecting discussion communities on vaccination in twitter, *Future Gener. Comput. Syst.* 66 (2017) 125–136, <http://dx.doi.org/10.1016/j.future.2016.06.032>.
- [20] Y. Kim, S.N. Jang, J.L. Lee, Co-occurrence network analysis of keywords in geriatric frailty, *J. Korean Acad. Community Health Nurs.* 29 (2018) 429–439, <http://dx.doi.org/10.12799/jkacn.2018.29.4.429>.
- [21] L. Tang, B. Bie, D. Zhi, Tweeting about measles during stages of an outbreak: A semantic network approach to the framing of an emerging infectious disease, *Am. J. Infect. Control* 46 (2018) 1375–1380, <http://dx.doi.org/10.1016/j.ajic.2018.05.019>.
- [22] M. Pérez-Pérez, G. Pérez-Rodríguez, F. Fdez-Riverola, A. Lourenço, Using twitter to understand the human bowel disease community: Exploratory analysis of key topics, *J. Med. Internet Res.* 21 (2019) e12610, <http://dx.doi.org/10.2196/12610>.
- [23] N.H. Cho, J.E. Shaw, S. Karuranga, Y. Huang, J.D. da Rocha Fernandes, A.W. Ohlrogge, B. Malanda, IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045, *Diabetes Res. Clin. Pract.* 138 (2018) 271–281, <http://dx.doi.org/10.1016/j.diabres.2018.02.023>.
- [24] M.C. Riddle, W.H. Herman, The cost of diabetes cared an elephant in the room, *Diabetes Care* 41 (2018) 929–932, <http://dx.doi.org/10.2337/dci18-0012>.
- [25] C. Bommer, V. Sagalova, E. Heesemann, J. Manne-Goehler, R. Atun, T. Barnighausen, J. Davies, S. Vollmer, Global economic burden of diabetes in adults: Projections from 2015 to 2030, *Diabetes Care* 41 (2018) 963–970, <http://dx.doi.org/10.2337/dc17-1962>.
- [26] WHO, Global report on diabetes, 2016, http://www.who.int/about/licensing/copyright_form/index.html (accessed 28 October 2019).
- [27] R. Kandimalla, V. Thirumala, P.H. Reddy, Is alzheimer's disease a type 3 diabetes? A critical appraisal, *Biochim. Biophys. Acta. Mol. Basis Dis.* 1863 (2017) 1078–1089, <http://dx.doi.org/10.1016/j.bbadis.2016.08.018>.
- [28] G. Xu, B. Liu, Y. Sun, Y. Du, L.G. Snetselaar, F.B. Hu, W. Bao, Prevalence of diagnosed type 1 and type 2 diabetes among US adults in 2016 and 2017: Population based study, *BMJ* 362 (2018) <http://dx.doi.org/10.1136/bmj.k1497>.
- [29] D.M. Maahs, N.A. West, J.M. Lawrence, E.J. Mayer-Davis, Epidemiology of type 1 diabetes, *Endocrinol. Metab. Clin. North Am.* 39 (2010) 481–497, <http://dx.doi.org/10.1016/j.ecl.2010.05.011>.
- [30] L. Chen, D.J. Magliano, P.Z. Zimmet, The worldwide epidemiology of type 2 diabetes mellitus—present and future perspectives, *Nat. Rev. Endocrinol.* 8 (2012) 228–236, <http://dx.doi.org/10.1038/nrendo.2011.183>.
- [31] Y. Yamamoto, Twitter4J - A java library for the twitter API, 2018, <http://twitter4j.org/en/index.html> (accessed 10 July 2018).
- [32] J. Raffo, WorldWide gender-name dictionary, 2016, <https://ideas.repec.org/c/wip/eccode/10.html>.
- [33] R. Rothe, R. Timofte, L. Van Gool, DEX: Deep expectation of apparent age from a single image, in: *IEEE Int. Conf. Comput. Vis. Work.* 2015, pp. 10–15, https://www.cv-foundation.org/openaccess/content_iccv_2015_workshops/w11/papers/Rothe_DEX_Deep_EXpectation_ICCV_2015_paper.pdf (accessed 10 July 2018).
- [34] A. Tharwat, Classification assessment methods, *Appl. Comput. Inform.* (2018) <http://dx.doi.org/10.1016/j.aci.2018.08.003>.
- [35] D. LaMacchia, Twitter-text library, 2017, <https://github.com/twitter/twitter-text> (accessed 10 September 2019).
- [36] Aristotelis, Hunspell dictionary of english medical terms, 2016, <https://github.com/glutanimate/hunspell-en-med-glut> (accessed 10 August 2019).
- [37] L. Yujian, L. Bo, A normalized levenshtein distance metric, *IEEE Trans. Pattern Anal. Mach. Intell.* 29 (2007) 1091–1095, <http://dx.doi.org/10.1109/TPAMI.2007.1078>.
- [38] C.D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S.J. Bethard, D. McClosky, The Stanford CoreNLP Natural Language Processing Toolkit, Association for Computational Linguistics (ACL), 2015, <http://dx.doi.org/10.3115/v1/p14-5010>.
- [39] C.J. Hutto, E. Gilbert, VADER: A parsimonious rule-based model for sentiment analysis of social media text, in: *Proc. Eighth Int. AAAI Conf. Weblogs Soc. Media, The AAAI Press, Michigan*, 2014.
- [40] D.R. Amancio, F.N. Silva, L.da F. Costa, Concentric network symmetry grasps authors' styles in word adjacency networks, *EPL (Europhys. Lett.)* 110 (2015) 68001, <http://dx.doi.org/10.1209/0295-5075/110/68001>.
- [41] C. Akimushkin, D.R. Amancio, O.N. Oliveira, Text authorship identified using the dynamics of word co-occurrence networks, *PLoS One* 12 (2017) e017052, <http://dx.doi.org/10.1371/journal.pone.0170527>.

- [42] F.N. Silva, D.R. Amancio, M. Bardosova, Lda F. Costa, O.N. Oliveira, Using network science and text analytics to produce surveys in a scientific topic, *J. Informetr.* 10 (2016) 487–502, <http://dx.doi.org/10.1016/j.joi.2016.03.008>.
- [43] R.Z. Mohammad, A. Abbasi, H. Liu, *Social Media Mining: An Introduction*, first ed., 2014, <https://www.cambridge.org/us/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/social-media-mining-introduction> (accessed 1 August 2019).
- [44] A. Clauset, M.E.J. Newman, C. Moore, Finding community structure in very large networks, *Phys. Rev. E* 70 (2004) 066111, <http://dx.doi.org/10.1103/PhysRevE.70.066111>.
- [45] G. Su, A. Kuchinsky, J.H. Morris, D.J. States, F. Meng, Glay: community structure analysis of biological networks, *Bioinformatics* 26 (2010) 3135–3137, <http://dx.doi.org/10.1093/bioinformatics/btq596>.
- [46] M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, *Phys. Rev. E* 69 (2004) 026113, <http://dx.doi.org/10.1103/PhysRevE.69.026113>.
- [47] D.J. Watts, S.H. Strogatz, Collective dynamics of ‘small-world’ networks, *Nature* 393 (1998) 440–442, <http://dx.doi.org/10.1038/30918>.
- [48] A.P. Masucci, G.J. Rodgers, Network properties of written human language, *Phys. Rev. E* 74 (2006) 026102, <http://dx.doi.org/10.1103/PhysRevE.74.026102>.
- [49] Y.-R. Cho, A. Zhang, Identification of functional hubs and modules by converting interactome networks into hierarchical ordering of proteins, *BMC Bioinformatics* 11 (Suppl 3) (2010) S3, <http://dx.doi.org/10.1186/1471-2105-11-S3-S3>.
- [50] R.M. Ferreira, J.L. Rybarczyk-Filho, R.J.S. Dalmolin, M.A.A. Castro, J.C.F. Moreira, L.G. Brunnet, R.M.C. de Almeida, Preferential duplication of intermodular hub genes: An evolutionary signature in eukaryotes genome networks, *PLoS One* 8 (2013) e56579, <http://dx.doi.org/10.1371/journal.pone.0056579>.
- [51] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, 2003, <http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf> (accessed 1 August 2019).
- [52] C. Sievert, K.E. Shirley, Ldavis: A method for visualizing and interpreting topics, 2014, <https://www.aclweb.org/anthology/W14-3110> (accessed 1 August 2019).
- [53] J. Chuang, C.D. Manning, J. Heer, Termite: Visualization techniques for assessing textual topic models, 2012, <http://vis.stanford.edu/files/2012-Termite-AVI.pdf> (accessed 1 August 2019).
- [54] S. Kullback, R.A. Leibler, On information and sufficiency, *Ann. Math. Stat.* 22 (1951) 79–86, <http://dx.doi.org/10.1214/aoms/117729694>.
- [55] F. Martin, M. Johnson, More Efficient Topic Modelling Through a Noun Only Approach, in: *Proc. Australas. Lang. Technol. Assoc. Work. 2015*: 111–115. <https://www.aclweb.org/anthology/U15-1013> (accessed 12 September 2019).
- [56] H.M. Wallach, I. Murray, R. Salakhutdinov, D. Mimno, Evaluation methods for topic models, in: *Proc. 26th Annu. Int. Conf. Mach. Learn. - ICM'09*, ACM Press, ew York, New York, USA, 2009, pp. 1–8, <http://dx.doi.org/10.1145/1553374.1553515>.
- [57] S. El-Sappagh, D. Kwak, F. Ali, K.-S. Kwak, DMTO: a realistic ontology for standard diabetes mellitus treatment, *J. Biomed. Semant.* 9 (2018) 8, <http://dx.doi.org/10.1186/s13326-018-0176-y>.
- [58] W.A. Kibbe, C. Arze, V. Felix, E. Mittraka, E. Bolton, G. Fu, C.J. Mungall, J.X. Binder, J. Malone, D. Vasant, H. Parkinson, L.M. Schriml, Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data, *Nucleic Acids Res.* 43 (2015) D1071–D1078, <http://dx.doi.org/10.1093/nar/gku1011>.
- [59] D.M. Dooley, E.J. Griffiths, G.S. Gosal, P.L. Buttigieg, R. Hoehndorf, M.C. Lange, L.M. Schriml, F.S.L. Brinkman, W.W.L. Hsiao, Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration, *Npj Sci. Food* 2 (2018) 23, <http://dx.doi.org/10.1038/s41538-018-0032-6>.
- [60] A. Niknejad, A. Comte, W. Dahdul, A. Dececchi, N. Ibrahim, D. Blackburn, E. Segerdell, C. Mungall, M. Haendel, Uber anatomy ontology, 2019, <https://bioportal.bioontology.org/ontologies/UBERON> (accessed 1 August 2019).
- [61] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, Drugbank 5.0: a major update to the drugbank database for 2018, *Nucleic Acids Res.* 46 (2018) D1074–D1082, <http://dx.doi.org/10.1093/nar/gkx1037>.
- [62] F.M. Couto, A. Lamurias, MER: a shell script and annotation server for minimal named entity recognition and linking, *J. Cheminform.* 10 (2018) 58, <http://dx.doi.org/10.1186/s13321-018-0312-9>.
- [63] E.C. Davenport, N.A. El-Sanhurry, Phi/Phimax: Review and synthesis, *Educ. Psychol. Meas.* 51 (1991) 821–828, <http://dx.doi.org/10.1177/001316449105100403>.
- [64] C. Howarth, A. Gazis, D. James, Associations of type 1 diabetes mellitus, maternal vascular disease and complications of pregnancy, *Diabetes Med.* 24 (2007) 1229–1234, <http://dx.doi.org/10.1111/j.1464-5491.2007.02254.x>.
- [65] X. Hua, N. Carvalho, M. Tew, E.S. Huang, W.H. Herman, P. Clarke, Expenditures and prices of antihyperglycemic medications in the United States: 2002–2013, *JAMA - J. Am. Med. Assoc.* 315 (2016) 1400–1402, <http://dx.doi.org/10.1001/jama.2016.0126>.
- [66] R.J. Sigal, G.P. Kenny, D.H. Wasserman, C. Castaneda-Sceppa, Physical activity/exercise and type 2 diabetes, *Diabetes Care* 27 (2004) 2518–2539, <http://dx.doi.org/10.2337/diacare.27.10.2518>.
- [67] J. Tay, N.D. Luscombe-Marsh, C.H. Thompson, M. Noakes, J.D. Buckley, G.A. Wittert, W.S. Yancy, G.D. Brinkworth, Comparison of low- and high-carbohydrate diets for type 2 diabetes management: A randomized trial, *Am. J. Clin. Nutr.* 102 (2015) 780–790, <http://dx.doi.org/10.3945/ajcn.115.112581>.
- [68] P.S. Leung, The potential protective action of vitamin d in hepatic insulin resistance and pancreatic islet dysfunction in type 2 diabetes mellitus, *Nutrients* 8 (2016) <http://dx.doi.org/10.3390/nu8030147>.
- [69] G. Lastra, S. Syed, L.R. Kurukulasuriya, C. Manrique, J.R. Sowers, Type 2 diabetes mellitus and hypertension: an update, *Endocrinol. Metab. Clin. North Am.* 43 (2014) 103–122, <http://dx.doi.org/10.1016/j.ecl.2013.09.005>.
- [70] C.R. Collins, K. Stephenson, A circle packing algorithm, *Comput. Geom.* 25 (2003) 233–256, [http://dx.doi.org/10.1016/S0925-7721\(02\)00099-8](http://dx.doi.org/10.1016/S0925-7721(02)00099-8).
- [71] G. Shaw, A. Karami, *Computational Content Analysis of Negative Tweets for Obesity, Diet, Diabetes, and Exercise*, John Wiley & Sons, Ltd., 2017, <http://dx.doi.org/10.1002/pr2.2017.14505401039>.
- [72] R. Wynn, S.O. Oyeyemi, J.-A.K. Johnsen, E. Gabarron, Tweets are not always supportive of patients with mental disorders, *Int. J. Integr. Care* 17 (2017) 149, <http://dx.doi.org/10.5334/ijic.3261>.
- [73] A.D. Association, Prevention or delay of type 2 diabetes, *Diabetes Care* 40 (2017) S44–S47, <http://dx.doi.org/10.2337/dc17-S008>.
- [74] V.C. Emnacen, *Be smart drink water a guide for school principals in restricting the sale and marketing of sugary drinks in and around schools*, 2016.
- [75] J. Naumann, D. Biehler, T. Lüty, C. Sadaghiani, Prevention and therapy of type 2 diabetes—what is the potential of daily water intake and its mineral nutrients?, *Nutrients* 9 (2017) <http://dx.doi.org/10.3390/nu9080914>.
- [76] R. Gualdani, M.M. Cavalluzzi, G. Lentini, S. Habtemariam, The chemistry and pharmacology of citrus limonoids, *Molecules* 21 (2016) <http://dx.doi.org/10.3390/molecules21111530>.
- [77] J. Rosenstock, C. Mathieu, H. Chen, R. Garcia-Sanchez, G.L. Saraiva, Dapagliflozin versus saxagliptin as add-on therapy in patients with type 2 diabetes inadequately controlled with metformin, *Arch. Endocrinol. Metab.* 62 (2018) 424–430, <http://dx.doi.org/10.20945/2359-3997000000056>.
- [78] A. Magbri, M. El-Magbri, K. Suljuki, S. Rashid, Two options the sweetest among them is bitter: Fournier-gangrene associated with sodium-glucose co-transporter 2-inhibitors, *Glob. J. Urol. Nephrol.* (2019) <http://dx.doi.org/10.28933/gjun-2019-06-2605>.
- [79] S. Adimadhyam, G.T. Schumock, G.S. Calip, D.E. Smith Marsh, B.T. Layden, T.A. Lee, Increased risk of mycotic infections associated with sodium-glucose co-transporter 2 inhibitors: a prescription sequence symmetry analysis, *Br. J. Clin. Pharmacol.* 85 (2019) 160–168, <http://dx.doi.org/10.1111/bcp.13782>.
- [80] L. Celardo, M.G. Everett, Network text analysis: A two-way classification approach, *Int. J. Inf. Manage.* 51 (2019) 102009, <http://dx.doi.org/10.1016/j.ijinfomgt.2019.09.005>.



Gael Pérez Rodríguez is a Ph.D. in Computer Science of the University of Vigo. He is currently researching at the fields of text mining, social mining, machine learning and artificial intelligence, applied to biomedical areas.



Martín Pérez Pérez is a Ph.D. in Computer Science of the University of Vigo. He is currently researching at the fields of text mining, social mining, machine learning and artificial intelligence, applied to biomedical areas.



Florentino Fdez-Riverola is a Full Professor of the Department of Computer Science at the University of Vigo (Spain) and Coordinator of the New Generation Computer Systems group (SING, <http://sing-group.org>), which is dedicated to the research and development of cutting-edge computational methodologies and applications.



Anália Maria Garcia Lourenço is a faculty member of the Department of Computer Science and a researcher affiliated to the Biomedical Research Centre (CINBIO), at the University of Vigo and the Centre of Biological Engineering, at the University of Minho. Her main research interests include computational intelligence, bioinformatics and systems biology.