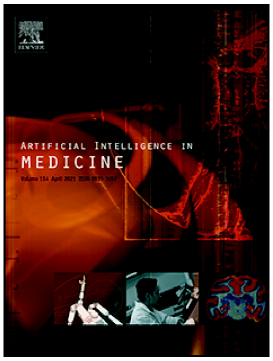


A framework to extract biomedical knowledge from gluten-related tweets: the case of dietary concerns in digital era

Martín Pérez-Pérez, Gilberto Igrejas, Florentino Fdez-Riverola, Anália Lourenço

PII: S0933-3657(21)00124-X

DOI: https://doi.org/10.1016/j.artmed.2021.102131

Reference: ARTMED 102131

To appear in: *Artificial Intelligence In Medicine*

Received date: 4 September 2020

Revised date: 27 April 2021

Accepted date: 22 June 2021

Please cite this article as: M. Pérez-Pérez, G. Igrejas, F. Fdez-Riverola, et al., A framework to extract biomedical knowledge from gluten-related tweets: the case of dietary concerns in digital era, *Artificial Intelligence In Medicine* (2021), https://doi.org/10.1016/j.artmed.2021.102131

A framework to extract biomedical knowledge from gluten-related tweets: the case of dietary concerns in digital era

Martín Pérez-Pérez[1,2,3][0000-0003-1349-6562] *, Gilberto Igrejas[4,5,6] [0000-0002-6365-0735] Florentino Fdez-Riverola [1,2,3][0000-0002-3943-8013] and Anália Lourenço[1,2,3,7][0000-0001-8401-5362]

[1] ESEI, Department of Computer Science, University of Vigo, Edificio Politécnico, Campus Universitario As Lagoas s/n 32004, Ourense, Spain
{martiperez, riverola, analia}@uvigo.es
[2] CINBIO, The Biomedical Research Centre, University of Vigo, Campus Univesitario Lagoas-Marcosende, 36310, Vigo, Spain
[3] SING, Next Generation Computer Systems Group, Galicia Sur Health Research Institute, SERGAS-UVIGO, Spain
[4] Department of Genetics and Biotechnology, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal
gigrejas@utad.pt
[5] Functional Genomics and Proteomics Unit, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal
[6] LAQV-REQUIMTE, Faculty of Science and Technology, Nova University of Lisbon, Lisbon, Portugal
[7] CEB, Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

* Corresponding author: Martín Pérez [Tlf.: +34 988 387013, Fax: +34 988 387001]

Abstract. Big data importance and potential are becoming more and more relevant nowadays, enhanced by the explosive growth of information volume that is being generated on the Internet in the last years. In this sense, many experts agree that social media networks are one of the internet areas with higher growth in recent years and one of the fields that are expected to have a more significant increment in the coming years. Similarly, social media sites are quickly becoming one of the most popular platforms to discuss health issues and exchange social support with others. In this context, this work presents a new methodology to process, classify, visualise and analyse the big data knowledge produced by the *sociome* on social media platforms. This work proposes a methodology that combines natural language processing techniques, ontology-based named entity recognition methods, machine learning algorithms and graph mining techniques to: (*i*) reduce the irrelevant messages by identifying and focusing the analysis only on individuals and patient experiences from the public discussion; (*ii*) reduce the lexical noise produced by the different ways in how users express themselves through the use of domain ontologies; (*iii*) infer the demographic data of the individuals through the combined analysis of textual, geographical and visual profile information; (*iv*) perform a community detection and evaluate the health topic study combining the semantic processing of the public discourse with knowledge graph representation techniques; and (*v*) gain information about the shared resources combining the social media statistics with the semantical analysis of the web contents. The practical relevance of the proposed methodology has been proven in the study of 1.1 million unique messages from more than 400,000 distinct users related to one of the most popular dietary fads that evolve into a multibillion-dollar industry, i.e., gluten-free food. Besides, this work analysed one of the least research fields studied on Twitter concerning public health (i.e., the allergies or immunology diseases as celiac disease), discovering a wide range of health-related conclusions.

Keywords: Social media, sociome profiling, text mining, graph mining, machine learning, health for informatics.

1. Introduction

Big data importance and potential are becoming more and more relevant nowadays, enhanced by the explosive growth of information volume that is being generated on the Internet in recent years [1,2]. In this sense, many experts agree that social media networks are one of the internet areas with the higher growth in recent years and one of the fields that are expected to have a more significant increment in the coming years[3–5]. This growth, as well as the popularization of the social media sites among the population as platforms to share and obtain information, is allowing the application of different methodologies to infer relevant knowledge into diverse domains [6–9]. In the same way, social media sites are quickly becoming one of the most popular platforms to discuss health issues and exchange social support with others. These platforms create the opportunity to interact among large groups of people that share similar interests and suffer the same afflictions [10–12]. In this sense, recent studies have demonstrated the relevant role of social media platforms in the health of people's behaviour, the diffusion of dangerous health trends, and the divulgation of misinformation [13–15].

One of the most studied social media platforms in the different research areas is Twitter, given its spread of use, public nature, and socio-technical flexibility [16,17]. Twitter is an interactive social media platform established in 2006 that allows users to send 280-character messages, or tweets, to one another. Every day, 500 million tweets are sent by more than 300 million active worldwide users [18]. Due to this, Twitter is a valuable source of information for biomedical researchers interested in capturing up-to-date data about a specific health area and harnessing the platform for the study, recruitment or intervention of different medical concerns and awareness campaigns. Twitter-based health research is a growing field, evidenced by the increasing number of publications per year and the diversity of funding organisations involved in these studies [19]. However, the free text style of tweets and user profile data presents severe challenges to information processing and domain-specific knowledge inference [20,21]. Similarly, a high volume of published messages on the platform can be shared or being influenced by organizations, businesses and other stakeholders that try to redirect the community conversations to a specific issue [22,23]. So, the characterisation of the user role and the profiling of the different accounts that are part of the community improve understanding and analysis of the discussed topics [24].

Within this context, the present paper focuses on the problem of interpreting the health-related public debate by examining the discussion footprints that people around the world exposed on social media (or *sociome)* applying a combination of natural language processing (NLP), ontology-based named entity recognition (NER), machine learning (ML) and different knowledge inference techniques to process, classify, visualise and analyse the big data knowledge produced on social media platforms. Therefore, the main contribution of this work lies in developing a new methodology to be able to: (*i*) identify the individuals and patient experiences from the public discourse reducing the numerous Twitter messages generated by spammers and general or commercial messages disseminated by stakeholders or informative accounts. This process reduces the influence that these messages could have in the semantical analysis of the social dialogue, focusing the study on real experiences; (ii) reduce the lexical noise normalising the different ways that users express their health-related issues (e.g. usage of synonyms or acronyms) through the use of domain ontologies; (*iii*) infer the demographic data of the individuals in a health-related community combining the semantical, geographical and visual profile information; (*iv*) perform a community detection study by the semantic analysis of each user discourse combined with knowledge graph representation techniques; (*v*) gain information about the shared resources analysing the semantic information of the web content.

2. Related work

The increasing popularity of social media networks and their impact on the health and commercial domains are being analysed and studied nowadays [25,26]. The opportunity provided by the immense amount of data generated by users in their interactions with others worldwide offers the possibility to obtain new knowledge using different computational techniques. In this line, Albert Park et al. [27] proposed unsupervised mining and visualization techniques to analyse the content of the user messages published on Reddit to compare the mental health communities and design and guide new patient education programs. In the work of Karami et al. [28], the authors applied unsupervised topic discovery methods based on the Latent Dirichlet Allocation technique to characterise diabetes, diet, exercise, and obesity comments on Twitter. More recently, Lenzi et al. [29] proposed unsupervised text mining techniques to acquiring diabetes-centred information on health care and compared the results of their study to standard methodologies, such as questionnaire research. Along the same lines, the work of Shaw et al. [30] combined unsupervised topic models and sentiment analysis to analyse the negative tweets related to obesity, diet, diabetes, and exercise. Compared to these previous studies, the current proposal combines the user profile categorisation and the semantical analysis to focus attention on individual and patient experiences. This approach reduces the discussion noise of general or commercial messages disseminated by stakeholders or newsletter accounts, minimizing their effect in the semantic analysis. Complementarily, the current work proposes a health-related topic identification technique based on domain ontologies to identify the medical domain topics discussed on the social networks and a more informative knowledge representation technique to visualize the discussed topics as well as the relationship between them. In this regard, the proposed ontology-based technique takes advantage of the semantical capabilities proportioned by the ontologies to normalise the lexical difference in which users express their health-related experiences. This idea is partially supported by the work of Masmoudi et al. [31], which explored the use of ontology-based approaches to analyse the radicalization indicators in online messages and exposed the benefits of text mining approaches over other counterparts.

Regarding the community analysis, Beguerisse-Díaz et al. [32] combined non-grouping techniques from anthropology, network science and information retrieval to detect the most influential and contributing people in the twitter diabetes communities and to discover different health-related implications for public health professionals and policymakers. More recently, Bello-Orgaz et al. [33] proposed a graph methodology based on the user interactions to detect the different user communities of the public vaccination discussion on Twitter. Similarly, Sathiyakumari et al. [34] proposed the use of different graph metrics to identify groups and subgroups in social networks. Compared to these previous works, the presented approach introduces an alternative community detection technique that combines user interaction graphs with the semantic analysis of the overall user messages as a whole. In this sense, this technique groups individuals in communities based on the main discussed topics identified in their messages and represents the community interactions as a graph to evaluate the most relevant users of each community and how the information flows.

In terms of user geolocation, Zheng et al. [35] provided a complete survey of different location prediction techniques on Twitter. Compared to existing alternatives, the current work bases the geolocation of the users in the combined analysis of three fundamental registers of the user profile: (*i*) the semantic processing of the declared location, (*ii*) the declared time zone, and (*iii*) the GPS coordinates. From another perspective, several authors addressed the gender categorisation problem by using different multimodal gender identification methodologies on Twitter, in which the best results were obtained by combining the profile textual information and the user image [36,37]. In this regard, the current work proposes a more accurate methodology that also uses geolocated information to identify the user gender.

From another perspective, Pérez-Pérez et al. [38] published an exploratory analysis of 24,634 tweets related to human bowel disease proposing different text mining and user characterisation techniques to discover relevant health outcomes to support decision making among the different user roles. Complementarily, Ke et al. [39] proposed a gender and resource analysis to study the scientist discussion on social media. Compared to these previous works, the current approach: (*i*) introduces a novel technique to analyse and represent the relevance and the main topic of the shared resources based on their content, (*ii*) proposes a new methodology to analyse the different user communities and their public discourse centring the study on individuals and patients, (*iii*) introduces a novel individual, and patient recognition process, (*iv*) proposes a novel gender recognition technique and (*v*) uses a novel geolocation account recognition that considers the user profile information and the published GPS coordinates.

In other social media analysis, Bian et al. [40] analysed the social media community related to Lynch Syndrome using different machine learning methods to understand the correlation between promotional health-related information and laypeople's discussions. However, as the authors said in the manuscript limitations section, the obtained results would have been more precise by classifying the user accounts based on their profile (i.e. demographics and gender) and using more advanced natural language processing tools and machine learning models to normalise and discover the discussed key topics. In this sense, the current work addresses these issues in greater depth by proposing a novel approach covering these aspects.

Finally, in terms of gluten-related analysis on social media, Puerta et al. [41] carried out a manual curated text analysis of a subset of 3,000 Spanish tweets focused on gluten-free proposing the use of co-occurrence networks to analyse the massive information available on social media networks as Twitter. Compared to this study, the current work performs an analysis of 1.1 million of English messages related to gluten-free from more than 400,000 distinct users, proposing a patient characterisation method to discover distinct health issues and dietary concerns focused on individuals and patient experiences.

Overall, the combination of the different techniques used in the current work makes this proposal a complete methodology for the analysis of the social media knowledge from a biomedical perspective focused on the individual and patient experiences.

3. The case study: the gluten-free food community

As recommendation of UE the Farm to Fork Strategy as an important step in ensuring a sustainable, fair, and resilient food system, which is central to achieve the goals set out in the European Green Deal; emphasizes the inextricable links between healthy people, healthy societies, and a healthy planet [42]. The importance of seed security and diversity notably promoting EU-grown plant proteins delivering locally sourced food and feed stuffs with high nutritional value while granting farmers access to quality seeds for plant varieties adapted to the pressures of climate change, including traditional and locally adapted varieties while ensuring access to innovative plant breeding to contribute to healthy seeds and protect plants against harmful pests and disease, in line with the environmental objectives and the 'do no harm' principle of the Green Deal [43].

Concerning public health, allergies and immunology diseases are among the most minor reported topics on Twitter studies [19]. On this subject, one of the fastest-growing autoimmune disorders in the last years and one of the most common genetic diseases in the West is celiac disease (CD) [44,45]. CD is a severe autoimmune reaction to eating gluten, a protein found in wheat, barley and rye. However, CD is not the only allergy that has adverse reactions to this protein. Wheat allergy and non-celiac gluten sensitivity or wheat intolerance syndrome are also classified among gluten-related disorders. The only existent treatment for these chronic diseases is to follow a gluten-free diet (GFD). Similarly, GFD is being explored as an effective treatment in other chronic diseases, with a significant reduction of their symptoms. Recent studies examine the benefits of this diet as a complementary therapy to reduce the ailments caused by some bowel diseases, such as inflammatory bowel syndrome and irritable bowel disease [46–50]. On the other hand, some researchers have explored the application of the GFD on a broader spectrum of illnesses and syndromes with or without intestinal pathology. In this sense, this nutritional therapy has been tested as a treatment of diabetes [51], schizophrenia [52] and other mental diseases [53].

Conversely, many people follow a self-prescribed GFD, even though the majority have not been previously diagnosed as having gluten disorders. These individuals rely on claims that a GFD improves general health, i.e., motivation is personal rather than medical. Social influencers also supported this claim. For example, nearly 50% of 910 athletes (including world-class and Olympic medallists) adhere to GFD because they perceive it as more healthy and providing energy benefits [53]. The public perception that a GFD promotes improved general health, besides the current confusion about the advantages and disadvantages of this alimentary trend, are some of the reasons that the GFD has become one of the most popular dietary fads and evolved into a multibillion-dollar industry [54–56].

The increasing popularity of GFD has important implications for children and parents. The social popularisation of a GFD without a medical prescription can lead to nutritional risk behaviours associated with inadequate macronutrient intake and dietary imbalances [57].

In terms of health information, research questionnaire-derived data indicate that individuals prefer as sources of health-related information Internet, print media sources, cookbooks, disease support groups, and other patient's experiences over medical books and even the family doctor [58]. Therefore, social media platforms play an essential role in proper medical education of society and signalling potential risks in alimentary fads. In this context, this paper aims to study the public *sociome* message related to the GFD, characterising the gluten-free community and their social message over the Twitter social platform to discover relevant health outcomes.

4. Materials and methods

4.1 General workflow

Figure 1 depicts the workflow implemented in the present study to retrieve, process, and analyse gluten-related tweets, which consists of four fundamental tasks: (*i*) data collection and filtering, (*ii*) data processing, (*iii*) the application of complemented knowledge inference techniques and (*iv*) the study of the obtained knowledge.

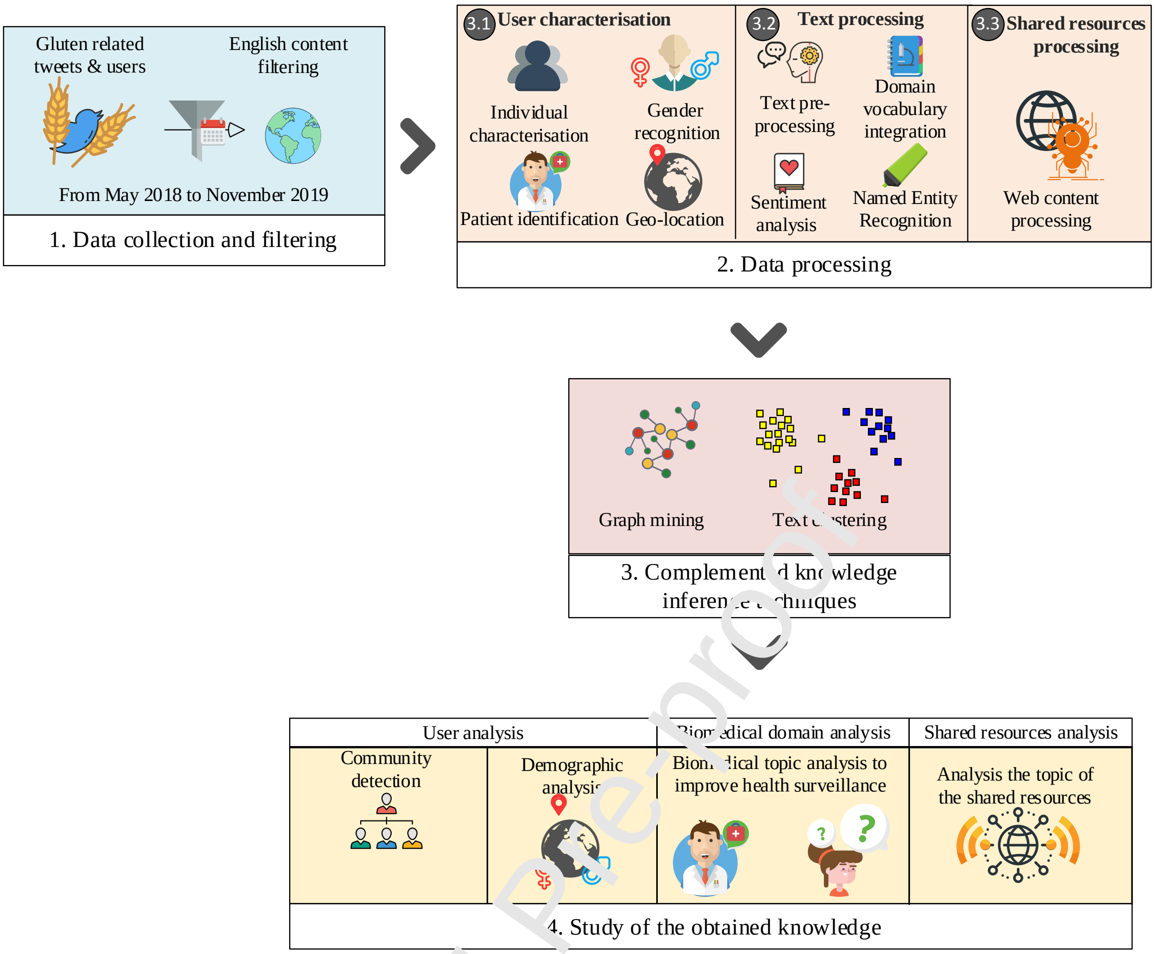

Gluten related tweets & users
English content filtering
From May 2018 to November 2019
1. Data collection and filtering
3.1 User characterisation
Individual characterisation
Gender recognition
Patient identification
Geo-location
3.2 Text processing
Text pre-processing
Domain vocabulary integration
Sentiment analysis
Named Entity Recognition
3.3 Shared resources processing
Web content processing
2. Data processing
Graph mining
Text clustering
3. Complemented knowledge inference techniques
User analysis
Community detection
Demographic analysis
Biomedical domain analysis
Biomedical topic analysis to improve health surveillance
Shared resources analysis
Analysis the topic of the shared resources
4. Study of the obtained knowledge

Figure 1. The workflow implemented to retrieve, process, analyse and study the gluten disease-related tweets.

Initially, tweets from May 1, 2018, to November 23, 2019, containing the keyword "Gluten" were collected via the Twitter4J, a Java library for the Twitter API [59]. This task retrieved 1.1 million unique gluten-related tweets in English from more than 400,000 distinct users, removing suspended accounts. Figure 2 shows the distribution of the final set of tweets along the time.

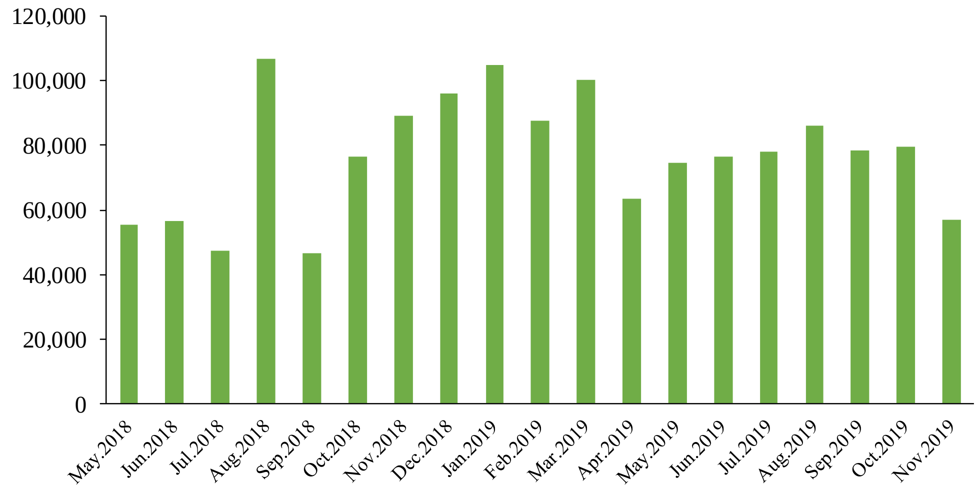


Figure 2. Gluten-related tweets distribution along time from May 1, 2018, to November 23, 2019.

The second task addressed the processing of the Twitter messages and the user data by applying text mining (TM) and ML methods to assist to (*i*) the user classification methodology described in section 4.2 that comprises the user profile characterisation, namely to identify individuals, patient accounts, user gender and user location; (*ii*) the processing methodology of the messages that involves the ontology-based entity recognition and the sentiment analysis methods explained in section 4.3; and (*iii*) the processing methodology of the shared resources (i.e., shared links) explained in section 4.4 that comprises the semantic analysis of the website contents to evaluate what kind of information was more relevant and shared among the communities. In the third task, different knowledge inference techniques based on graph mining methods and machine learning methods were performed to obtain a deeper understanding of the conversations and to complement the results of the previous task. Finally, in the last task, all the big data knowledge retrieved from processing the *sociome* were structured for their visualisation and analysis.

4.2 User characterisation

All data presented in the user profiles were collected and processed to profile the different accounts involved in the conversations. The aim was to distinguish the distinct users in terms of their role as individuals and non-individuals (i.e., organisations, stakeholders, educational, informative or commercial accounts that do not represent a personal experience), their gender (i.e., male or female) and their geolocation. The following subsections detail the different methodologies applied in user characterisation processing. The agreement rates of the different characterisation algorithms were calculated using the standard measures of recall, precision and F-score.

Correspondingly, recall is the percentage of correctly labelled positive results in overall positive cases, i.e., it is a measure of the ability of a system to identify positive cases.

$$recall = \frac{TP}{(TP+FN)} \tag{1}$$

Precision is the percentage of correctly labelled positive results overall positive labelled results, i.e., it is a measure of the reproducibility of a classifier of the positive results.

$$precision = \frac{TP}{(TP+FP)} \tag{2}$$

And, the F-score (or balanced F-measure) is the harmonic mean between precision and recall.

$$F - score = 2 * \frac{precision \times recall}{precision + recall} \tag{3}$$

4.2.1 Individual characterisation

The developed algorithm determines that a user is an individual if one of the following sequential criteria steps is fulfilled: (*i*) the user account name was recognised as a person's name based on a lookup over a worldwide name dictionary [60]; (*ii*) the face of a person was recognised in the user profile image by the application of a face detector pre-trained ML model [61]; (*iii*) Twitter identified the user as a contributor or a translator; (*iv*) the account description was written in the first person, (i.e., pronouns and their variant forms); (*v*) the user description contained emojis or emoticons (i.e., organization profiles tend to be described more formally); or (*vi*) the user account name or the description included English honorifics (e.g., Ms. or Mr.).

To better understanding the implemented procedure, supplementary material 1 shows the flow diagram to classify a specific user as an individual. The straightforward application of these criteria led to the labelling of 425,160 user accounts (87% of total users) as "Individual" and 62,771 (13% of users) as "Non-individuals" with an F-score of 0,84 calculated over a sample subset of 1000 random user profiles manually labelled.

4.2.3 Patient identification

This part of the algorithm was based on the matching of characteristic phrasing. In particular, it was devised to identify people who explicitly express suffering from the disease (e.g., "I have celiac disease") or relatives who talk about the condition in a third-person way (e.g., "My child has gluten allergy") but not for pets (e.g., "My dog has gluten sensitivity"). To achieve this, the matching rules tried to find a pronoun and/or a list of verbs (e.g., "He has", "I am diagnosed" or "I suffer") related with a mention of the disease and/or its variants (e.g., "coeliac", "gluten sensitivity" or "gluten allergy"). The application of these heuristics led to the labelling of 13,504 users as "Patient" (3% of "individual" accounts) and 411,656 users as "Individuals with no health information found" (97% of "individual" accounts) with an F-score of 0,89 calculated over a sample subset of 1000 random user manually labelled.

4.2.2 Gender recognition

Gender is an essential biological variable affecting experimental outcomes as well as health and disease. To produce precise and reproducible results applicable to both men and women, gender should be considered a relevant biological variable from basic and preclinical research [62,63]. In this regard, the following procedure was applied to assign the gender to the individual accounts previously identified. First, if there was a perfect match against the gender-name database [60] (i.e., a unique gender was associated with the specific name), then the gender was set. Otherwise, the name and the user geolocation were combined to identify the gender. If the gender probability of the name was greater than 80%, then the gender was set. Finally, if all of the previous criteria were not satisfied, then the procedure continues checking if there was a unique detected face in the user profile picture. If a unique recognised face exists, then a deep learning image recogniser, pre-trained over more than 500,000 public face images extracted from Wikipedia and IMDb with an accuracy of 96%, was applied to infer the user gender [64]. In any other cases, the gender was set as "Unknown". To better understanding the implemented procedure, supplementary material 2 shows the flow diagram to identify the user gender. The application of these criteria led to the labelling of 166,223 user accounts (39% of “individual” accounts) as “female”, 144,016 user accounts (34% of “individual” accounts) as “male”, and 112,514 (27% of “individual” accounts) as “Unknown” with an F-score of 0.86 calculated over a sample subset of 1000 random user profiles manually labelled.

4.2.4 Geolocation

The geographic structure of a population is a crucial determinant to carry out public information campaigns and study the impact of healthcare stakeholder's community awareness campaigns in different demographic areas [65]. There are two primary ways to identify the location of the users. The simplest way is to collect the coordinates of the location from their Twitter messages. However, this is only possible if users enable the geolocation option on their accounts. The second way is to analyse the self-reported location in user profiles. However, user self-reported location is usually free content text, which frequently brings consistency issues to use it in demographic research studies (for example, one person can specify "LA" while others may identify the same area as "Los Angeles"). Another issue to consider was the presence of locations that do not provide valuable information, such as “in my house” or “at the moon”, and city names related to multiple countries (for instance, "Sydney" is the name of a city in Australia but also in Canada).

Therefore, the applied Geolocation strategy used the GeoNames database [66], which contained more than 10 million topographical names and was accessible through a free web service, follow the next heuristic. First, all user profiles containing a Global Positioning System (GPS) coordinates (e.g., “iPhone: 47.595398, -122.328018”) were used to geolocate the user by a sorted distance proximity algorithm, applying the Equation 1:

$$\begin{aligned} Distance_{Km} = 6371 * \ arcos(cos(\ radians(ComparedLatitude) \ ) \\ * \ cos(\ radians(\ TweetLatitude \ ) \ ) \\ * \ cos(\ radians(\ TweetLongitude) \\ - \ radians(ComparedLongitude) \ ) \\ + \ sin(\ radians(ComparedLatitude) \ ) \\ * \ sin(\ radians(\ TweetLatitude \ ) \ ) \end{aligned} \tag{4}$$

Where $Distance$ is the circle distances between two pairs of coordinates $(ComparedLatitude, ComparedLongitude)$ and $(TweetLatitude, TweetLongitude)$ measured in kilometres.

If there were no identified (GPS) coordinates in the user profile, then the free text location in the account description was searched against all location names at the GeoNames database. If the result was not accurate enough (i.e., matching multiple database entries), the time zone and the UTC offset were used to assist in the definition of the location. When different areas had a similar name and the same time zone, but in distinct countries, the location labelled as “Not located”. Finally, in those cases that the user was “Not located” and any of their Twitter messages indicated GPS coordinates, then their geolocation was searched by proximity to all

GPS coordinates contained at the GeoNames database. The application of the proposed algorithm led to the labelling 164,596 accounts (that was 51% of 317,184 individuals with some kind of location provided) with an F-score of 0,95 calculated over a sample subset of 1000 random user profiles manually labelled.

4.3 Text processing

To analyse the content of the dataset, it was essential to properly recognise the relevant (topic-related) terms mentioned and their related sentiment. For this purpose, different pre-processing and TM techniques were applied to assist (*i*) the in-house developed ontology-based named entity recogniser to establish relevant domain semantic categories; and (*ii*) the sentiment analyser to characterise the emotions among the identified terms.

4.3.1 Pre-processing and part of speech

To improve the analysis of the text, the following pre-processing steps were applied to reduce the text noise [67]. As a first step, elements that did not provide relevant information like some special characters (e.g., ‘&’, ‘(‘, ‘)’, ‘*’, ‘+’, ‘<’, or ‘>’), user mentions (represented with ‘@’), hashtags (represented with ‘#’), Uniform Resource Locators (URLs) and emojis were removed from texts. In the special case of emojis, they were only removed to assist the developed in-house ontology-based named entity recognizer for establishing relevant domain semantic categories. All these operations were carried out using the Twitter-Text library [68]. In the next step, three or more consecutive and identical characters at word tokens were standardised (e.g., ‘alleeeeeeergy’ to ‘alleergy’). Then, a spelling error correction was carried out using the Hunspell dictionary [69]. Editing was done automatically by selecting the suggested word with the highest similarity to the original (incorrect) term calculated with the Normalised Levenshtein algorithm [70]. In the next step, different text processing methodologies were performed to prepare texts for clustering and named entity recognition (NER) steps. In detail, the following operations were carried act: (*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*) expansion of abbreviations and shorthand terms (e.g., GFD to Gluten-free diet). (*iv*) part of speech (POS) tagging (i.e., to identify the lexical category of each token); (*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 were also considered. All the previous steps were implemented using the Stanford CoreNLP pipeline [71].

4.3.2 Domain vocabulary integration

The following domain-related ontologies and dictionaries were applied to recognise, extract and standardised the semantic domain concepts from the Twitter messages: the FoodOn ontology [72], the Physical Activity Ontology [73], the National Cancer Institute Thesaurus ontology [74], the Symptom Ontology [75], the Foundational Model of Anatomy ontology [76], the Medical Subject Headings Ontology, the Chemical Entities of Biological Interest lexicon [77], the Disease Ontology [78], the DrugBank lexicon [79], the Kegg lexicon [80], and an expert-manually curated list of food diets. Overall, a lexicon of 295,646 entries supported the entity recognition task. Concepts were semantically grouped into the following categories “Disease”, “Food & Nutrition”, “Anatomy”, “Drug & Chemical compounds”, “Symptoms”, “Physical activity”, “Plants”, “Diet” and “Dietary Supplements”.

4.3.3 Named entity recognition

The named entity recognition pipeline was implemented in-house and entailed a dictionary lookup, as well as pattern and rule-based recognition. An inverted recognition technique was used in actual entity recognition [81]. This technique uses the words in the text as patterns to be matched against the lexicon. This approximation is adequate for the type of texts analysed in this work due to their short length compared to the size of the lexicon. Moreover, recognition preference was given to the longest possible n-gram (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).

4.3.4 Sentiment analysis

The sentiment of the Twitter message was analysed using the Valence Aware Dictionary and sEntiment Reasoner (VADER) for Python [82], a lexicon and rule-based sentiment analysis tool adjusted explicitly to the detection sentiments expressed in social media. VADER uses a parsimonious rule-based model to assess the sentiment of tweets based on the sentiment lexicon that is used in the social media domain. In addition, it translates UTF-8 encoded emojis to text to be used as relevant sentiment information. The predicted sentiment (i.e., compound score) is computed by summing the valence scores of each word in the lexicon, adjusted according to emotion-related rules, and then normalised to have values between -1 (most extreme negative emotion) and +1 (most extreme positive emotion).

4.4 Shared resources processing

Nowadays, social platforms have become relevant media for information sharing; however, people have to use external links (URLs) to enrich and support the published information due to the usual character limitation. For this reason, social media users usually tend to employ short URLs services (e.g., bit.ly or ow.ly) to reduce the string size of the links [83]. This means that it is more difficult to carry out an analysis of the leading platforms that users use to support their messages or that they consider relevant information to be shared among the community. In this regard, some works have demonstrated that the content of websites can be a valuable source of information in different domain areas to provide new information about the social communities [84]. To analyse the shared resources, a web crawler supported the automatic retrieval of website contents. The purpose was to study the website content to examine what resources supported the community discussion (i.e., the main category of the resource), what were the most shared website content categories (i.e., resources contained in more volume of Twitter messages) and what were the most popular resource categories (i.e., resources that reached most retweet and favourites).

4.4.1 Web content processing

To obtain the expanded URL and a tiny description of the shared resources, a web crawler pipeline was applied with the JSOUP Java library [85]. The main objective was to obtain a brief description of the general topic of the site, using the website content if the HyperText Markup Language (HTML) meta attributes (i.e., keywords, description or Open Graph protocol tags) were not present [86]. The execution of this pipeline allowed the retrieval of the content of 139,485 pages, excluding: (*i*) those that were no longer available; (*ii*) recursive links to the own social platform; and (*iii*) links to Instagram and Amazon since they did not allow crawling to their websites (a login or a captcha protected them).

4.5 Complemented knowledge inference techniques

To complement the obtained knowledge after processing the different social media data, the following inferred knowledge techniques were integrated. Twitter messages and website contents can be explored in terms of clustering similarities [87]. To group related texts, a combination of Term Frequency–Inverse document frequency (TF-IDF) and K-means methods were applied. The TF-IDF measure expresses the relevance of a term in a dataset in regards to the total number of times that a term appears in the global documents and in a particular document. It is expressed as follows (Equation 2):

$$tf - idf(t, d, D) = tf(t, d) \times idf(t, D) \tag{5}$$

Where t is the evaluated term and d is one document in de dataset D.
The $tf(t, d)$ expresses the ratio number of terms t in a document d, and is expressed as (Equation 3):

$$tf(t, d) = \frac{n_t}{\sum_k n_k} \tag{6}$$

where n_t is the number of occurrences of the term t in a document d and n_k is the number of all terms in a document d.
The $idf(t, D)$ expresses the logarithmic ratio number of terms t in the dataset D, and is expressed as (Equation 4):

$$idf(t, D) = \ log \frac{|D|}{|\{d_i \in D | t \in d_i\}|} \tag{7}$$

Finally, to discover groups of documents that cover similar subjects, a k-means clustering algorithm was applied. K-means is an unsupervised machine-learning algorithm to distribute a set of k clusters based on the distance measure between them. Documents that have similar features $\vec{d}$ are grouped in the same cluster, but should have highly dissimilar features with the other clusters. To calculate the distance between two documents, i.e., talk about similar topics, different distance measure based on the document TF-IDF space vector was applied. The primary distance functions for text document distance calculation are Euclidean, Manhattan distance, Minkowski, cosine and humming [88]. In this sense, to find the best solution for each case, the different metrics have been tested and the one with the lowest Davies Bouldin index has been chosen in each case. The Davies Bouldin index finds out for every cluster which cluster is the most similar. After it summarizes the maximum cluster similarities to create a single index DB (Equation 5). If the index is low, the clusters are not very similar to each other, which means that they are compact and well separated.

$$DB = \frac{1}{n_c}\sum_{i=1}^{n_c} R_i, \tag{8}$$

where R_i is defined as Equation 6:

$$R_i = \underset{j=1\ldots.n_c, i\neq j}{Max}(R_{ij}), i{=}1\ldots.n_c \tag{9}$$

where $R_{ij}$is the similarity measure of clusters and has to satisfy the following five conditions (i) $R_{ij} \geq 0$; (ii) $R_{ij} = R_{ji}$; (iii) if $S_i = 0$and $S_j = 0$ then $R_{ij} = 0$; (iv) if $S_i > S_j$ and $d_{ij} = d_{ik}$ then $R_{ij} > R_{ik}$; (v) if $S_i = S_j$ and $d_{ij} < d_{ik}$ then $R_{ij} > R_{ik}$ where S_i is the dispersion measure of a cluster and $d_{ij}$is the cluster dissimilarity measure.

On the other hand, graph representation and mining methods were issued to measure the relevance of (i) individual terms as well as term-term pairs and (ii) to measure the importance of the different user communities. Specifically, this analysis was applied to evaluate the co-occurrence of semantically meaningful terms (i.e., whenever two terms were found in the same tweet, these two terms were considered to share an edge) and to evaluate the user interactions via user mentions and retweets. Knowledge graphs are generally described in terms of the number of vertexes and edges. In this sense, the following equations formally define the obtained knowledge graphs. Therefore, a concept c, that comprise a set of terms (n-grams) of similar meaning represents a vertex v at the co-occurrence network :

$$c = \{t_0, t_1, \cdots t_{k-1}\} \tag{10}$$

where t_i is the i-th term associated with concept c and k represents the total number of terms that denoted the same concept. The existence of both concepts in the same message denotes a relationship between the concept c_i and c_j, generating an edge in the knowledge network. To evaluate how frequently was this relationship (i.e., the edge relevance), Equation 11 measures how many times the concepts c_i and c_j were used together in the overall dataset D, being defined as follows:

$$co - occurrence(c_i, c_j) = \sum_{i=0}^{n}\sum_{j=i+1}^{n} f(i, j, D) \tag{11}$$

where $f(i, j, D)$ is defined as Equation 12:

$$f(i,j,D) = \begin{cases} 1, if\ \exists n\left[c_i\ \in Message_n \wedge\ c_j \in Message_n\right] \\ \qquad\qquad\qquad 0, otherwise \end{cases} \tag{12}$$

Similarly, the notion of user-interaction knowledge measures how frequently two users, u_i and u_j, were connected in a conversation. In this sense, each user u represents a vertex v in the user-interaction network, whereas a registered retweet $RT_{u_iu_j}$ or mention $MT_{u_iu_j}$ represents an edge denoting a relationship between them. To evaluate how frequently was this relationship (i.e., the edge relevance), Equation 13 measures how many times the user u_i and u_j were connected in the overall dataset D, being defined as follows:

$$user - interaction(u_i, u_j) = \sum_{i=0}^{n} \sum_{j=i+1}^{n} f(i,j,D) \tag{13}$$

where $f(i,j,D)$ is defined as:

$$f(i,j,D) = \begin{cases} 1, if\ \exists n\left[RT_{u_iu_j} \in Message_n\ or\ MT_{u_iu_j} \in Message_n\right] \\ \qquad\qquad\qquad 0, otherwise \end{cases} \tag{14}$$

where $RT_{u_iu_j}$ denotes a retweet from the user u_i to the user u_j and $MT_{u_iu_j}$ denotes a mention from the u_i to the user u_j in a message.

To compute the accumulated relevance of a vertex v (i.e. a concept c or a specific user u) in each knowledge network, the degree of centrality was used by measuring the total amount of edges that a vertex has with the other vertexes, being defined as follows:

$$D_c(v_i) = d_i \tag{15}$$

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

Similarly, in the case of directed graphs as the user knowledge network, the degree of centrality measures the sum of the out-degree, $deg^{+}(v)$, and the in-degree, $deg^{-}(v)$, begin defined as follows:

$$D_c(v_i) = \sum_{v \in V} deg^{+}(v) \sum_{v \in V} deg^{-}(v) \tag{16}$$

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

Finally, to represent the knowledge network in an intuitive way, the software Gephi [89] was used together with the circle pack layout (also called nested circle layout) to group vertexes based on their information (i.e., degree or cluster) in a hierarchical way.

5. Results

5.1 Community detection: Gluten-related communities

Conversations on social media create inherent communication graphs with identifiable contours as people replying and mentioning one another in their tweets. These conversational structures differ depending on the nature of the conversation, the interests, and the relevance of people participating in the discussion. In this sense, the representation of the different sub-communities and their interests and interactions allows gaining a better understanding of how information flows and who were the discussion leaders. Figure 3 illustrates the interactions between the different communities of users clustered by the content of their Twitter messages.

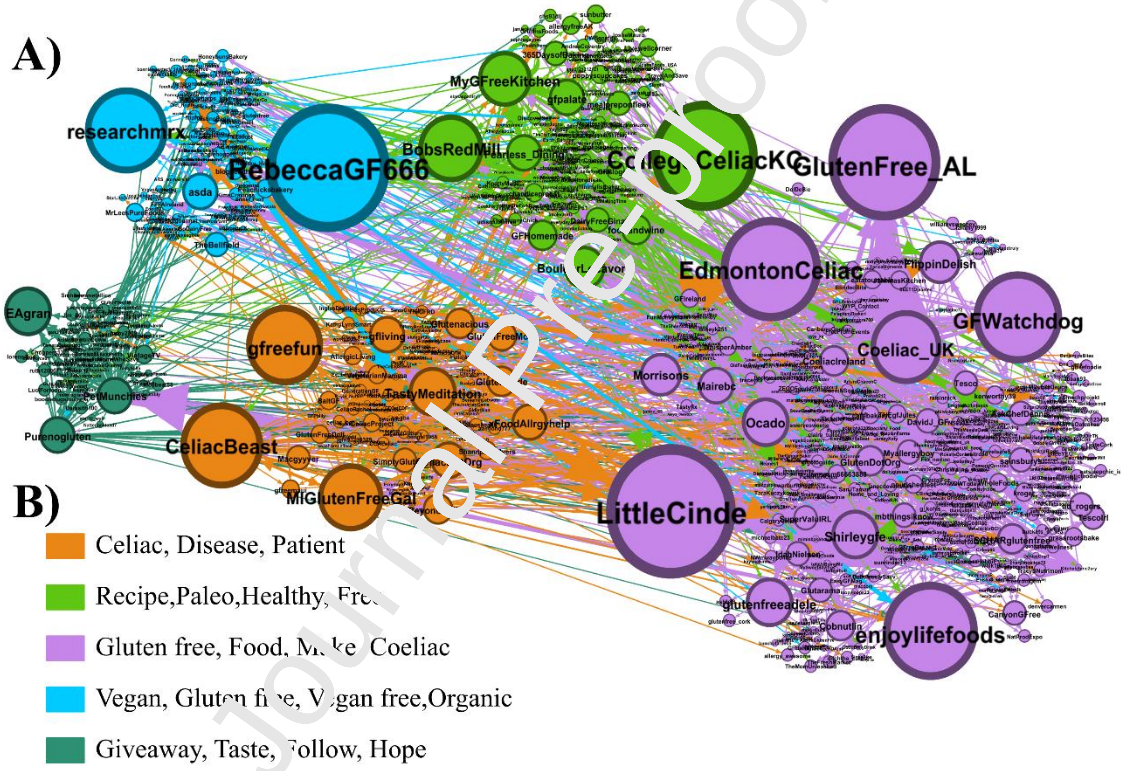


Figure 3. (A) Graph showing the detected communities based on the content of their Twitter messages. The vertex colour represents each community, whereas the edges represent the interactions between them (i.e., mentions and retweets). The vertex size is based on the in-degree metric, whilst the edge size stands for the number of retweets or mentions between a pair of users, and the edge colour represents the source cluster. (B) Illustrate the main topics of each community.

Figure 3A depicts the discovered interactions between the different users constructed with users that had more than ten interactions between them. This graph contains 910 vertexes (i.e., unique users) and 1,341 edges (i.e., retweets and mentions). Edges were weighted based on the number of retweets and mentions between a source and a target user, whereas the vertex size is based on their in-degree. The colour of the vertexes represents the community where each user is classified based on the content of their Twitter messages. Figure 3B details the main words of each cluster.

Attending to the interactions between users, the purple community contains the largest number of users (47.5% - 427 different users). This community had a more general character (i.e., more different topics were talked about) but mainly discussed gluten-free food and food intolerances. The next largest group was the light green community (17.8% - 160 different users), and their conversations were mainly focused on topics related to recipes, foods and diets, such as the Keto diet, the Paleo diet, the vegetarian diet and free diets (e.g., dairy-free, gluten-free, low carb). Below these, there was the blue community (13.4% - 118 different users) related to topics such as vegan lifestyle, vegan diets, healthy or organic food and recipes. Very close was the orange community (12.2% - 110 different users) focused on people concerns and centred into news, research and health-related issues such as disease, diet or gluten sensitivity. On the last position, it was the dark green community (9.24% - 83 different users), whose central themes were mostly related to a commercial aspect. Overall, the dark green or "commercial" community was the one that received fewer interactions, since their Twitter messages were more focused on advertising, whereas the purple or "general" group was the community that attracts more attention because it accumulated more knowledge by having a broader population of users who talked about more diverse topics.

5.2 Demographic analysis

5.2.1 Geographical distribution of gluten-related communities

In the present investigation, 164,596 out of the 317,184 individual users that had indicated a location (51% of individuals with some kind of location provided) were geographically distributed around the world. In general, most of the users were located in the United States, the United Kingdom, Canada, and Australia, whereas the bigger percentage of patient ratio (Patients detected/country community users) identified were located in Finland, Canada, France and Australia. The highest reported gluten-related communities in terms of the continent was America, Asia and Europe, whereas the more significant percentage of patient ratio detected were located in America, Oceania and Europe. In general, the user location and the number of patients obtained were consistent with current knowledge about the prevalence of the CD in the principal regions of the world [90]. However, the case of Canada should be highlighted, since, in recent years, they have not been identified as countries with a high incidence of the population with gluten-related diseases, but the previous research studies suggest that there is some delay in the diagnosis of CD [91].

5.2.2 Gender analysis of biomedical topics

Table 1 summarizes the more significant differences between topics exposed by males and females. For the construction of the table, only terms with a volume of difference between genders greater than 40 users were included and a perceptual difference between genders greater than 10%. %F is the percentage of females who have mentioned the term in some of their Twitter messages, whereas %M is the percentage of males who have mentioned the term in some of their Twitter messages. #F is the total females who have mentioned the term in some of their Twitter messages, whereas #M is the total males who have mentioned the term in some of their Twitter messages. #F-#M is the difference between the number of males and females who have mentioned the term. Finally, %F-%M is the difference percentage between men and women who have mentioned the term in some of their Twitter messages (i.e., bigger positive numbers represent a greater number of females that mentioned the term whereas bigger negatives numbers represent a greater number of males that mentioned the term).

Table 1. Top differentiating terms between females and males in gluten-related Twitter messages. **F** represents females and **M** represents males.

Annotated term	**Semantic category**	**%F**	**%M**	**#F**	**#M**	**#F-#M**	**%F-%M**
Pie crust	Food & Nutrition	80.5	19.5	190	46	144	61.0
Lip	Anatomy	77.1	22.9	162	48	114	54.3
Coffee cake	Food & Nutrition	76.1	23.9	150	47	103	52.3
Paraben	Drug & Chemical compounds	76.1	23.9	220	69	151	52.2
Small intestine	Anatomy	74.5	25.5	373	128	245	48.9
Immune system	Anatomy	74.2	25.8	480	167	313	48.4
Brown Sugar	Food & Nutrition	73.5	26.5	150	54	96	47.1
Appetizer	Food & Nutrition	73.3	26.7	302	110	192	46.6
Applesauce	Food & Nutrition	72.2	27.8	104	40	64	44.4
lavender oil	Compound	72.0	28.0	139	54	85	44.0
influenza	Symptom	72.0	28.0	113	44	69	43.9
Migraine	Symptom	71.9	28.1	383	150	233	43.7
Peppermint	Food & Nutrition	71.7	28.3	180	71	109	43.4
Codein	Drug & Chemical compounds	71.3	28.7	201	81	120	42.6
Nausea	Symptom	70.9	29.1	100	41	59	41.8
Beer	Food & Nutrition	43.0	57.0	1819	2408	-589	-13.9
Stout	Food & Nutrition	36.3	63.8	58	102	-44	-27.5
Ale	Food & Nutrition	30.0	70.0	133	311	-178	-40.1

In general, males talked in a higher percentage than females about alcoholic beverages such as Ale, Beer or Stout, whereas females discuss in a higher rate than males about: (*i*) related beauty products such as "paraben" and "lavender oil"; (*ii*) foods and food elaboration such as "pie crust", "coffee cake" or "brown Sugar"; and (*iii*) symptoms such as "migraines" and "nauseas". Regarding the interest of men in gluten-free beer, this study is in line with the market analysis that suggests that there is increasing consumer demand for gluten-free beer, which can be a profitable niche market for multinational breweries and an opportunity for arable farmers to increase alternative grain production [92,93] On the other hand, the association of beauty products and gluten suggests that campaigns were raising by small cosmetic producers, with a greater female audience, taking advantage of the healthy appearance of gluten-free products to obtain market profitability. In this respect, there is no clear evidence that the application of gluten-containing beauty products are harmful when there are no allergies and when there is no contact with the oral pathways [94]. Concerning the obtained difference of percentage of males and females that mentioned the term "migraines", this outcome was in line with different scientific studies reporting that the prevalence of migraines in individuals with CD is much higher in females than in males [95,96]. Attending to the overall annotated semantic categories, females have talked in a higher percentage than males about related symptoms and diseases (around 23%). This data complies with different research studies and reports that claim that women tend to search more online health information and tend to have fewer issues sharing personal and health problems, whereas men can be inclined to "just get on with it" (i.e., "man up") and ignore their problems [54,97–100].

5.3 Biomedical topics interactions exposed by the community

4.3.1 Biomedical topics interactions exposed by individuals

Understanding the semantic interactions among the main discussed topics of individual conversations and their related emotions can help to understand the current state of gluten-related diseases and improve public health surveillance based on the public *sociome* discussion. In this sense, Figure 4 presents a knowledge graph composed of a total of 2,836 vertexes (unique concepts) and 64,226 edges depicting how the different domain topics were interconnected in the conversations of individuals. Explicit mentions to celiac (e.g., "celiac sprue" or "Coeliac") or gluten-related disease (e.g., "Gluten intolerance" or "Gluten allergy"), and non-content bearing, generalist terms (e.g., "Gluten" or "Food"), were removed to show a more readable graph and be able to discover more relevant information by eliminating information that was a priori trivial (e.g., Gluten is related with celiac). The colour of the vertexes represents the semantic category of the concept (i.e., red stands for "Drug & Chemical compounds", orange represents "Diets", light green relates to "Food & Nutrition", dark green relates to "Diseases", light blue stands for "Anatomy", dark blue relates to "Physical activity", purple represents "Symptoms", light brown represents "Plants" and yellow relates to "Dietary Supplements"). In contrast, the edge colour stands for the overall sentiment associated, i.e., green indicates a positive sentiment, grey represents neutral sentiment and red stands for negative sentiment. The vertex size is based on the degree of the term (i.e., bigger vertexes represent terms that have a greater number of associations), whereas the edge size was calculated based on the users that use both terms at the same Twitter message (i.e., thicker edges represent a stronger association between two vertexes).

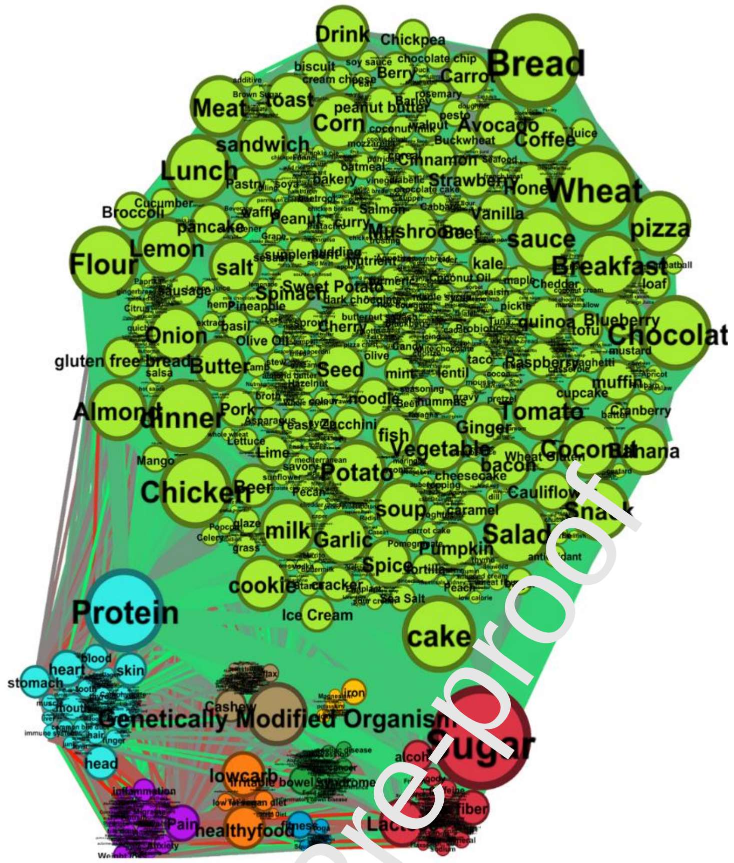


Figure 4. Knowledge graph representing the co-mention annotated terms. The size of the vertexes is based on the number of users mentioning the term, while vertex colour represents the corresponding semantic category. The edge colour indicates the sentiment of the majority of the associated tweets, and the edge size represents the number of users that use both terms in the same Twitter message.

Regarding the overall categories of the graph, “Food & Nutrition” and “Compounds” were the semantic categories with the highest number of unique concepts (i.e., 1,737 and 308, respectively). On the other hand, the most relevant terms of each category were, in order of relevance: “bread” (Food & Nutrition). “sugar” (Drug & Chemical compounds), “protein” (Anatomy), “genetically modified organism” (Plants), “lowcarb” (Diet), pain (Symptom), “inflammatory bowel disease” (Disease), “iron” (Dietary Supplements) and “fitness” (Physical activity). For a more exhaustive analysis of this semantic knowledge graph, a sentiment-degree filter was applied. Figure 5 shows the most relevant word interactions with positive sentiment (i.e., the bigger number of users talking about both terms expressing positive emotions) and the most relevant word interactions with a negative sentiment.

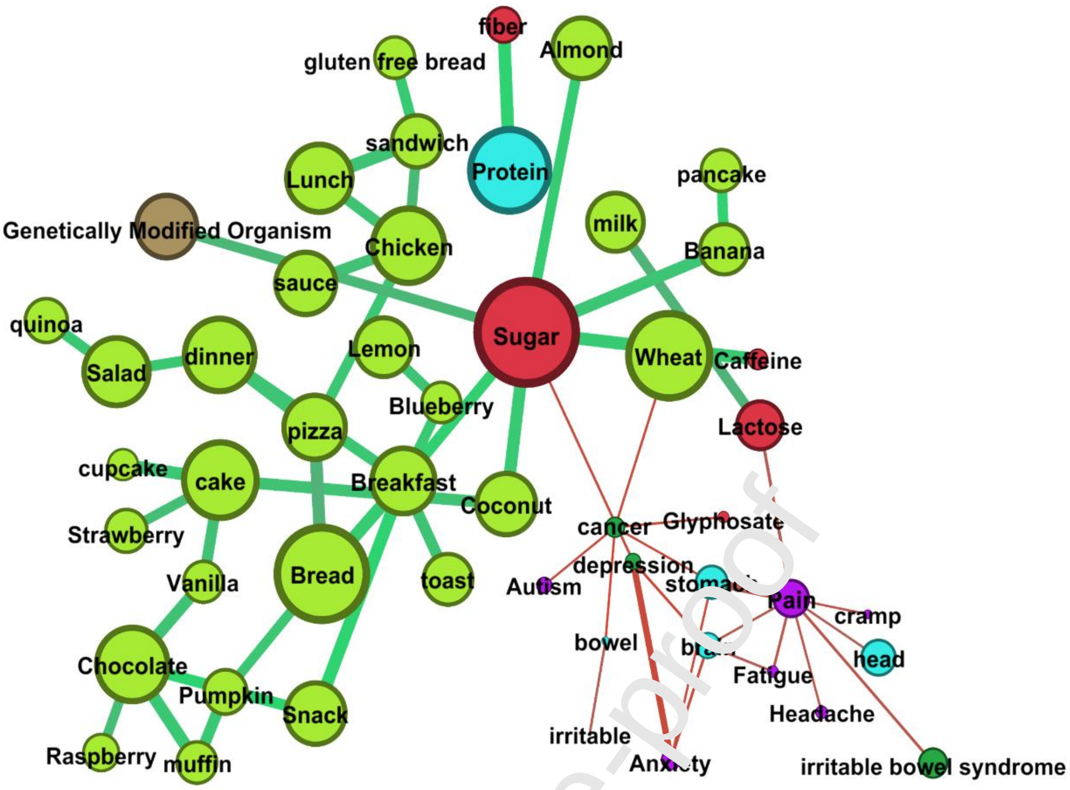
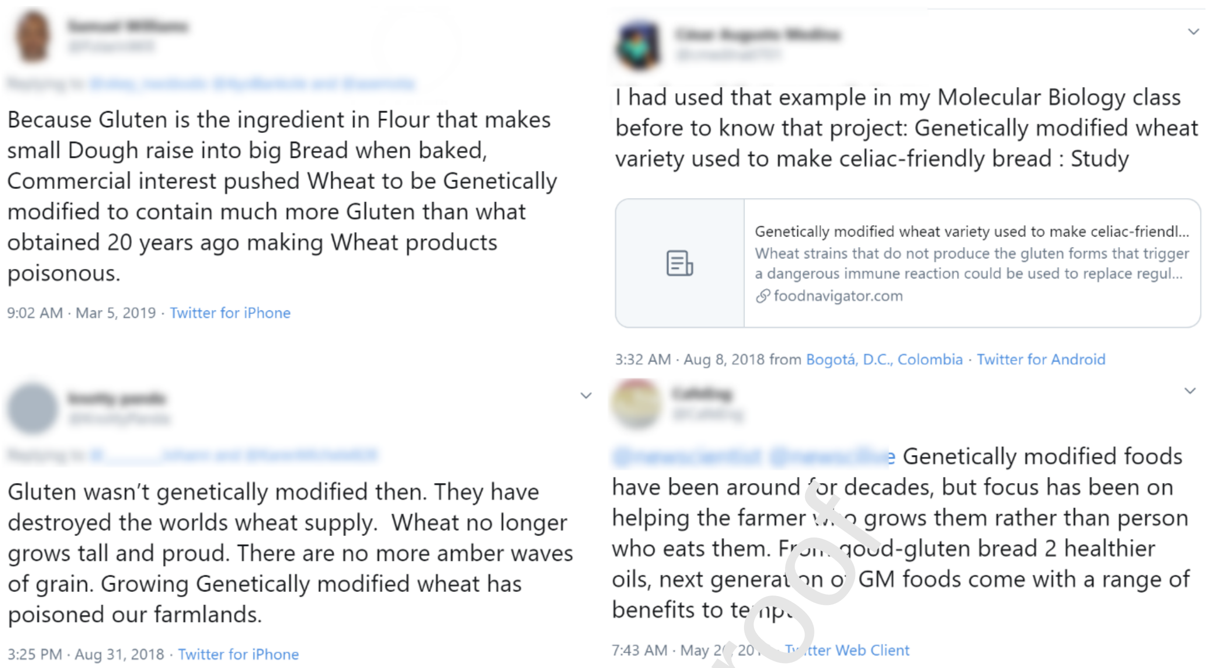


Figure 5. Sub-graph representing the top positive and negative co-mention annotated terms. The size of the vertexes is based on the number of users mentioning the term, while vertex colour represents the corresponding semantic category. The edge colour indicates the sentiment of the majority of the associated tweets, and the edge size represents the number of users that use both terms in the same Twitter message.

Regarding the interactions related to positive emotions, Food & Nutrition stands out over the other categories due to the tweeting of a large number of healthy recipes, i.e., low in sugar, gluten-free and often lactose-free. Another term that stands out among the positive interactions was “genetically modified organisms” (GMO) due to some users put their hopes in that it could be the solution in the creation of new free gluten products. However, this was a controversial topic because other users discuss GMO as the cause of gluten allergies (Figure 6). In this sense, and taking into account the importance that GMO foods may have in the future diet of people who are allergic or intolerant to gluten, it is necessary to start raising patient awareness campaigns from recognised sources of information about the benefits and absence of evidence of any health hazards [101,102].

To conclude, the final relevant positive interaction was “protein” and “fiber”, explained by a high number of Twitter messages containing gluten-free recipes associated with a healthy lifestyle and other food trending topics as “superfood” or “lowcarb”.



Because Gluten is the ingredient in Flour that makes small Dough raise into big Bread when baked, Commercial interest pushed Wheat to be Genetically modified to contain much more Gluten than what obtained 20 years ago making Wheat products poisonous.

9:02 AM · Mar 5, 2019 · Twitter for iPhone

Gluten wasn't genetically modified then. They have destroyed the worlds wheat supply. Wheat no longer grows tall and proud. There are no more amber waves of grain. Growing Genetically modified wheat has poisoned our farmlands.

3:25 PM · Aug 31, 2018 · Twitter for iPhone

I had used that example in my Molecular Biology class before to know that project: Genetically modified wheat variety used to make celiac-friendly bread : Study

Genetically modified wheat variety used to make celiac-friendl...
Wheat strains that do not produce the gluten forms that trigger a dangerous immune reaction could be used to replace regul...
foodnavigator.com

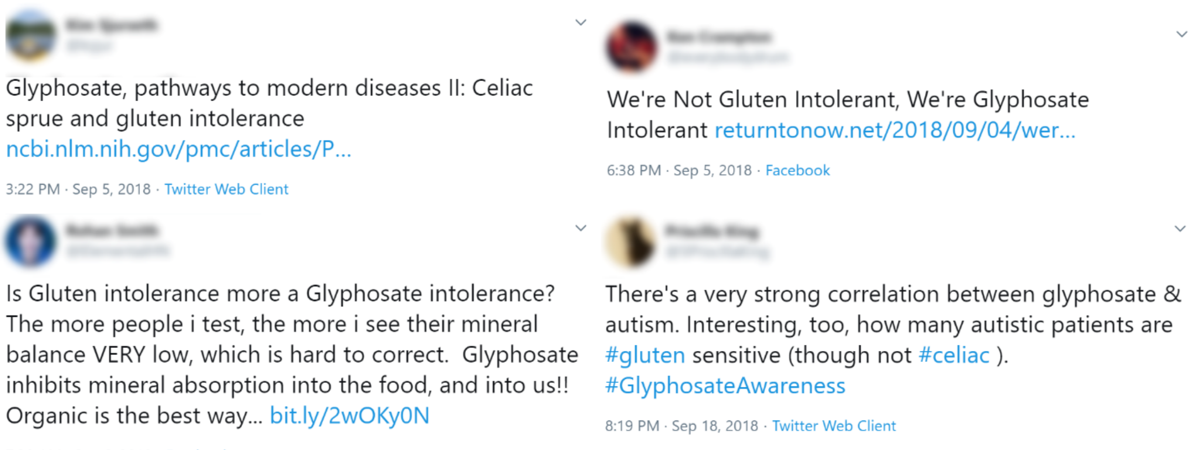
3:32 AM · Aug 8, 2018 from Bogotá, D.C., Colombia · Twitter for Android

Genetically modified foods have been around for decades, but focus has been on helping the farmer who grows them rather than person who eats them. From good-gluten bread 2 healthier oils, next generation of GM foods come with a range of benefits to tempt

7:43 AM · May 2[illegible] 201[illegible] · Twitter Web Client

Figure 6. Example gluten-related Twitter messages that show the current debate in the community concerning genetically modified organisms (GMO). On the left, opinions against them and on the right opinions that support them.

Regarding the interactions related to negative emotions, symptoms stand out over the other categories. In this sense, there were groups of symptoms related to gluten-related diseases that were supported by science such as, "headache" and "stomach pain", and another group of medical issues that had a lack of clear evidence regarding gluten or gluten-free diets, such as "autism" or "depression" [103,104]. Finally, regarding the compounds, "Glyphosate" was shown as one of the compounds that most appeared to be related to negative sentiments. This semantic connection happens due to the frequent Twitter messages sharing news in the community, associating the compound glyphosate to CD (Figure 7). This relationship arises due to the result of a study published in 2013, which says that certain environmental factors may be involved in the development of inflammatory bowel diseases and some compound as glyphosate causes intestinal alterations in animals with a similar characteristic of gluten allergies [105]. This research article was misrepresented to suggest that glyphosate is the cause of CD when the relation between them has not been proven based on experimental evidence [106].



[redacted]

Glyphosate, pathways to modern diseases II: Celiac sprue and gluten intolerance ncbi.nlm.nih.gov/pmc/articles/P...

3:22 PM · Sep 5, 2018 · Twitter Web Client

[redacted]

We're Not Gluten Intolerant, We're Glyphosate Intolerant returntonow.net/2018/09/04/wer...

6:38 PM · Sep 5, 2018 · Facebook

[redacted]

Is Gluten intolerance more a Glyphosate intolerance? The more people i test, the more i see their mineral balance VERY low, which is hard to correct. Glyphosate inhibits mineral absorption into the food, and into us!! Organic is the best way... bit.ly/2wOKy0N

[redacted]

There's a very strong correlation between glyphosate & autism. Interesting, too, how many autistic patients are #gluten sensitive (though not #celiac).
#GlyphosateAwareness

8:19 PM · Sep 18, 2018 · Twitter Web Client

Figure 7. Example of gluten-related tweets claiming that glyphosate is a cause of some diseases as autism or CD.

4.3.2 Community comparison: Biomedical topics interactions exposed by patients

By comparing the conversations of all individuals against the discussion of patients, it was possible to detect which concepts were more important in the patient community and which were less relevant to patients and may be overemphasised due to the fads or the beliefs of the general community. In this sense, Figure 8 shows the knowledge graph obtained exclusively from the patient Twitter messages (left) and the knowledge graph obtained from the processing of the whole gluten community (right).

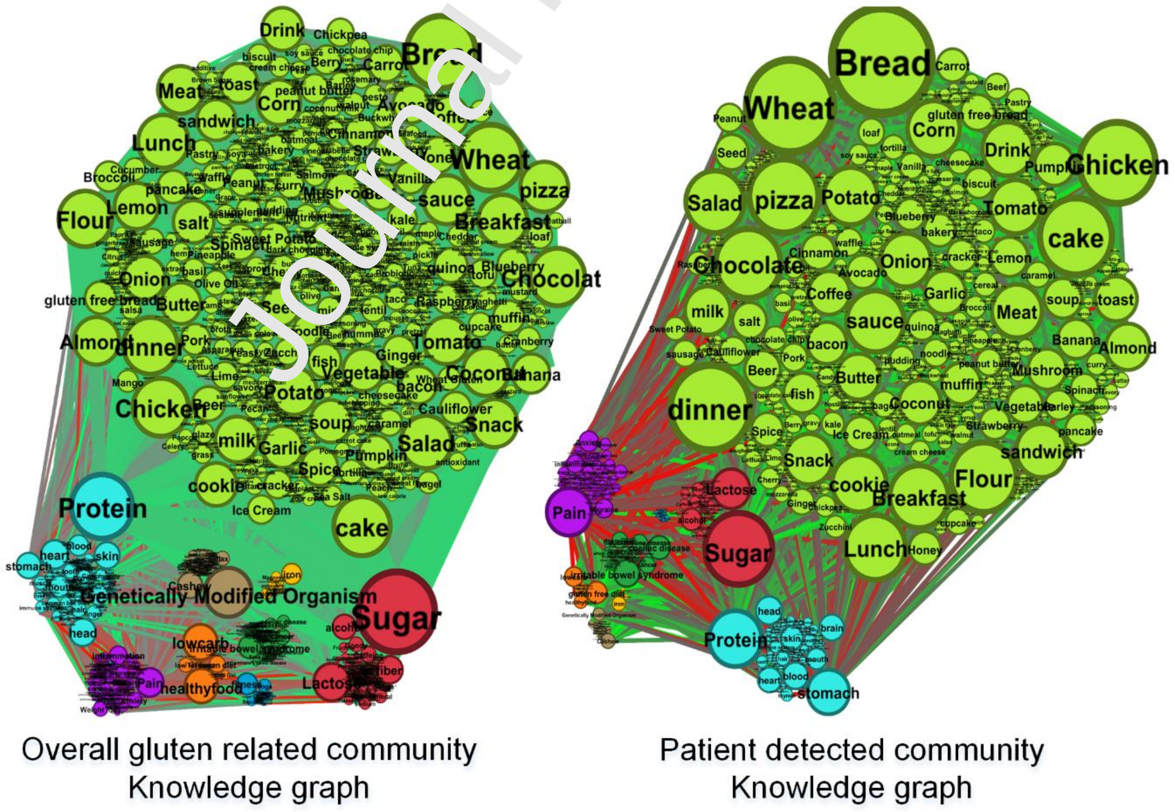


Figure 8. Knowledge graphs representing the co-mention annotated terms at the general gluten-related community (left) and the patient community (right). The size of the vertexes is based on the number of users mentioning the term, while vertex colour represents the corresponding semantic category. The edge colour indicates the sentiment of the majority of the associated tweets, and the edge size represents the number of users that use both terms in the same Twitter message.

Although, at first glance, both graphs were very similar, there were significant differences between them. In general terms, domain topics like symptoms and diseases acquired greater relevance in the patient Twitter messages, whereas topics as plants, sports and diets had a more significant impact on the general community. Focusing on specific terms, the volume of Twitter messages talking about "genetically modified organisms" and "vegan diets" was proportionally higher in the whole community of individuals than in the case of patients (i.e., the topological coefficient of vertexes it was more significant in the gluten-related knowledge graph). On the other hand, the volume of Twitter messages talking about "depression" was more significant in the patient conversations than in the general discussions of individuals. In contrast, the number of Twitter messages that talked about "autism", "healthy food", and "lowcarb" was less representative in the patient conversations. This evidences that a relevant proportion of the general community messages not associated with patients were related to (*i*) users that perceive that the GFD promotes improved public health, (*ii*) users that combined or associated the GFD with other diets like the low-carb diet to have a healthier lifestyle, and (*iii*) users that associated GFD with other diseases.

5.4 External resources analysis

511,040 different URLs from the whole dataset of tweets, that is, around 42% of all Twitter messages had almost one link, and at least 29% of all users had ever shared an URL. Considering the availability of the URLs, 263,432 of all shared URLs (51%) were not accessible at the time of this study, or it was not possible to retrieve any information (e.g., Amazon or Facebook). Table 2 summarizes the top 10 resource domains ordered by the total number of "individual" accounts that share the resource, besides the top 10 resource domains that reached the most significant number of community interactions (retweets and favourites).

Table 2. The top 10 resource domains sorted by the total number of "individual" accounts that share the resource besides the top 10 resource domains that reach the most significant number of community interactions (retweets and favourites).

Top 10 shared domains	instagram.com, facebook.com, amazon.com, youtube.com, pinterest.com, celiac.com, glutenfreeworks.com, simplygluten-free.com, yoursecretrecipes.com, purewow.com
Top 10 domains with more interactions (retweets and favourites)	instagram.com, youtube.com, minimalistbaker.com, kitchensanctuary.com, julianbakery.com, arbonne.com, facebook.com, ketodietapp.com, theonion.com, nowandgen.com, amazon.com, glutenfreeliving.com

As shown in Table 2, the gluten-free community on Twitter had a great connection with other popular social media platforms as Instagram or Facebook. This interconnection is motivated by the importance and impact that social media influencers have on the gluten-free community [107]. The second place of importance were the websites related to recipes, blogs about healthy foods, and electronic e-commerce. In the last level of incidence, there were links to more formal resources of information or with a greater degree of reliability, such as public news channels or links to scientific articles.

To analyse the overall objective of all shared resources, Figure 9 summarizes the main website clusters obtained through the processing of their retrieved content (excluding other social media platforms and Amazon.com).

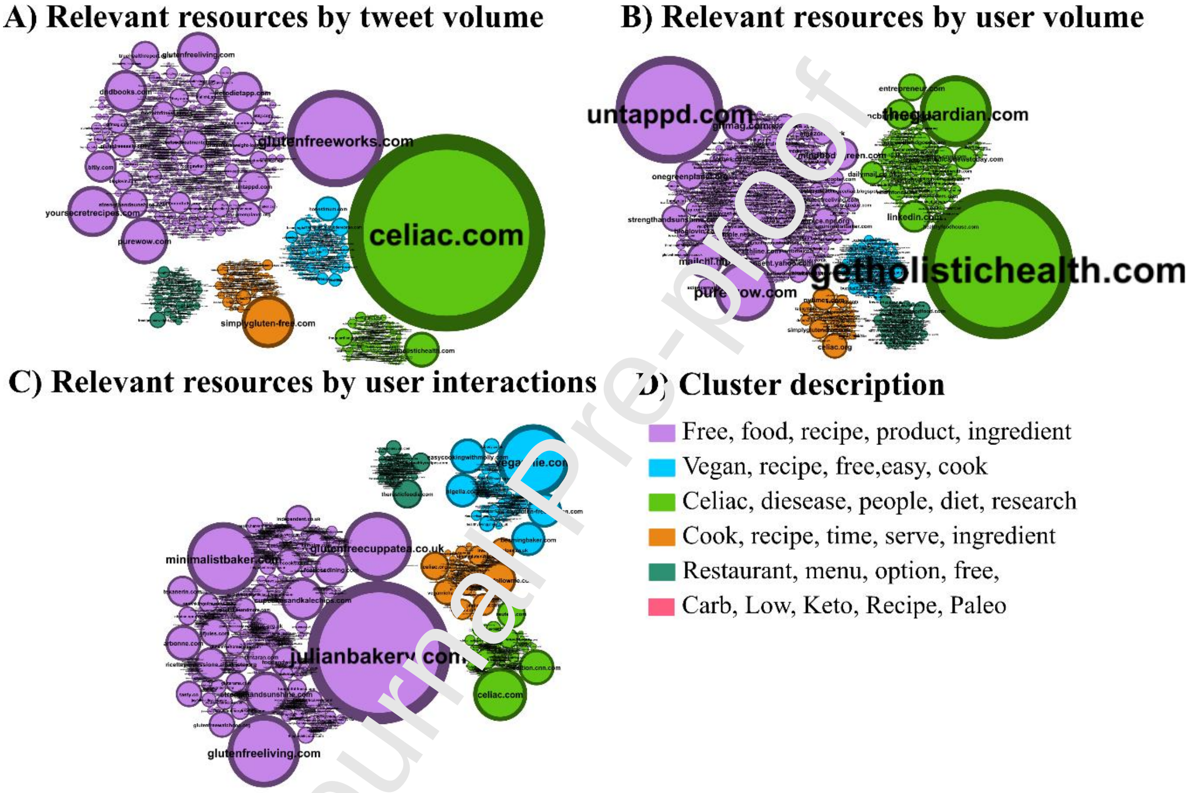


Figure 9. (A) Represents the top shared resources sized by the number of Twitter messages that contain each resource. (B) Depicts the top shared resources sized by the number of users that share each resource. (C) Denotes the top shared resources taking in to account the number of interactions reached, i.e., retweets and favourites. (D) Illustrates the main topics of each website category.

Figure 9A shows the top shared resources sized by the number of Twitter messages containing each resource (i.e., the number of times that link was shared) and gives information about the most advertised websites. On the other hand, Figure 9B depicts the top shared resources that were relevant to different users. Finally, Figure 9C provides information about the most influential resources taking into account the number of interactions reached (retweets and favourites). Joining the general knowledge offered by the different clusters, it can be said that the most relevant cluster category was related to recipes since they had the most significant relevance at the different clusters and reaches a high volume of favourites and retweets. Contrarily, resources associated with CD and other medical topics were contained in a large number of Twitter messages and also shared by many different users, but they were not highly shared with retweets and favourites. Finally, links related to veganism had the highest proportion of interactions concerning the volume of information that was shared within the community (i.e., their impact, retweets and favourites, were relatively high in proportion to the low volume of Twitter messages or users who have shared them).

6. Discussion

6.1 Principal findings

The developed methodology in this study demonstrates how social media platforms could be an excellent complementary source of information to study the current state of health issues and dietary concerns. Results obtained by the demographic analysis of the community and the characterisation of the individuals revealed how the number of patients in social media platforms followed a similar distribution by country established by previous studies. Besides, this study discovered countries where the public discussion among patients related to CD was growing (i.e., Canada). On the other hand, the gender analysis and the characterisation of the individuals provided symptomatology differences between males and females, such as the incidence of migraines in females. Taking into account the overall inferred knowledge presented in results, it was noted that the community of non-patient individuals usually related GFD with the idea of a healthier lifestyle or healthy food, and it was typically associated with the culture of physical exercise as a lifestyle. On the other hand, as can be seen in Figure 8, the patient community discussed in greater volume negative issues related to symptoms and diseases, unlike other thematic as diets. Besides, the mention of terms related to the topic "Plants" was significantly lower inside the patient community than outside. This fact, in combination with the impact denoted in Figure 9 (i.e., retweets and favourites) by the vegetarian shared resources, evidenced how relevant was gluten-free diets within the world of veganism and a vegetarian diet as a lifestyle. In this sense, these types of diets need to be taken into account in the development of new information campaigns about the possible nutritional risks and dietary unbalances that a self-prescription of GFD could be in combination with other dietary lifestyles. In the same way, it was possible to observe how the resources that refer to reliable sources of information were kept at a lower level of relevance, having a lower impact within the gluten community and, therefore, less outreach.

These results denoted the importance of being constant in the development of health management campaigns and the relevance that health organisations and stakeholders had in supporting truthful information and countering misinformation (such as the case of glyphosate or autism covered in this study). Our findings suggest that the lack of reliable information about gluten-free diets, GMO and gluten-related diseases continues to lead to or reinforce beliefs that result in a detriment to public health. In this sense, the community analysis, in conjunction with the study of shared resources, revealed how the collaboration with domain influencers and their support to health awareness campaigns could be a useful tool to reach a more significant number of social platforms and, therefore, gain access to a more substantial number of individuals.

6.2 Limitations

Not all Twitter users who follow a GFD will necessarily write tweets about their lifestyle or health concerns. Besides, the number of available tweets downloaded from the Twitter API is limited, with no assurance of a random or representative sample [46]. For this reason, it was not possible to perform an analysis of a larger volume of data. Thus, a full data retrieval through automated dashboard vendors or using a paid service of the Twitter API, may provide further insights. Therefore, the veracity of the data written by the users on social media platforms may not be authentically real. However, there is an increasing number of patients turning to popular social media sites to share illness experiences or seek advice from others with similar health conditions [108]. We also emphasize that the amount of data that can be collected from these studies is significantly higher than alternative and complementary methods, such as surveys. Remark also that this study was only focused on English messages from the Twitter social platform, and it did not take into account individuals under the Twitter age restrictions (i.e., 13 years). Finally, it should be noted that if the proposed techniques are extended to support a greater variety of languages, like Chinese or Spanish, as well as other social media platforms, like Facebook or Instagram, then it may be possible to provide complementary findings. However, the public data access to these other social media platforms is greatly limited [109,110].

7. Conclusions

This work presents a new methodology to process, classify, visualise and analyse the big data knowledge produced by the *sociome* on social media platforms. The practical relevance of the proposed big data analysis methodology has been proven in the study of 1.1 million unique gluten-related messages from more than 400,000 distinct users. The work is centred on the research of one of the most popular dietary fads that evolve into a multibillion-dollar industry, the GFD. Besides, this work explores one of the least reported research fields studied on Twitter concerning public health, i.e., the allergies or the immunology diseases as the CD. In this sense, even though our case study was focused on gluten-related tweets, the proposed methodology has the potential to be applied to a more general field of study.

Consequently, this real-world case study exemplifies the broad range of non-trivial and practical knowledge that the proposed methodology can gather. This methodology has shown how the use of a standard domain vocabulary, the design of effective *sociome* profiling strategies, and the usage of different classification and visualization techniques could enhance the study of the high volume of data produced on social media platforms to improve health surveillance. NLP, NER and ML techniques were used to classify the user profiles and their messages. In contrast, complementary methods as the reconstruction of knowledge graphs and the application of clustering methods enabled a holistic, multi-layered analysis to acquire new knowledge, looking into different levels of detail and perspective views.

Regarding the outcomes of the different techniques applied, the current methodology has proven to achieve an adequate performance to identify: (*i*) individuals (with an F-score of 0.84); (*ii*) patients (with an F-score of 0.89); (*iii*) the user gender (with an F-score of 0.86); and (*iv*) the user location (with an F-score of 0.95). With reference to the principal findings, the outcomes obtained from the semantic analysis carried out, combined with the reconstruction of knowledge graphs, pointed out a wide range of health-related conclusions. Considering the proposed community analysis, this study illustrates how public health organisations may find social media a valuable tool to obtain health-related information and raise health awareness campaigns to targeted populations. In this regard, the application of the present methodology could help to identify the most suitable community members to support a campaign, taking into account their public discourse.

To conclude, even though social media data analysis produces insights more rapidly, cheaply, and gives a complete picture of the public attitude, traditional methods like surveys usually provide higher quality, targeted, and relevant health outcomes. Conversely, online conversations in the social media patient communities centred on health behaviours probably provide more nuanced and realistic information about health-related attitudes and beliefs than traditional survey measure due to the relaxed and informal nature of conversations. So, science is more likely to benefit by combining both modes of analysis to understand and explain changes in individual and collective behaviour

Finally, although part of the methodologies developed in this study is language-independent, future work will be centred on integrating different language ontologies to process more messages and to be able to analyse a larger volume of the world population

8. Acknowledgements

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

This work was supported by: the Associate Laboratory for Green Chemistry-LAQV, which is financed by national funds from and the Portuguese Foundation for Science and Technology(FCT) under the scope of the strategic funding of [UID/QUI/50006/2019] and [UIDB/04469/2020] units, and BioTecNorte operation [NORTE-01-0145-FEDER-000004] funded by the European Regional Development Fund under the scope of Norte2020 - Programa Operacional Regional do Norte, 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 postdoctoral fellowship [ED481B-2019-032] of Martín Pérez-Pérez, funded by the Xunta de Galicia.

9. Conflict of Interest

None declared.

10. Abbreviations

- Gluten-free diet (GFD)
- Global Positioning System (GPS)

- HyperText Markup Language (HTML)
- Machine learning (ML)
- Name entity recognition (NER)
- Natural language processing (NLP)
- Part of speech (POS)
- Term frequency – Inverse document frequency (TF-IDF)
- Text mining (TM)
- Uniform Resource Locators (URL)
- Celiac Disease (CD)

11. References

[1] Lomet DB. The Future of Data Management. Computer (Long Beach Calif) 2017;50:12–3. https://doi.org/10.1109/MC.2017.3641630.

[2] Carnevale L, Celesti A, Galletta A, Dustdar S, Villari M. Osmotic computing as a distributed multi-agent system: The Body Area Network scenario. Internet of Things 2019;5:130–9. https://doi.org/10.1016/j.iot.2019.01.001.

[3] Lv Z, Song H, Basanta-Val P, Steed A, Jo M. Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics. IEEE Trans Ind Informatics 2017;13:1891–9. https://doi.org/10.1109/TII.2017.2650204.

[4] Roser M, Ritchie H, Ortiz-Ospina E. Internet. Our World Data 2015.

[5] Kaya M, Kawash J, Khoury S, Day M-Y. Social Network Based Big Data Analysis and Applications 2018:0–22. https://doi.org/10.1007/978-3-319-78196-9.

[6] Gabarron E, Dorronzoro E, Rivera-Romero O, Wynn R. Diabetes on Twitter: A Sentiment Analysis. J Diabetes Sci Technol 2019;13:439–44. https://doi.org/10.1177/1932296818811679.

[7] Jeong B, Yoon J, Lee J-M. Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. Int J Inf Manage 2019;48:280–90. https://doi.org/10.1016/J.IJINFOMGT.2017.09.009.

[8] Antheunis ML, Tates K, Nieboer TE. Patients' and health professionals' use of social media in health care: Motives, barriers and expectations. Patient Educ Couns 2013;92:426–31. https://doi.org/10.1016/j.pec.2013.06.020.

[9] Corea F. Can Twitter Proxy the Investors' Sentiment? The Case for the Technology Sector. Big Data Res 2016;4:70–4. https://doi.org/10.1016/j.bdr.2016.05.001.

[10] Valdez RS, Brennan PF. Exploring patients' health information communication practices with social network members as a foundation for consumer health IT design. Int J Med Inform 2015;84:363–74. https://doi.org/10.1016/j.ijmedinf.2015.01.014.

[11] Rozenblum R, Greaves F, Bates DW. The role of social media around patient experience and engagement. BMJ Qual Saf 2017;26:845–8. https://doi.org/10.1136/bmjqs-2017-006457.

[12] Smailhodzic E, Hooijsma W, Boonstra A, Langley DJ. Social media use in healthcare: A systematic review of effects on patients and on their relationship with healthcare professionals. BMC Health Serv Res 2016;16:442. https://doi.org/10.1186/s12913-016-1691-0.

[13] Turner PG, Lefevre CE. Instagram use is linked to increased symptoms of orthorexia nervosa. Eat Weight Disord 2017;22:277–84. https://doi.org/10.1007/s40519-017-0364-2.

[14] Balmau O, Guerraoui R, Kermarrec AM, Maurer A, Pavlovic M, Zwaenepoel W. The fake news vaccine: A content-agnostic system for preventing fake news from becoming viral. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11704 LNCS, Springer; 2019, p. 347–64. https://doi.org/10.1007/978-3-030-31277-0_23.

[15] So J, Prestin A, Lee L, Wang Y, Yen J, Chou WYS. What Do People Like to "Share" About Obesity? A Content Analysis of Frequent Retweets About Obesity on Twitter. Health Commun 2016;31:193–206. https://doi.org/10.1080/10410236.2014.940675.

[16] Injadat MN, Salo F, Nassif AB. Data mining techniques in social media: A survey. Neurocomputing 2016;214:654–70. https://doi.org/10.1016/j.neucom.2016.06.045.

[17] Felt M. Social media and the social sciences: How researchers employ Big Data analytics. Big Data Soc 2016;3:205395171664582. https://doi.org/10.1177/2053951716645828.

[18] Pelovitz BRM. The 2014 2014:18–24. https://blog.twitter.com/official/en_us/a/2014/the-2014-yearontwitter.html (accessed January 28, 2020).

[19] Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, Merchant RM. Twitter as a tool for health research: A systematic review. Am J Public Health 2017;107:e1–8. https://doi.org/10.2105/AJPH.2016.303512.

[20] Bhoi A, Balabantaray RC. Named Entity Recognition from Social Media Text : A Comparative Study. Int J Control Theory Appl 2017;10:9–15.

[21] Ajao O, Hong J, Liu W. A survey of location inference techniques on Twitter. J Inf Sci 2015;41:855–64. https://doi.org/10.1177/0165551515602847.

[22] Bokunewicz JF, Shulman J. Influencer identification in Twitter networks of destination marketing organizations. J Hosp Tour Technol 2017;8:205–19. https://doi.org/10.1108/JHTT-09-2016-0057.

[23] Park H, Reber BH, Chon MG. Tweeting as health communication: Health organizations use of twitter for health promotion and public engagement. J Health Commun 2016;21:188–98. https://doi.org/10.1080/10810730.2015.1058435.

[24] Ikeda K, Hattori G, Ono C, Asoh H, Higashino T. Twitter user profiling based on text and community mining for market analysis. Knowledge-Based Syst 2013;51:35–47. https://doi.org/10.1016/j.knosys.2013.06.020.

[25] Conrad EJ, Becker M, Powell B, Hall KC. Improving Health Promotion Through the Integration of Technology, Crowdsourcing, and Social Media. Health Promot Pract 2020;21:228–37. https://doi.org/10.1177/1524839918811152.

[26] O’reilly M, Dogra N, Hughes J, Reilly P, George R, Whiteman N. Potential of social media in promoting mental health in adolescents. Health Promot Int 2019;34:981–91. https://doi.org/10.1093/heapro/day056.

[27] Park A, Conway M. Tracking Health Related Discussions on Reddit for Public Health Applications. AMIA . Annu Symp Proceedings AMIA Symp 2017;2017:1362–71.

[28] Karami A, Dahl AA, Turner-McGrievy G, Kharrazi H, Shaw G. Characterizing diabetes, diet, exercise, and obesity comments on Twitter. Int J Inf Manage 2018;38:1–6. https://doi.org/10.1016/J.IJINFOMGT.2017.08.002.

[29] Lenzi A, Maranghi M, Stilo G, Velardi P. The social phenotype: Extracting a patient-centered perspective of diabetes from health-related blogs. Artif Intell Med 2019;101:101727. https://doi.org/10.1016/j.artmed.2019.101727.

[30] Shaw G, Karami A. Computational content analysis of negative tweets for obesity, diet, diabetes, and exercise. Proc Assoc Inf Sci Technol 2017;54:357–65. https://doi.org/10.1002/pra2.2017.14505401039.

[31] Masmoudi A, Barhamgi M, Faci N, Saoud Z, Belhajjame K, Benslimane D, et al. An ontology-based approach for mining radicalization indicators from online messages. Proc. - Int. Conf. Adv. Inf. Netw. Appl. AINA, vol. 2018- May, Institute of Electrical and Electronics Engineers Inc.; 2018, p. 609–16. https://doi.org/10.1109/AINA.2018.00094.

[32] Beguerisse-Díaz M, McLennan AK, Garduño-Hernández G, Barahona M, Ulijaszek SJ. The ‘who’ and ‘what’ of #diabetes on Twitter. Digit Heal 2017;3:205520761668884. https://doi.org/10.1177/2055207616688841.

[33] Bello-Orgaz G, Hernandez-Castro J, Camacho D. Detecting discussion communities on vaccination in twitter. Futur Gener Comput Syst 2017. https://doi.org/10.1016/j.future.2016.06.032.

[34] Sathiyakumari K, Vijaya MS. Identification of subgroups in a directed social network using edge betweenness and random walks. Smart Innov. Syst. Technol., vol. 77, Springer Science and Business Media Deutschland GmbH; 2018, p. 115–25.

https://doi.org/10.1007/978-981-10-5544-7_12.

[35] Zheng X, Han J, Sun A. A Survey of Location Prediction on Twitter. IEEE Trans Knowl Data Eng 2018;30:1652–71. https://doi.org/10.1109/TKDE.2018.2807840.

[36] Rangel F, Rosso P, Montes-Y-Gómez M, Potthast M, Stein B. Overview of the 6th Author Profiling Task at PAN 2018: Multimodal gender identification in Twitter. CEUR Workshop Proc., 2018.

[37] Takahashi T, Tahara T, Nagatani K, Miura Y, Taniguchi T, Ohkuma T. Text and image synergy with feature cross technique for gender identification: Notebook for PAN at CLEF 2018. CEUR Workshop Proc., vol. 2125, JMIR Publications Inc.; 2018. https://doi.org/10.2196/14077.

[38] Pérez-Pérez M, Pérez-Rodríguez G, Fdez-Riverola F, Lourenço A. Using twitter to understand the human bowel disease community: Exploratory analysis of key topics. J Med Internet Res 2019;21. https://doi.org/10.2196/12610.

[39] Ke Q, Ahn Y-Y, Sugimoto CR. A systematic identification and analysis of scientists on Twitter. PLoS One 2017;12:e0175368. https://doi.org/10.1371/journal.pone.0175368.

[40] Bian J, Zhao Y, Salloum RG, Guo Y, Wang M, Prosperi M, et al. Using social media data to understand the impact of promotional information on laypeople’s discussions:a case study of lynch syndrome. J Med Internet Res 2017;19:e414. https://doi.org/10.2196/jmir.9266.

[41] Puerta P, Laguna L, Vidal L, Ares G, Fiszman S, Tárrega A. Co-occurrence networks of Twitter content after manual or automatic processing. A case-study on “gluten-free.” Food Qual Prefer 2020;86:103993. https://doi.org/10.1016/j.foodqual.2020.103993.

[42] EU Commission. Farm to Fork Strategy, for a fair, healthy and environmentally-friendly food system. Eur Comm 2020:COM(2020) 381 final.

[43] EU Commission. The European Green Deal. Eur Comm 2019:COM(2019) 640 final.

[44] Rubio–Tapia A, Kyle RA, Kaplan EL, Johnson DR, Page W, Erdtmann F, et al. Increased Prevalence and Mortality in Undiagnosed Celiac Disease. Gastroenterology 2009;137:88–93. https://doi.org/10.1053/J.GASTRO.2009.03.059.

[45] Ludvigsson JF, Rubio-Tapia A, van Dyke CT, Melton LJ, Zinsmeister AR, Lahr BD, et al. Increasing incidence of celiac disease in a North American population. Am J Gastroenterol 2013;108:818–24. https://doi.org/10.1038/ajg.2013.60.

[46] Limketkai BN, Sepulveda R, Hing T, Shah ND, Choe M, Limsui D, et al. Prevalence and factors associated with gluten sensitivity in inflammatory bowel disease. Scand J Gastroenterol 2018;53:147–51. https://doi.org/10.1080/00365521.2017.1409364.

[47] Manceñido Marcos N, Pajares Villarroya R, Salinas Moreno S, Arribas López MR, Comas Redondo C. The association between de novo inflammatory bowel disease and celiac disease. Rev Esp Enferm Dig 2020;112:7–11. https://doi.org/10.17235/reed.2019.5535/2018.

[48] Volta U, Pinto-Sanchez MI, Boschetti E, Caio G, De Giorgio R, Verdu EF. Dietary triggers in irritable bowel syndrome: Is there a role for gluten? J Neurogastroenterol Motil 2016;22:547–57. https://doi.org/10.5056/jnm16069.

[49] Aziz I, Trott N, Briggs R, North JR, Hadjivassiliou M, Sanders DS. Efficacy of a Gluten-Free Diet in Subjects With Irritable Bowel Syndrome-Diarrhea Unaware of Their HLA-DQ2/8 Genotype. Clin Gastroenterol Hepatol 2016;14:696-703.e1. https://doi.org/10.1016/j.cgh.2015.12.031.

[50] Herfarth HH, Martin CF, Sandler RS, Kappelman MD, Long MD. Prevalence of a gluten-free diet and improvement of clinical symptoms in patients with inflammatory bowel diseases. Inflamm Bowel Dis 2014;20:1194–7.

https://doi.org/10.1097/MIB.0000000000000077.

[51] Haupt-Jorgensen M, Holm L, Josefsen K, Buschard K. Possible Prevention of Diabetes with a Gluten-Free Diet. Nutrients 2018;10:1746. https://doi.org/10.3390/nu10111746.

[52] Levinta A, Mukovozov I, Tsoutsoulas C. Use of a gluten-free diet in schizophrenia: A systematic review. Adv Nutr 2018;9:824–32. https://doi.org/10.1093/ADVANCES/NMY056.

[53] Brietzke E, Cerqueira RO, Mansur RB, McIntyre RS. Gluten related illnesses and severe mental disorders: a comprehensive review. Neurosci Biobehav Rev 2018;84:368–75. https://doi.org/10.1016/j.neubiorev.2017.08.009.

[54] Newberry C, McKnight L, Sarav M, Pickett-Blakely O. Going Gluten Free: the History and Nutritional Implications of Today's Most Popular Diet. Curr Gastroenterol Rep 2017;19. https://doi.org/10.1007/s11894-017-0597-2.

[55] Reilly NR. The Gluten-Free Diet: Recognizing Fact, Fiction, and Fad. J Pediatr 2016;175:206–10. https://doi.org/10.1016/j.jpeds.2016.04.014.

[56] Masih J, Sharma A, Teodor R. Study on Consumer Behaviour and Economic Advancements of Gluten-free Products. Niger Orig Res Artic Masih Sharma 2016;AJEA:24737. https://doi.org/10.9734/AJEA/2016/24737.

[57] Vici G, Belli L, Biondi M, Polzonetti V. Gluten free diet and nutrient deficiencies: A review. Clin Nutr 2016;35:1236–41. https://doi.org/10.1016/j.clnu.2016.05.002.

[58] Silvester JA, Weiten D, Graff LA, Walker JR, Duerksen DR. Is it gluten-free? Relationship between self-reported gluten-free diet adherence and knowledge of gluten content of foods. Nutrition 2016;32:777–83. https://doi.org/10.1016/j.nut.2016.01.021.

[59] Yamamoto Y. Twitter4J - A Java library for the Twitter API 2018.

[60] Raffo J. Worldwide Gender-Name Dictionary 2016.

[61] Lienhart R, Kuranov A, Pisarevsky V. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2003;2781:297–304. https://doi.org/10.1007/978-3-540-45243-0_39.

[62] Lee SK. Sex as an important biological variable in biomedical research. BMB Rep 2018;51:167–73. https://doi.org/10.5483/BMBRep.2018.51.4.034.

[63] Day S, Mason R, Lagosky S, Rochon PA. Integrating and evaluating sex and gender in health research. Heal Res Policy Syst 2016;14:75. https://doi.org/10.1186/s12961-016-0147-7.

[64] Rothe R, Timofte R, Van Gool L. Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks. Int J Comput Vis 2018;126:144–57. https://doi.org/10.1007/s11263-016-0940-3.

[65] Di Fino EMA, Defago MD, Scavuzzo CM. Spatial analysis applied to nutritional epidemiology. 2019 18th Work. Inf. Process. Control. RPIC 2019, Institute of Electrical and Electronics Engineers Inc.; 2019, p. 105–10. https://doi.org/10.1109/RPIC.2019.8882136.

[66] Geonames.org. GeoNames Database. Retrieved 26072016 2016.

[67] Jianqiang Z, Xiaolin G. Comparison research on text pre-processing methods on twitter sentiment analysis. IEEE Access 2017;5:2870–9. https://doi.org/10.1109/ACCESS.2017.2672677.

[68] Satoshi Nakagawa DL and K-LS. Twitter-Text Library 2017. https://github.com/twitter/twitter-text.

[69] Aristotelis. Hunspell Dictionary of English Medical Terms 2016. https://github.com/glutanimate/hunspell-en-med-glut.

[70] Yujian L, Bo L. A Normalized Levenshtein Distance Metric. IEEE Trans Pattern Anal Mach Intell 2007;29:1091–5. https://doi.org/10.1109/TPAMI.2007.1078.

[71] Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit, Association for Computational Linguistics (ACL); 2015, p. 55–60. https://doi.org/10.3115/v1/p14-5010.

[72] Dooley DM, Griffiths EJ, Gosal GS, Buttigieg PL, Hoehndorf R, Lange MC, et al. Food on: A harmonized food ontology to increase global food traceability, quality control and data integration. Npj Sci Food 2018;2:1–10. https://doi.org/10.1038/s41538-018-0032-6.

[73] Kim H, Mentzer J, Taira R. Developing a Physical Activity Ontology to Support the Interoperability of Physical Activity Data. J Med Internet Res 2019;21:e12776. https://doi.org/10.2196/12776.

[74] Golbeck J, Fragoso G, Hartel F, Hendler J, Oberthaler J, Parsia B. The National Cancer Institute's Thesaurus and Ontology. SSRN Electron J 2018. https://doi.org/10.2139/ssrn.3199007.

[75] Schriml LM. Symptom Ontology 2018. http://www.obofoundry.org/ontology/symp.html%0Ahttps://bioportal.bioontology.org/ontologies/SYMP (accessed December 11, 2019).

[76] Rosse C, Mejino JL V. The Foundational Model of Anatomy Ontology. Anat. Ontol. Bioinforma., Springer London; 2008, p. 59–117. https://doi.org/10.1007/978-1-84628-885-2_4.

[77] de Matos P, Alcántara R, Dekker A, Ennis M, Hastings J, Haug K, et al. Chemical entities of biological interest: An update. Nucleic Acids Res 2009;38. https://doi.org/10.1093/nar/gkp886.

[78] Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, et al. Disease Ontology 2015 update: An expanded and updated database of Human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res 2015;43:D1071–8. https://doi.org/10.1093/nar/gku1011.

[79] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Res 2018;46:D1074–82. https://doi.org/10.1093/nar/gkx1037.

[80] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017;45:D353–61. https://doi.org/10.1093/nar/gkw1092.

[81] Couto FM, Lamurias A. MER: a shell script and annotation server for minimal named entity recognition and linking. J Cheminform 2018;10:58. https://doi.org/10.1186/s13321-018-0312-9.

[82] Hutto CJ, Gilbert E. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proc. Eighth Int. AAAI Conf. Weblogs Soc. Media, Michigan: The AAAI Press; 2014.

[83] Pavel A, Palade V, Iqbal R, Hintea D. Using short URLs in tweets to improve twitter opinion mining. Proc. - 16th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2017, vol. 2017- Decem, Institute of Electrical and Electronics Engineers Inc.; 2017, p. 965–70. https://doi.org/10.1109/ICMLA.2017.00-28.

[84] Wan H, Moens M-F, Luyten W, Zhou X, Mei Q, Liu L, et al. Extracting relations from traditional Chinese medicine literature via heterogeneous entity networks. J Am Med Inform Assoc 2016;23:356–65. https://doi.org/10.1093/jamia/ocv092.

[85] Hedley J. jsoup: Java HTML Parser 2017. https://jsoup.org/.

[86] Haugen A. Abstract: The Open Graph Protocol Design Decisions, 2010, p. 338–338. https://doi.org/10.1007/978-3-642-17749-1_25.

[87] Pérez-Rodríguez G, Pérez-Pérez M, Fdez-Riverola F, Lourenço A. Online visibility of software-related web sites: The case of biomedical text mining tools. Inf Process

Manag 2019;56:565–83. https://doi.org/10.1016/j.ipm.2018.11.011.

[88] Tun YM. Comparision of Different Distance Measure Methods in Text Document Clustering. Int J Res Eng 2018;5. https://doi.org/10.21276/ijre.2018.5.7.2.

[89] Bastian M, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks. Third Int AAAI Conf Weblogs Soc Media 2009. https://doi.org/10.1136/qshc.2004.010033.

[90] Brusca I. Overview of biomarkers for diagnosis and monitoring of celiac disease. Adv. Clin. Chem., vol. 68, Academic Press Inc.; 2015, p. 1–55. https://doi.org/10.1016/bs.acc.2014.12.006.

[91] Canadian Celiac Association (CCA). Delays in Celiac Disease Diagnoses Remain Painful and Costly for Canadians n.d. https://www.globenewswire.com/news-release/2019/05/01/1813587/0/en/Delays-in-Celiac-Disease-Diagnoses-Remain-Painful-and-Costly-for-Canadians.html (accessed December 11, 2019).

[92] Hager AS, Taylor JP, Waters DM, Arendt EK. Gluten free beer - A review. Trends Food Sci Technol 2014;36:44–54. https://doi.org/10.1016/j.tifs.2014.01.001.

[93] Jeong B, Yoon J, Lee JM. Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. Int J Inf Manage 2019;48:280–90. https://doi.org/10.1016/j.ijinfomgt.2017.09.009.

[94] Thompson T, Grace T. Gluten in Cosmetics: Is There a Reason for Concern? J Acad Nutr Diet 2012;112:1316. https://doi.org/10.1016/j.jand.2012.07.011.

[95] Dimitrova AK, Ungaro RC, Lebwohl B, Lewis SK, Tennyson CA, Green MW, et al. Prevalence of Migraine in Patients With Celiac Disease and Inflammatory Bowel Disease. Headache J Head Face Pain 2013;53:344–55. https://doi.org/10.1111/j.1526-4610.2012.02260.x.

[96] Pulido O, Zarkadas M, Dubois S, MacIsaac K, Cantin I, La Vieille S, et al. Clinical features anmptovery on a gluten-free diet in Canadian adults with celiac disease. vol. 27. 2013. https://doi.org/10.1155/2013/741740.

[97] Stern MJ, Cotten SR, Drentea P. The Separate Spheres of Online Health: Gender, Parenting, and Online Health Information Searching in the Information Age. J Fam Issues 2012;33:1324–50. https://doi.org/10.1177/0192513X11425459.

[98] Baker P, Dworkin SL, Tong S, Banks I, Shand T, Yamey G. The men’s health gap: Men must be included in the global health equity agenda. Bull World Health Organ 2014;92:618–20. https://doi.org/10.2471/BLT.13.132795.

[99] Institute NC. HINTS 2017: Women and Health Information Seeking. 2018.

[100] Ashley Jane. Women more likely to report ill health than men 2010. http://news.bbc.co.uk/2/hi/health/8588686.stm (accessed February 5, 2020).

[101] García-Molina MD, Giménez MJ, Sánchez-León S, Barro F. Gluten free wheat: Are we there? Nutrients 2019;11:487. https://doi.org/10.3390/nu11030487.

[102] Ozuna CV, Barro F. Safety evaluation of transgenic low-gliadin wheat in Sprague Dawley rats: An alternative to the gluten free diet with no subchronic adverse effects. Food Chem Toxicol 2017;107:176–85. https://doi.org/10.1016/j.fct.2017.06.037.

[103] Zingone F, Swift GL, Card TR, Sanders DS, Ludvigsson JF, Bai JC. Psychological morbidity of celiac disease: A review of the literature. United Eur Gastroenterol J 2015;3:136–45. https://doi.org/10.1177/2050640614560786.

[104] Catassi C. Gluten Sensitivity. Ann Nutr Metab 2015;67:15–26. https://doi.org/10.1159/000440990.

[105] Samsel A, Seneff S. Glyphosate, pathways to modern diseases II: Celiac sprue and gluten intolerance 2013. https://doi.org/10.2478/intox-2013-0026.

[106] Mesnage R, Antoniou MN. Facts and Fallacies in the Debate on Glyphosate Toxicity. Front Public Heal 2017;5. https://doi.org/10.3389/fpubh.2017.00316.

[107] Top 30 Gluten Free Influencers – The Gluten Free Awards n.d. http://theglutenfreeawards.com/top30glutenfreeinfluencers/ (accessed February 4, 2020).

[108] Benetoli A, Chen TF, Aslani P. How patients' use of social media impacts their interactions with healthcare professionals. Patient Educ Couns 2018;101:439–44. https://doi.org/10.1016/j.pec.2017.08.015.

[109] Keller MS, Mosadeghi S, Cohen ER, Kwan J, Spiegel BMR. Reproductive health and medication concerns for patients with inflammatory bowel disease: Thematic and quantitative analysis using social listening. J Med Internet Res 2018;20:e206. https://doi.org/10.2196/jmir.9870.

[110] Muralidhara S, Paul MJ. #Healthy Selfies: Exploration of Health Topics on Instagram. JMIR Public Heal Surveill 2018;4:e10150. https://doi.org/10.2196/10150.

Highlights

- A methodology to process, classify, visualise and analyse the big data knowledge produced by the *sociome* on social media platforms for Health Informatics is proposed.
- The case study is the analysis of the gluten-free food community on Twitter.
- *Sociome* profiling are applied to characterise the different accounts involved in the conversations and analyse the public discussion in terms of role (i.e. individuals and non-individuals), gender (i.e. male or female) and geo-location.
- Ontology engineering, Natural language processing, named entity recognition, machine learning and graph mining techniques are applied.
- Results may help to identify alimentary risk trends, market opportunities, to improve the awareness campaigns of dietary concerns and to identify demographic patterns.

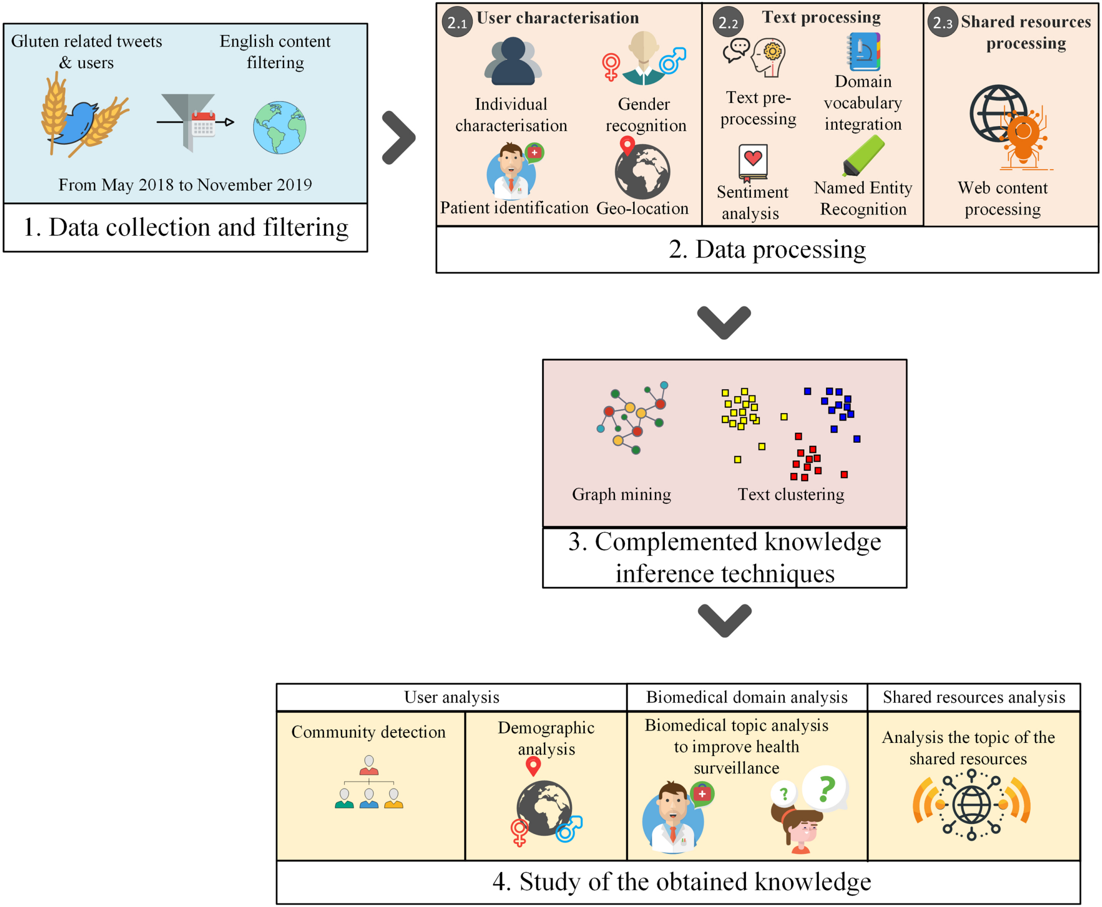

Gluten related tweets & users
English content filtering
From May 2018 to November 2019
1. Data collection and filtering
2.1 User characterisation
Individual characterisation
Gender recognition
Patient identification
Geo-location
2.2 Text processing
Text pre-processing
Domain vocabulary integration
Sentiment analysis
Named Entity Recognition
2.3 Shared resources processing
Web content processing
2. Data processing
Graph mining
Text clustering
3. Complemented knowledge inference techniques
User analysis
Biomedical domain analysis
Shared resources analysis
Community detection
Demographic analysis
Biomedical topic analysis to improve health surveillance
Analysis the topic of the shared resources
4. Study of the obtained knowledge

Figure 1

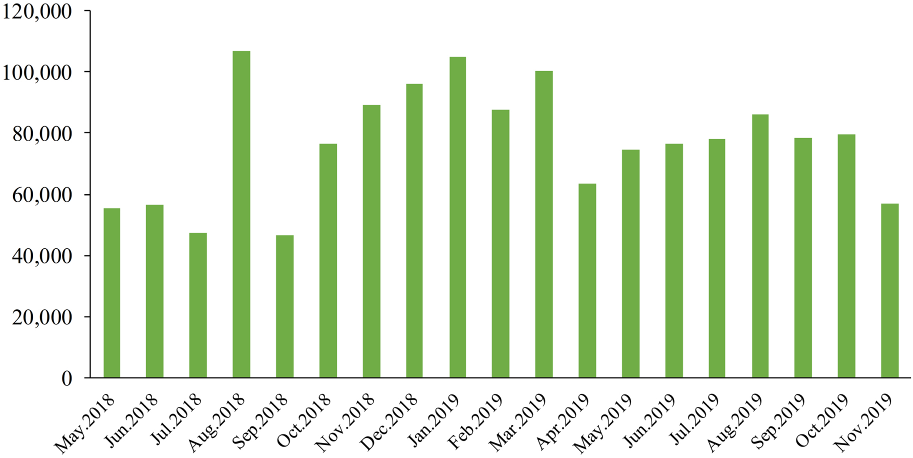

120,000
100,000
80,000
60,000
40,000
20,000
0
May.2018
Jun.2018
Jul.2018
Aug.2018
Sep.2018
Oct.2018
Nov.2018
Dec.2018
Jan.2019
Feb.2019
Mar.2019
Apr.2019
May.2019
Jun.2019
Jul.2019
Aug.2019
Sep.2019
Oct.2019
Nov.2019

Figure 2

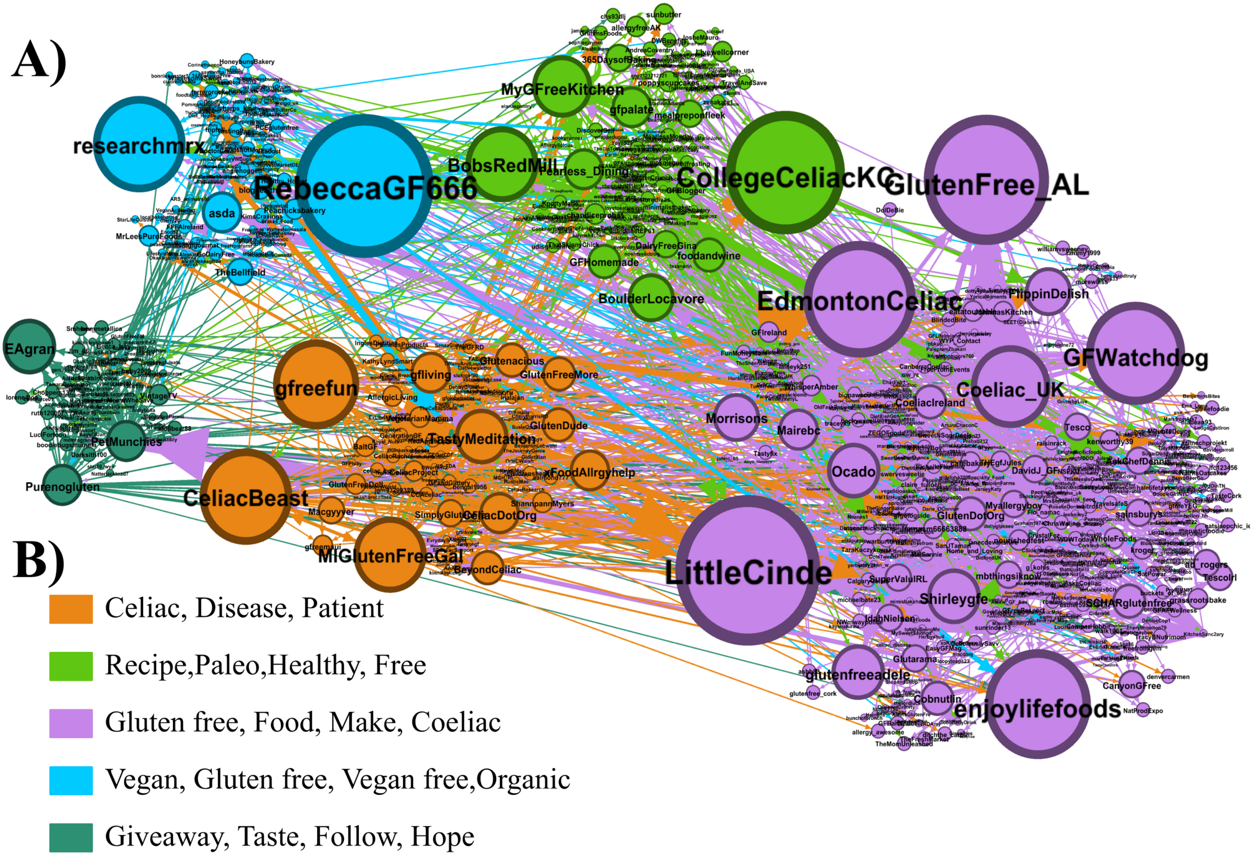

A)
researchmrx
RebeccaGF666
asda
TheBellfield
MrLeesPureFoods
MyGFreeKitchen
gfpalate
BobsRedMill
Fearless_Dining
365DaysofBaking
sunbutter
allergyfreeAK
DairyFreeGina
foodandwine
GFHomemade
BoulderLocavore
CollegeCeliacKC
GlutenFree_AL
EdmontonCeliac
FlippinDelish
GFWatchdog
Coeliac_UK
CoeliacIreland
Morrisons
Mairebc
Ocado
Tesco
LittleCinde
Shirleygfe
glutenfreeadele
enjoylifefoods
CanyonGFree
GlutenDotOrg
Myallergyboy
EAgran
PetMunchies
Purenogluten
gfreefun
gfliving
Glutenacious
GlutenFreeMore
TastyMeditation
GlutenDude
xFoodAllrgyhelp
CeliacBeast
MIGlutenFreeGal
BeyondCeliac
CeliacDotOrg
B)
🟧 Celiac, Disease, Patient
🟩 Recipe,Paleo,Healthy, Free
🟪 Gluten free, Food, Make, Coeliac
🟦 Vegan, Gluten free, Vegan free,Organic
🟩 Giveaway, Taste, Follow, Hope

Figure 3

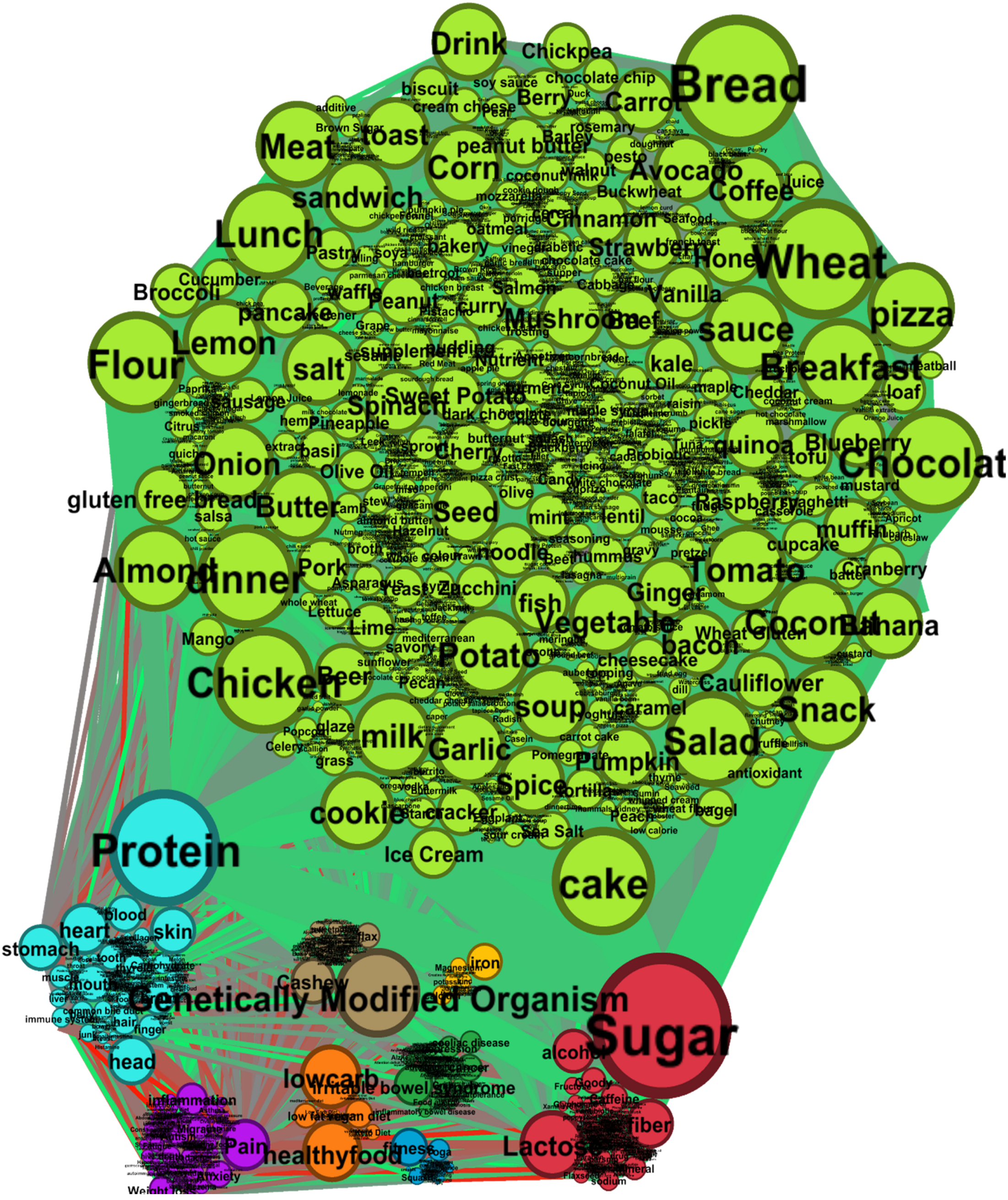

Drink
Chickpea
Bread
biscuit
soy sauce
chocolate chip
Berry
Carrot
cream cheese
additive
Meat
toast
peanut butter
rosemary
Barley
Brown Sugar
Corn
coconut milk
walnut
pesto
Avocado
Coffee
Juice
sandwich
mozzarella
Buckwheat
Lunch
Cinnamon
Seafood
oatmeal
porridge
Pastry
soya
bakery
vinegar
Strawberry
Honey
Wheat
chocolate cake
Broccoli
Cucumber
waffle
Peanut
curry
chicken breast
Salmon
Cabbage
Vanilla
pizza
pancake
Pistachio
Mushroom
Beef
sauce
Flour
Lemon
salt
supplement
pudding
Nutrient
kale
Breakfast
meatball
Red Meat
Coconut Oil
maple
Cheddar
loaf
Paprika
Sausage
Lemon Juice
Sweet Potato
Spinach
dark chocolate
raisin
coconut cream
hot chocolate
marshmallow
Citrus
hemp
Pineapple
pickle
Blueberry
extract
basil
Cherry
quinoa
tofu
Chocolate
Onion
Olive Oil
Candy
white chocolate
mustard
gluten free bread
Butter
lamb
Seed
olive
mint
Lentil
taco
Raspberry
spaghetti
casserole
salsa
muffin
Apricot
hot sauce
cupcake
Nutmeg
Hazelnut
broth
noodle
seasoning
hummus
gravy
pretzel
Almond
dinner
Pork
Tomato
Cranberry
Asparagus
yeast
Zucchini
fish
lasagna
Ginger
batter
whole wheat
Lettuce
Lime
Vegetable
Coconut
Banana
Mango
mediterranean
savory
Wheat Gluten
bacon
custard
sunflower
Potato
Chicken
Beer
Pecan
cheesecake
Cauliflower
cheddar cheese
soup
Yoghurt
caramel
Snack
glaze
Popcorn
milk
Garlic
Salad
Celery
carrot cake
truffle
shellfish
grass
Pomegranate
Pumpkin
thyme
antioxidant
Spice
Tortilla
Seaweed
cookie
Starch
cracker
Eggplant
Peach
bagel
sour cream
Sea Salt
low calorie
Ice Cream
cake
Protein
blood
heart
skin
stomach
collagen
tooth
carbohydrate
thyroid
muscle
mouth
liver
common bile duct
immune system
hair
finger
junk
head
flax
Cashew
iron
Magnesium
potassium
Genetically Modified Organism
Sugar
alcohol
celiac disease
depression
cancer
lowcarb
irritable bowel syndrome
Goody
Fructose
Caffeine
inflammation
low fat
vegan diet
Keto Diet
inflammatory bowel disease
fiber
Migraine
Pain
healthyfood
fitness
yoga
Lactose
autism
anxiety
mineral
Flaxseed
sodium
Weight loss

Figure 4

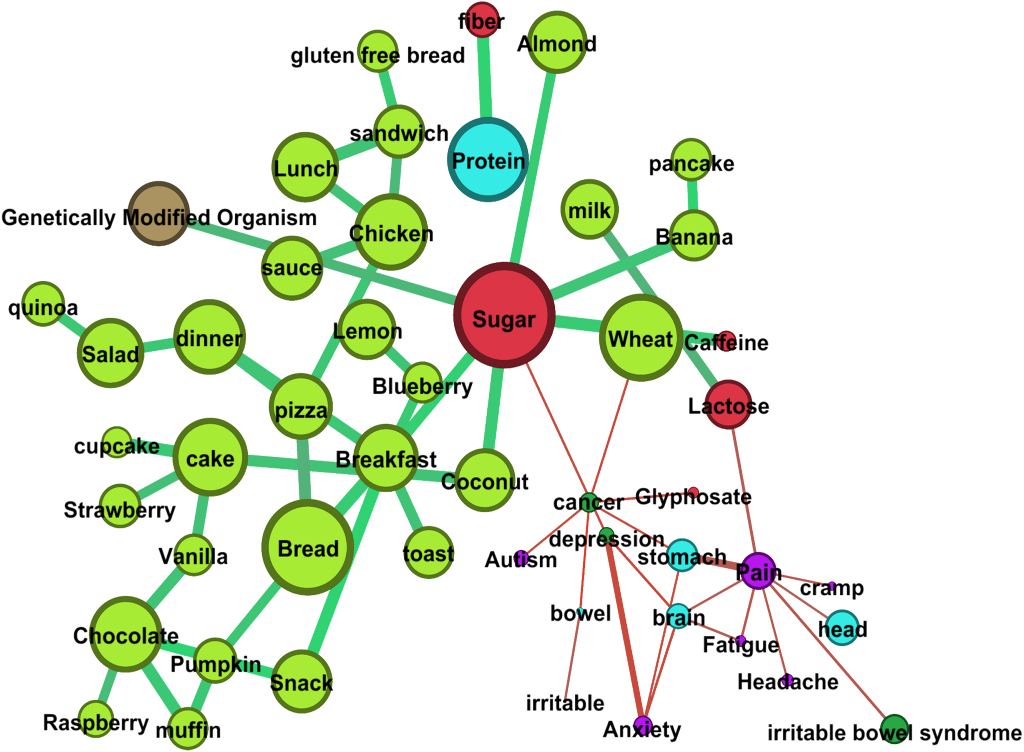
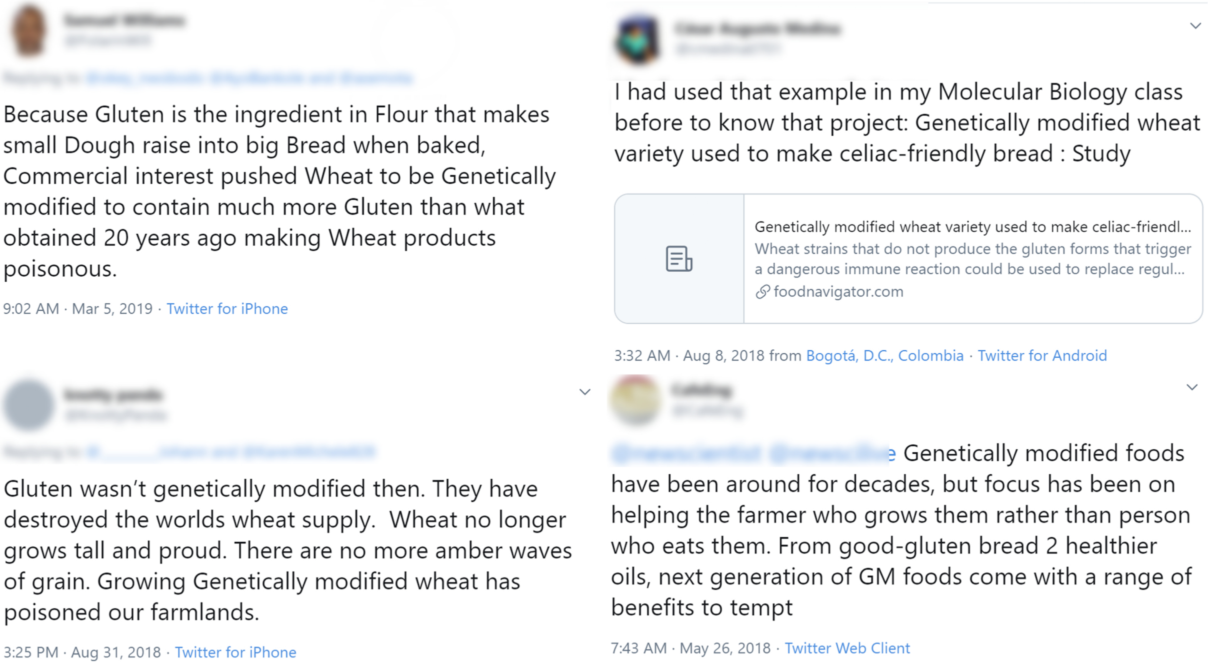

fiber
Almond
gluten free bread
sandwich
Protein
pancake
Lunch
milk
Banana
Genetically Modified Organism
Chicken
sauce
quinoa
Sugar
Wheat
Caffeine
Salad
dinner
Lemon
Blueberry
Lactose
pizza
cupcake
cake
Breakfast
Coconut
cancer
Glyphosate
Strawberry
depression
Autism
stomach
Pain
Vanilla
Bread
toast
cramp
bowel
brain
head
Chocolate
Fatigue
Pumpkin
Snack
Headache
irritable
Raspberry
muffin
Anxiety
irritable bowel syndrome

Figure 5



[redacted]

Because Gluten is the ingredient in Flour that makes small Dough raise into big Bread when baked, Commercial interest pushed Wheat to be Genetically modified to contain much more Gluten than what obtained 20 years ago making Wheat products poisonous.

9:02 AM · Mar 5, 2019 · Twitter for iPhone

[redacted]

I had used that example in my Molecular Biology class before to know that project: Genetically modified wheat variety used to make celiac-friendly bread : Study

Genetically modified wheat variety used to make celiac-friendl...
Wheat strains that do not produce the gluten forms that trigger a dangerous immune reaction could be used to replace regul...
🔗 foodnavigator.com

3:32 AM · Aug 8, 2018 from Bogotá, D.C., Colombia · Twitter for Android

[redacted]

Gluten wasn't genetically modified then. They have destroyed the worlds wheat supply. Wheat no longer grows tall and proud. There are no more amber waves of grain. Growing Genetically modified wheat has poisoned our farmlands.

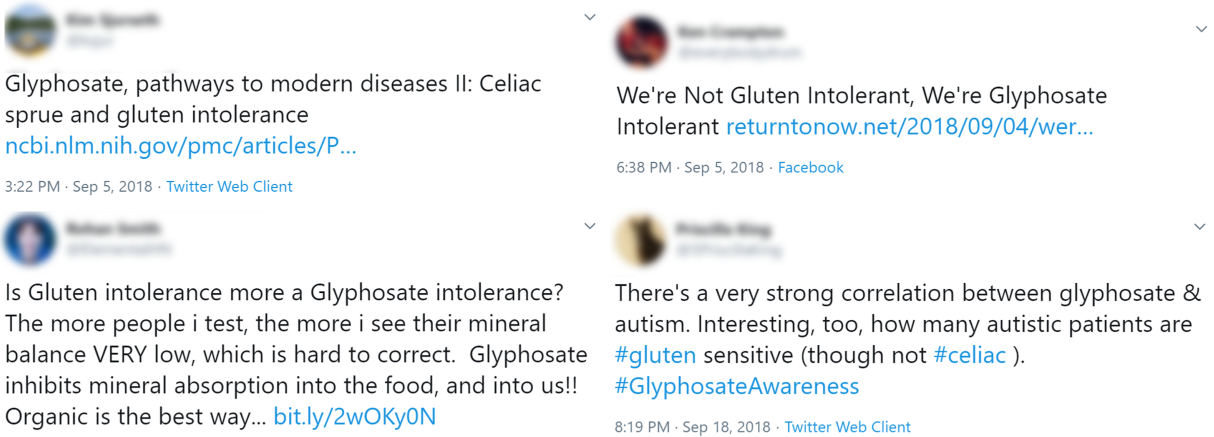
3:25 PM · Aug 31, 2018 · Twitter for iPhone

[redacted]

[redacted] Genetically modified foods have been around for decades, but focus has been on helping the farmer who grows them rather than person who eats them. From good-gluten bread 2 healthier oils, next generation of GM foods come with a range of benefits to tempt

7:43 AM · May 26, 2018 · Twitter Web Client

Figure 6



[redacted]

Glyphosate, pathways to modern diseases II: Celiac sprue and gluten intolerance ncbi.nlm.nih.gov/pmc/articles/P...

3:22 PM · Sep 5, 2018 · Twitter Web Client

[redacted]

We're Not Gluten Intolerant, We're Glyphosate Intolerant returntonow.net/2018/09/04/wer...

6:38 PM · Sep 5, 2018 · Facebook

[redacted]

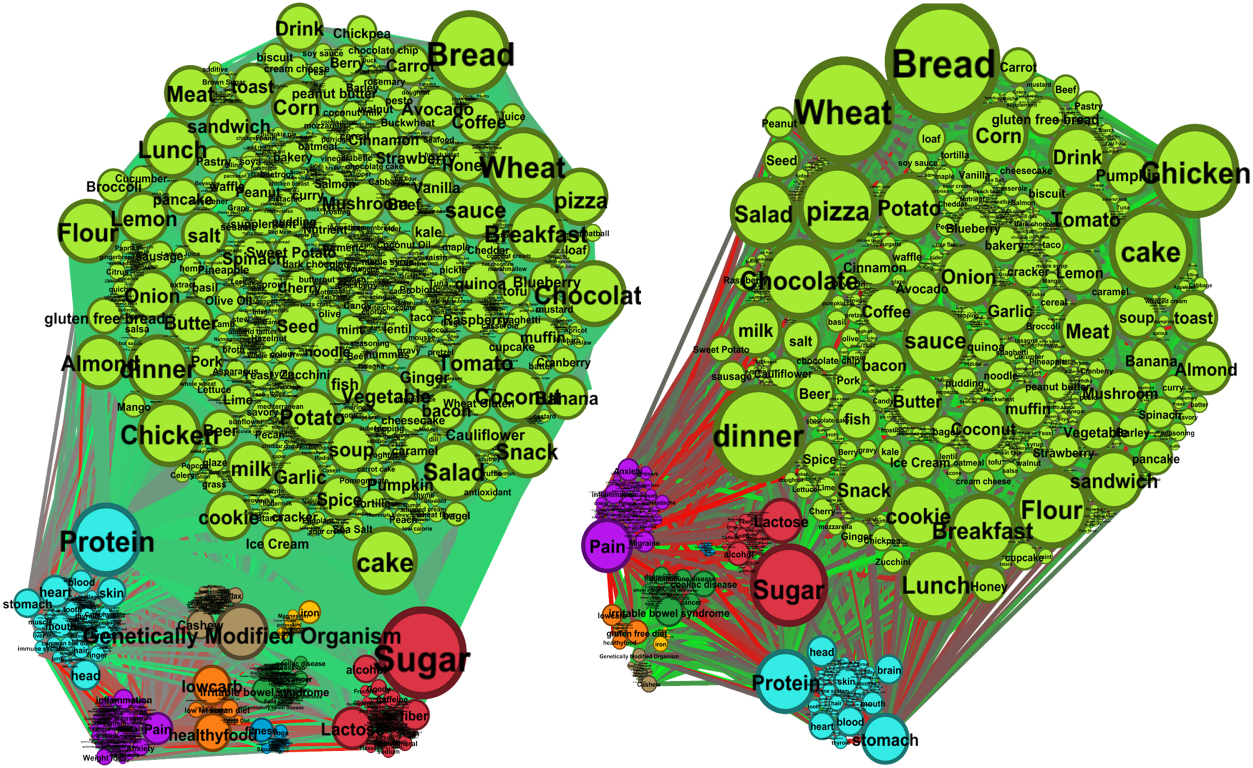
Is Gluten intolerance more a Glyphosate intolerance? The more people i test, the more i see their mineral balance VERY low, which is hard to correct. Glyphosate inhibits mineral absorption into the food, and into us!! Organic is the best way... bit.ly/2wOKy0N

[redacted]

There's a very strong correlation between glyphosate & autism. Interesting, too, how many autistic patients are #gluten sensitive (though not #celiac). #GlyphosateAwareness

8:19 PM · Sep 18, 2018 · Twitter Web Client

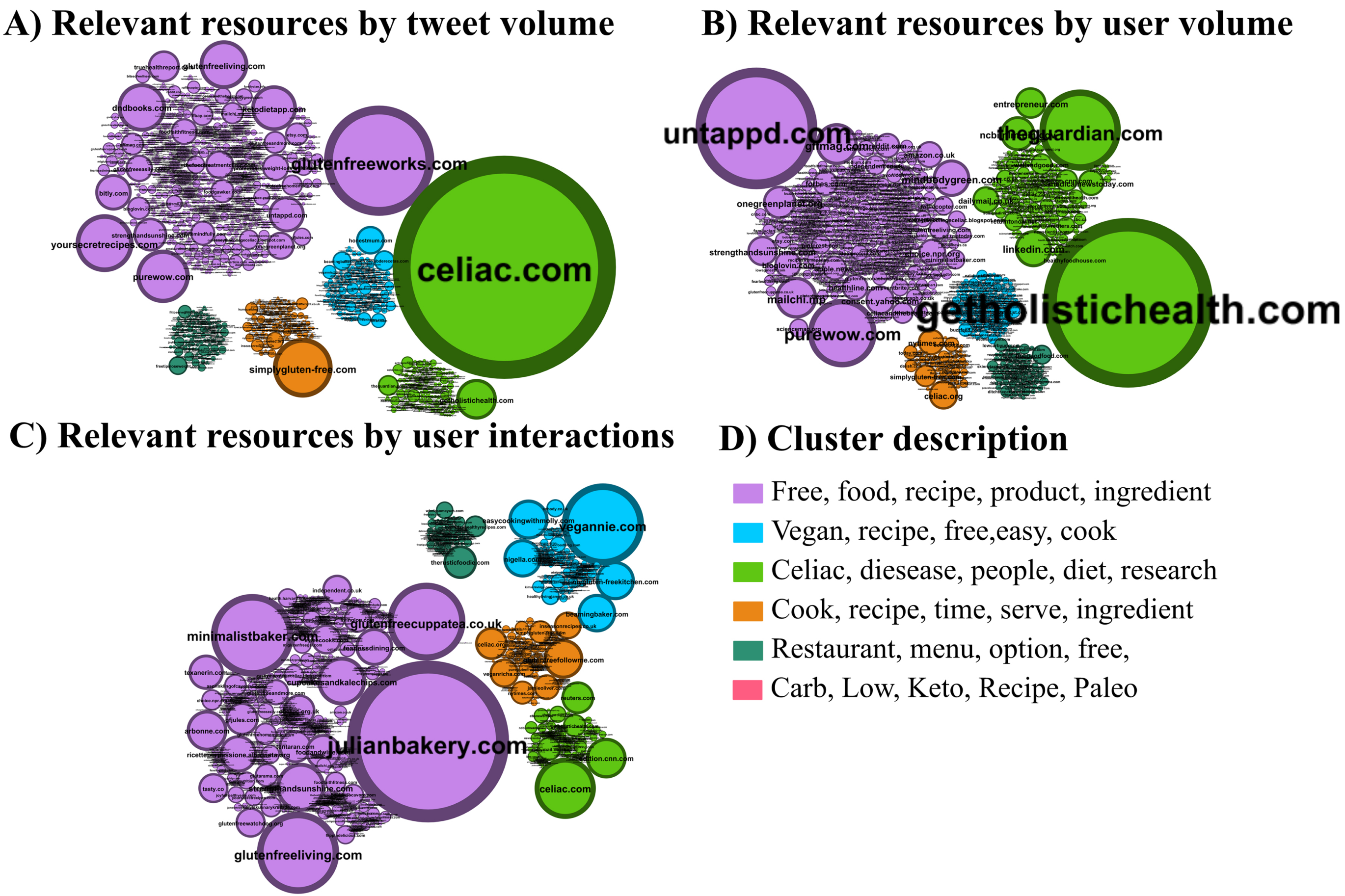
Figure 7



Overall gluten related community
Knowledge graph

Patient detected community
Knowledge graph

Figure 8


A) Relevant resources by tweet volume

B) Relevant resources by user volume

C) Relevant resources by user interactions

D) Cluster description

Figure 9