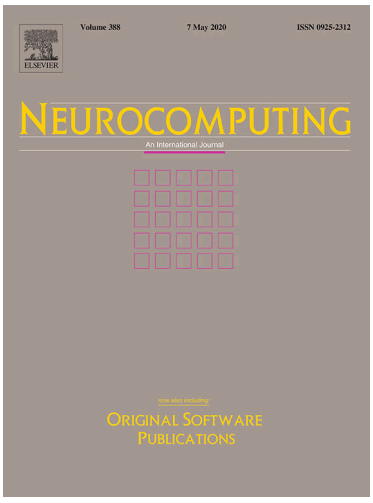
Journal Pre-proofs



Boosting biomedical document classification through the use of domain entity recognizers and semantic ontologies for document representation: the case of gluten bibliome

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

PII: S0925-2312(21)01651-9
DOI: https://doi.org/10.1016/j.neucom.2021.10.100
Reference: NEUCOM 24531

To appear in: *Neurocomputing*

Received Date: 25 January 2021
Revised Date: 5 October 2021
Accepted Date: 8 October 2021

Please cite this article as: M. Pérez-Pérez, T. Ferreira, A. Lourenço, G. Igrejas, F. Fdez-Riverola, Boosting biomedical document classification through the use of domain entity recognizers and semantic ontologies for document representation: the case of gluten bibliome, *Neurocomputing* (2021), doi: https://doi.org/10.1016/j.neucom.2021.10.100

Boosting biomedical document classification through the use of domain entity recognizers and semantic ontologies for document representation: the case of gluten bibliome

Martín Pérez-Pérez[1,2] [0000-0003-1349-6562], Tânia Ferreira[4,5] [0000-0002-3974-0474], Anália Lourenço[1,2,3] [0000-0001-8401-5362], Gilberto Igrejas[4,5,6] [0000-0002-6365-0735] Florentino Fdez-Riverola [1,2] [0000-0002-3943-8013] *

[1] CINBIO, Universidade de Vigo, Department of Computer Science, ESEI - Escuela Superior de Ingeniería Informática, 32004 Ourense, España {martiperez, riverola, analia}@uvigo.es
[2] SING Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, Spain
[3] CEB, Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal
[4] Department of Genetics and Biotechnology, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal gigrejas@utad.pt, tania.rmf@hotmail.com
[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

* Corresponding author: Florentino Fdez-Riverola [Tlf.: +34 988 387015, Fax: +34 988 387001]

Abstract

The increasing number of scientific research documents published keeps growing at an unprecedented rate, making it increasingly difficult to access practical information within a target domain. This situation is motivating a growing interest in applying text mining techniques for the automatic processing of text resources to structure the information that helps researchers to find information of interest and infer knowledge of practical use. However, the automatic processing of research documents requires the previous existence of large, manually annotated text corpora to develop robust and accurate text mining processing methods and machine learning models. In this context, semi-automatic extraction techniques based on structured data and state-of-the-art biomedical tools appear to have significant potential to enhance curator productivity and reduce the costs of document curation. In this line, this work proposes a semi-automatic machine learning workflow and a NER+Ontology boosting technique for the automatic classification of biomedical literature. The practical relevance of the proposed approach has been proven in the curation of 4,115 gluten-related documents extracted from PubMed and contrasted against the word embedding alternative. Comparing the results of the experiments, the proposed NER+Ontology technique is an effective alternative to other state-of-the-art document representation techniques to process the existing biomedical literature.

Keywords: Literature mining, document classification, semi-automatic curation, ontology-based representation, gluten bibliome.

1. Introduction

The recent technological improvements emerged, and the reduction of costs to apply new scientific techniques are generating a vast amount of information associated with the area of Biomedicine [1]. This increase is followed by a corresponding publication of textual knowledge in the form of technical studies, posts and books (also known as bibliome), which keeps growing at an unprecedented rate and exceeding the ability of researchers to digest it [2,3]. At the same time, it is becoming increasingly difficult for the general population to contrast the media misinformation and find reliable sources of information based on the empirical evidence exposed in the bibliome [4,5]. In this sense, current bioinformatics challenges pass through the combination of vast amounts of structured, semi-structured, weakly structured data and unstructured information to build new sources of knowledge that could be explored by the general public and help researchers to discover the knowledge of practical use [6].

In this context, text mining (TM) techniques and machine learning (ML) approaches are being explored as procedures to recognize the relevant parts of the bibliome, allowing the effective search of information, the discovering of hidden interactions between biomedical entities and the assistance in obtaining new knowledge and inferring hypothesis for further biomedical research documents [7,8]. However, the automatic processing of the bibliome requires the previous existence of large, manually annotated text corpora, or structured biomedical databases, to develop robust and accurate workflows that use TM and ML algorithms to process all data automatically. The relevance of the annotated corpora has been highly discussed in diverse manual curation tasks that have been set up to construct gold standards to evaluate and develop derived computational algorithms [9–13]. These efforts have a high impact on delivering new and more robust computational algorithms and biomedical databases, but it also requires high human costs both in time and money [14,15].

The time-consuming nature of manual curation, along with the exponential growth of biomedical literature, strongly limits the number of publications that database curators can revise [16,17]. In addition, the limitations of keyword-based search techniques to rank biomedical articles in a specific problem domain require the application of robust text mining workflows that further consider the role of the biomedical entities discussed in the bibliome. The objective is to create novel computational methods that enable discovering important scientific publications considering the relevance of the biochemical interactions reported. The relevance of this computational support is of utmost importance to discover and analyze the health-related and pharmacological interactions supported by the literature [18]. Consequently, several tasks have promoted the development of novel computational methods to assist in the automatic literature classification and reduce the manual curation efforts [19,20]. In a similar line, different studies propose a hybrid curation process or semi-automatic tasks to promote that experts revise documents that have been automatically processed [21,22]. These semi-automatic approaches combine predictive knowledge extraction methods with manual expert annotation to reduce the required curation work [23].

In this context, this work proposes a biomedical document description (i.e. vector-space) integrated into a semi-automatic curation workflow to enhance the classification efficiency in a real curation task. To this end, the current approach combines unsupervised and biomedical knowledge extraction methods with lexical normalization procedures to boost different state-of-the-art classifiers and assist in the manual curation of the bibliome. In order to evidence the merit of the proposed approach, several experiments were executed comparing the proposed document description technique against the well-known word embedding alternative through the use of state-of-the-art classifiers.

2. Related work

The limitations of keyword-based search techniques to rank relevant biomedical articles in a specific problem domain require investigating robust text mining techniques that further consider the biomedical knowledge contained. Traditional approaches explored unsupervised methods, named entity recognition techniques or domain ontologies to recognize the relevance of a document in a specific domain. For example, García et al. [24], Chen et al. [25], and Matos [26] applied unsupervised semantic similarities as bag-of-concepts to improve the automatic classification performance of biomedical studies. On the other hand, Jorge et al. [27] and Luo et al. [28] considered the integration of named entity recognizers to support the classification of the literature. In terms of semantic normalization to enhance vocabulary unification and to classify documents with similar content, Kulmanov et al. [29] and Ding et al. [30] performed an overview of different approaches that incorporated ontologies-based techniques to ML methods to compute the word similarity. In this line, the authors highlighted the additional inference and reasoning capacity that domain ontologies contribute to the ML area. For its part, Sanchez-Pi et al. [31] demonstrated how an experimental developed ontology-based classification algorithm obtained better performance in the medical area compared with a state-of-the-art ML method. Compared to these works, the main contribution of this study lies in proposing a novel document representation vector-space that takes advantage of combining different techniques (i.e. unsupervised text mining algorithms, named entity recognizers and the lexical capacities of a domain ontology) to boost several state-of-the-art classifiers.

Regarding the implementation of semi-automatic workflows to assist in the manual curation of the literature, different works explored the combination of automatic text mining methods with the manual work of experts to improve the curation accuracy and efficiency. For example, Kwon et al. [32] described the advantages of these workflows in a real curation scenario and discussed how these approaches reduced the annotation time for a beginner-intermediate level annotator. In the same line, Szostak et al. [33] compared a semi-automated workflow against a manual curation counterpart and proved that semi-automatic approaches reached similar results while reducing curation effort. This idea is also supported by Rinaldi et al. [34] that exemplified how text mining technologies could enhance the productivity of the curators. Finally, Winnenburg et al. [35] discussed how text mining methods could be tightly integrated with the manual annotation process to

scale up high-quality manual curation. In the same line, the current work considers the problem of curation and classification of biomedical bibliome as a whole and, in contrast to previous approaches, presents the integration of the proposed document vector-space in a semi-automatic curation workflow to improve the computational and manual performance. Therefore, the proposed semi-automatic workflow guides the manual curators with the extracted knowledge at the same time as it reduces the manual work in an escalated way by applying past curator decisions to filter irrelevant information automatically. On the whole, the proposed approach takes advantage of the combination of an accurate document representation vector-space with a semi-automatic workflow to reach a better performance by the continuous improvement of the applied knowledge inference techniques. In this sense, the implemented workflow was applied in a real curation task to improve the classification performance of gluten-related documents that could contain relevant biochemical interactions.

3. Case study

Concerning the alimentary proteins, more and more studies, as well as health awareness campaigns, keep advertising the existent association between nutrition and the increment of chronic diseases among the population. In this sense, the number of exploratory research studies testing the elimination of some alimentary proteins in specific diets to treat patients with (or without) an apparent nutritional association has highly increased in recent years. However, in return, the implications of suggestions of these exploratory experiments may be misunderstood or misused by bad actors in the social media platforms causing misinformation and high monetary and human costs [36–38]. One of the diets that are being tested as an experimental therapy for the treatment of different diseases, not only for handling gluten-related disorders, is the gluten-free diet (GFD) [39,40]. The difficulty in digesting the growing scientific information, not conclusive scientific evidence in these experimental studies and the influence of social media platforms have caused an increased spread of gluten-related misinformation in the last years [41,42]. This event induces many people to follow the GFD as a self-prescribed lifestyle, although most of them have not been previously diagnosed with a related disease [43,44].

In relation to the gluten protein, Figure 1 shows the increment in the number of scientific documents discussing this topic up to and including the publication tendency for the year 2030.

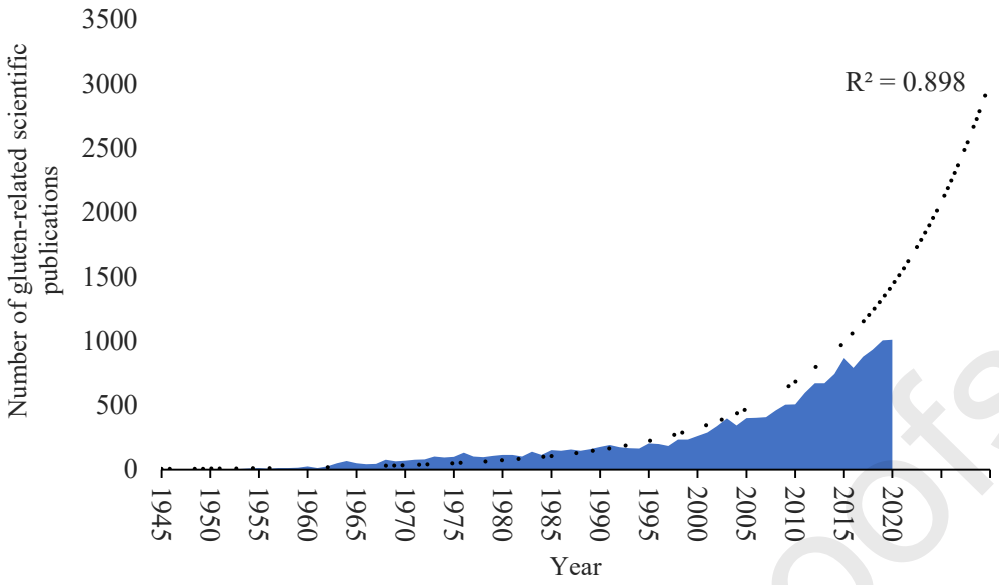


Figure 1: Number of gluten-related scientific works published annually in the PubMed database. The blue color represents the number of gluten-related documents year by year. The black line indicates the exponential publication tendency extrapolated to the year 2030.

Therefore, the need for computational approaches to support the classification and analysis of publications containing relevant biomedical interactions becomes increasingly important, especially to structure the recognized role of the different biochemical compounds in the body processes and diseases. In this line, recent studies explored the manual curation of the literature in different knowledge areas to generate new databases with relevant health-related interactions that provide practical and structured scientific information to the general population and researchers [45–47].

Accordingly, the proposed semi-automatic workflow was applied to the gluten bibliome with the goal of identifying studies that contain relevant health-related knowledge (i.e. documents that support meaningful biochemical interactions) to create a future database that assists researchers to make appropriate decisions and develop new hypotheses supported by the available bibliome.

4. Materials and methods

This section describes the proposed document description technique, as well as the integrated, iterative, and semi-automatic data curation workflow applied to the gluten-related bibliome. The proposed workflow comprises different sub-sequential rounds that support the application of past experiences to assist the document curation incrementally. In other words, manually curated portions of the dataset were incrementally applied to improve and fine tune the different predictive methods that supported the automatic classification and annotation of the remaining unclassified documents. This scalable and iterative approach established a semi-automatic workflow in which unprocessed documents were automatically filtered and annotated, considering the previous curator decisions. Therefore, the manual inferred knowledge of past iterations was automatically

propagated to the sub-sequential curation rounds in order to enhance the baseline classification performance, improve the inferred knowledge methods and reduce human efforts.

In this sense, the implemented workflow consists of the following fundamental phases (*i*) knowledge retrieval; (*ii*) document processing; and (*iii*) document classification. Figure 2 summarizes the different tasks comprising each phase, while the following subsections give details about the strategies applied in every case.

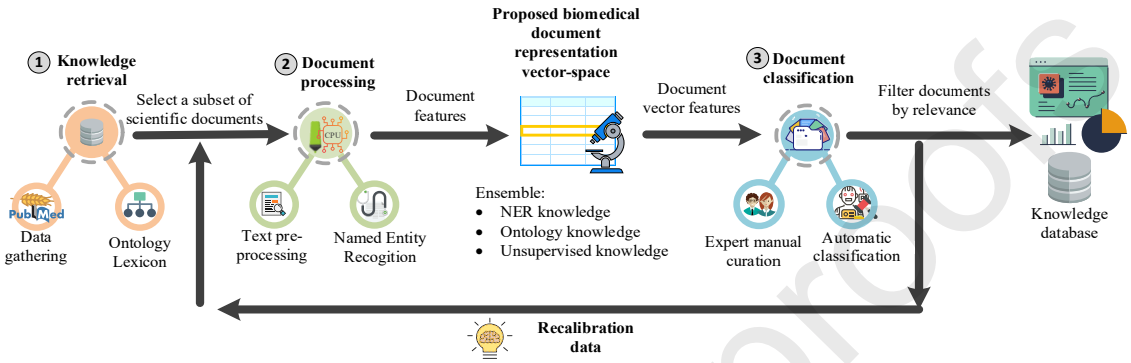


Figure 2: Schema of the semi-automatic curation workflow using the proposed document representation vector-space. The current approach presents a document description technique that combines (i) unsupervised text mining techniques, (ii) named entity recognizers and (iii) domain ontologies to create a document representation (i.e. vector-space) that boosts the automatic curation of the biomedical bibliome. This approach takes advantage of the integrated semi-automatic workflow using past experiences to automatically filter irrelevant documents and to improve the performance of the different processes.

4.1 Knowledge retrieval

As illustrated in Figure 2, the objective of this phase was to retrieve gluten-related scientific documents from the PubMed repository and to identify domain ontologies that were most suited to the scope of this work. The output of this phase provides an initial corpus of gluten-related documents to be further curated, plus a lexicon database to be used to annotate the retrieved documents and normalize the domain identified terms.

4.1.1 Data gathering, lexicon and domain word normalization

The National Center for Biotechnology Information (NCBI) Entrez Utilities Web services were used to access the PubMed library, search for potentially relevant documents, and download associated publication details, including the abstracts [48]. For the current study, the most relevant 4,115 abstracts (out of a total of 12,047 documents) were initially retrieved from the PubMed repository to be further processed.

The following domain-related ontologies and dictionaries were initially selected to recognize, extract and normalize the semantic domain concepts present in the selected documents: FoodOn ontology [49], Symptom (SYMP) ontology [50], Medical Subject Headings (MeSH) [51], Chemical Entities of Biological Interest (ChEBI) lexicon [52], Foundational Model of Anatomy (FMA) ontology [53], National Cancer Institute Thesaurus (NCIt) [54], Disease Ontology [55], DrugBank lexicon [56], KEGG [57];

PharmGKB [58], the protein catalogue of Uniprot [59] and an expert manually curated list of food diets.

Overall, a lexicon of 1,000,450 entries was generated in this step to support the later entity recognition task as well as the normalization of the different terms with the same domain meaning (e.g. normalize the distinct representations of a concept like "*Colonic hamartomatous polyp*" or "*Peutz Jeghers polyp*" to a unique central idea "*Peutz-Jeghers syndrome*").

4.2 Document processing

Once the initial set of documents was retrieved, and the required resources were correctly identified in the previous phase, documents were processed to identify and normalize the different domain-relevant concepts discussed. The output of this phase produces an automatically annotated corpus to be further revised by the experts but also contributes with valuable information suitable for use in the design of the supervised document classifiers.

4.2.1 Initial text pre-processing

Initially, different text pre-processing operations were applied to prepare documents for further exploration. In detail, the following operations were carried out: (*i*) tokenization (i.e. to split a set of text up into minimal meaningful elements); (*ii*) English stop words removal (i.e. elimination of frequent English words like "the" or "by"); (*iii*) n-gram computation (i.e. consider a contiguous sequence of *n* tokens as a concept); (*iv*) part of speech (POS) tagging (i.e. identification of the lexical category of each token); (*v*) small tokens removal (i.e. those having less than three characters); (*vi*) convert tokens to lowercase; and (*vii*) lemmatization (i.e. obtaining the lexeme form of the tokens). This initial document pre-processing was implemented using the well-known Stanford CoreNLP pipeline [60].

4.2.2 Named entity recognition

After the previous procedure, different but complementary NER methods were used to correctly identify mentions of critical entities in the target domain, notably anatomy terms (e.g. duodenal), cell types (e.g. T-cell), compounds (e.g. vitamin D), variety of diets (e.g. vegan), diseases (e.g. osteoporosis), food or food products (e.g. rice), genes (e.g. HLA-DQB1), organisms (e.g. Lactobacillus), proteins (e.g. IgA) and symptoms (e.g. ataxia). These automatically generated annotations were used to index document contents and to help to reduce the cost of their manual annotation. To carry out this operation, an ensemble of the following six state-of-the-art NER taggers was used:

LINNAEUS [61], an open-source stand-alone software system capable of recognizing and normalizing species name mentions with speed and accuracy. The software can be freely downloaded from http://linnaeus.sourceforge.net/.

ABNER [62], a statistical ML system using linear-chain conditional random fields (CRFs) for automatically tagging genes, proteins, and other entity names in a text. The software is freely available at http://pages.cs.wisc.edu/~bsettles/abner/.

OSCAR4 [63], an open-source chemistry analysis routines (OSCAR) developed since 2002 to recognize chemical names, reaction names, ontology terms, enzymes, chemical prefixes, and adjectives. The software can be freely downloaded from https://bitbucket.org/wwmm/oscar4/wiki/Home.

TMCHEM [64], another open-source alternative for identifying chemical names in biomedical literature, including chemical identifiers, drug brand and trade names, and systematic formats. TmChem achieved the highest performance in the BioCreative IV CHEMDNER task (over 87% F-measure), being accessible at https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/tmchem/.

DNORM [65], software that uses ML to recognize and normalize disease names in a biomedical text. DNorm achieved the best performance in the 2013 ShARe/CLEF shared task on disease normalization in clinical notes, being accessible at https://www.ncbi.nlm.nih.gov/research/bionlp/Tools/dnorm/.

In addition, with the goal of complementing the functionality offered by the previous state-of-the-art taggers, an in-house ontology-based NER able to perform dictionary lookups as well as pattern and rule-based recognition was also developed. It is based on an inverted recognition strategy that uses words as patterns to be matched against an ontology-based lexicon [66]. The proposed approach is suitable for the type of texts analyzed in the current 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. "*wheat gluten protein*" instead of "*gluten*"), while concepts that may be associated with more than one semantic category were ignored. Additionally, the implemented recognizer accepts perfect matches as well as lexical variations of the terms (i.e. lemmatized entries, abbreviations, and synonym normalization), being updated with the expert recommendations at the end of each curation round to improve its annotation performance (i.e. semantic type of the annotations and false-positive identified concepts).

4.2.3 Automatic text annotation

In order to integrate the previously commented alternatives with the goal of improving the global accuracy of annotations by taking into consideration their semantic context, the following strategy was applied.

Initially, all documents were annotated separately with each tagger, selecting the annotations containing more grams. In this way, the inconsistences (i.e. different annotations of the same token with incompatible semantic types) were solved by prioritizing the expert-derived knowledge incorporated into the in-house ontology-based NER over the successive annotation rounds. Alternatively, if there was not a match that could be solved by the ontology-based NER (e.g. a new concept that was not previously incorporated into the ontology lexicon), the confidence of the different taggers was used. This ontology-based normalization enables additional inference and reasoning capacity steps like standardizing different terms with similar meaning (i.e. synonyms) or inferring related semantical terms (i.e. deduce families of concepts). After that, a post-processing operation to enhance the annotation performance based on the semantic context of the identified terms and their recognized semantic category was executed. In order to carry

out this process, a rule-based annotation strategy was applied following the criteria of the expert curators (see Supplementary material 1). As an example, if a given tagger identified the anatomic part "*small intestinal*" and another tagger determined a symptom "*intraepithelial lymphocytosis*" in a nearest semantic context, then those annotations were joined into the most complex domain concept "*small intestinal intraepithelial lymphocytosis*". In addition, the capacity to identify semantical patterns in the context of the annotations allowed the recognition of more complex concepts. For example, supposing the name of a food or a protein is identified near to the word "sensitivity" or "intolerance", then the corresponding annotation is expanded to the associated symptom semantic category (e.g. "*egg intolerance"*).

4.3 Document classification

Following the proposed workflow illustrated in Figure 2, the semi-automatic annotation process concludes with the curation of the documents by experts. In this phase, experts revise the integrated NER annotations automatically generated in the previous phase and classify the content of the documents as relevant or irrelevant. The first round of this iterative task generates an initial gold standard (updated in subsequent annotation rounds), that is used to retrain and update the overall performance of the different methods used in the implemented workflow. To facilitate a better understanding of the actual integration of the manual curation step in the proposed workflow, Figure 3 provides details about the iterative curation process and classification strategy applied in the current work.

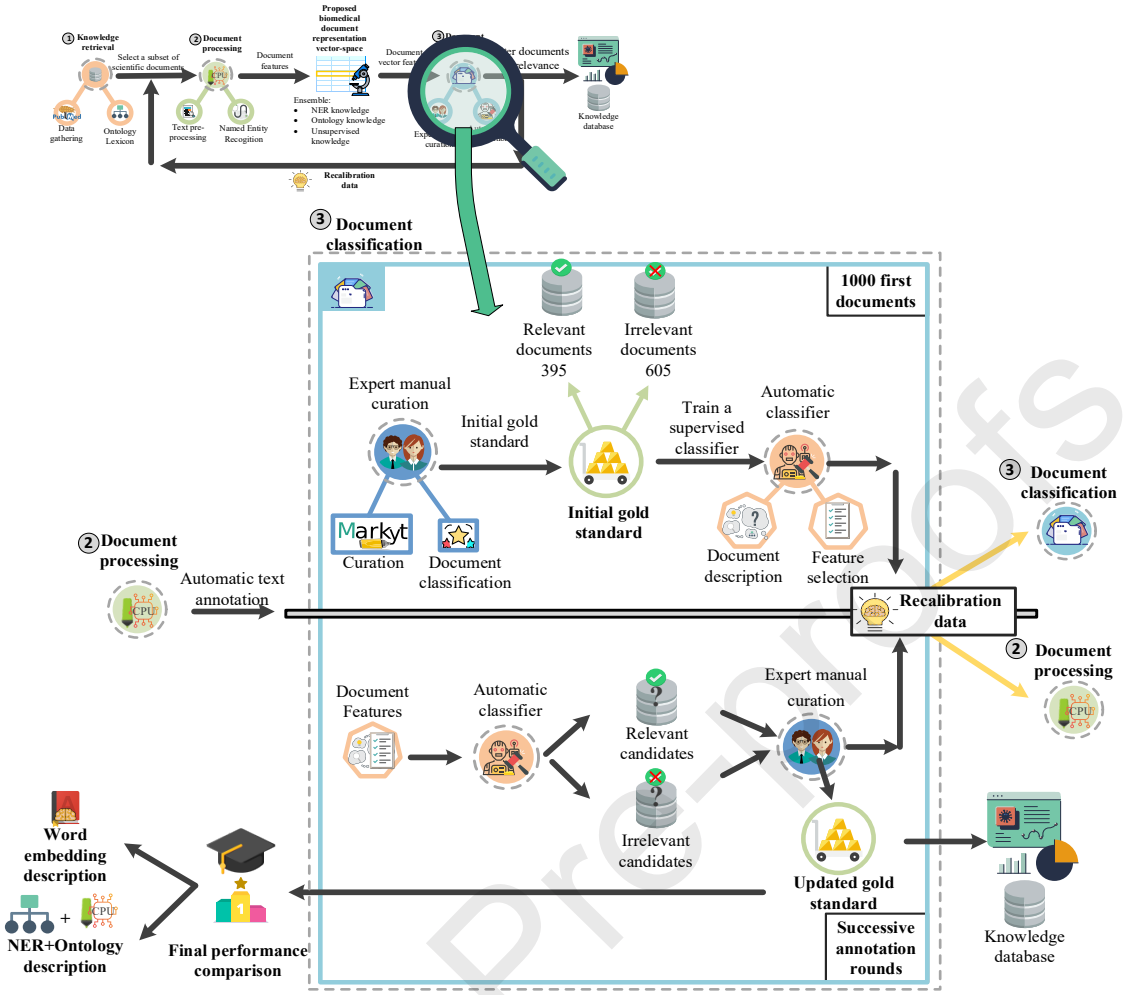


Figure 3: Iterative semi-automatic curation strategy. The curation phase starts with the automatic annotation and manual curation of the 1,000 PubMed records (belonging to the initial group of 4,115 documents). This process generates an updatable gold standard used to improve the different methods that form part of the proposed workflow.

4.3.1 Expert manual curation

In order to provide specific support to experts in the initial manual curation of the automatically generated annotations, and also for the later document classification phase, the Markyt annotation tool s was used. In detail, the Markyt framework contributed with useful information concerning the following aspects of the proposed workflow: (*i*) produced valuable insights to update both the ontology-based NER algorithm and the automatic text annotation of successive rounds; and (*ii*) made available relevant information about the manual classification of documents to improve the subsequent training and test of the automated classifier included in the workflow. Figure 4 shows the Markyt framework in action during the annotation of two given documents.

Figure 4: Snapshot of the Markyt annotation tool used in the proposed workflow. Illustrate the visualization of document contents, existing annotations, relevance classification (icon following to the PubMed id) and different available semantic types.

As previously commented, in order to obtain an initial gold standard for feeding the proposed workflow and generate a primary document classifier, a first round with 1,000 automatically annotated but unclassified documents was carried out (see Figure 3, top). The expert revision of this initial set of documents enabled (*i*) the development of the first classifier to assess the relevance of the follow-up documents belonging to the next rounds and (*ii*) established the basis for the automatic annotation rules. The application of this semi-automatic curation strategy helped to save manual efforts by identifying relevant biomedical entities and relevant documents based on previous experiences [67]. Following this iterative approach, inconsistencies, glitches, misses, and interpretation issues were fixed and documented by experts to enhance the global workflow performance (i.e. improving the vocabulary and matching rules supporting both the automatic annotation and the priority given to each available tagger).

4.3.2 Initial document description

With the goal of accurately representing each document for serving as input to the automatic classifier of the proposed workflow, a document attribute matrix was initially formed considering three different but complementary groups of attributes.

The first set of attributes is given by a word vector containing the most meaningful concepts of each document when considering the whole dataset (first set of descriptive columns in Table 1). For its generation, the term frequency-inverse document frequency (TF-IDF) measure of unigrams, bigrams, and trigrams was computed. TF-IDF measures the significance of a given word in a dataset regarding the total number of times that it appears in a particular document compared to the overall dataset (Equation 1).

$$\text{TF - IDF}(t,d,D) = \text{TF}(t,d) \times \text{IDF}(t,D) \quad (1)$$

where t is the evaluated term, d stands for any given document from the dataset, D, and TF (t,d) expresses the ratio of corresponding to the term, t, in a document, d, described as follows (Equation 2):

$$\mathrm{TF}(t,d) = \frac{n_t}{\sum_k n_k} \tag{2}$$

where n_t is the number of occurrences of the term, t, in a document, d, and n_k is the total number of terms in a document, d. Moreover, in Equation 1 IDF(t,D) stands for the logarithmic ratio of the term, t, in the dataset, D, being computed as follows (Equation 3):

$$\mathrm{IDF}(t,D) = \log \frac{|D|}{|\{d_i \in D | t \in d_i\}|} \tag{3}$$

Table 1: Initial vector representation describing each document.

Document	**TF-IDF terms**			**#Normalised annotations**			**#Domain count**		
	celiac	**allergy**	**…**	**Glu-D1**	**DOID:9892**	**…**	**Diseases**	**Compounds**	**…**
Doc 1	0.6	0.30	…	2	5	…	2	8	…
Doc 2	0.32	0	…	0	2	…	5	3	…
…	…	…	…	…	…	…	…	…	…
Doc *n*	…	…	…	…	…	…	…	…	…

The second group of attributes comprises the normalized label of all automatic annotations of each document (second set of descriptive columns in Table 1). In this way, the combination of the output of the different taggers, in conjunction with the distinct domain ontologies of the lexicon, enables that those annotations with a similar meaning can be computed as the same entry in this attribute group.
Finally, the third group of attributes includes the different semantic types annotated in each document (third set of descriptive columns in Table 1).

4.3.3 Final document representation

Regardless of the specific strategy used to extract attributes from any document to generate its vector representation (as the one proposed in the previous section), thousands of entries usually form it. This scenario requires the consideration of a precise feature selection procedure to identify the most informative features. In this sense, a combination of Information Gain (IG), Chi Square (χ^2), and the stability-correlation measure was applied in this work.
In detail, the IG of any feature, f_k, describing a class, c_i, represents the reduction in uncertainty about c_i when the value of f_k is known, and can be calculated as follows (Equation 4):

$$\mathrm{IG}(f_k,C_i) = \sum_{C \in \{c_i,\bar{c}_i\}} \sum_{t \in \{f_k,\bar{f}_k\}} P(f_i,c_i)\log\frac{P(f_k,c_k)}{P(f_k)P(c_i)} \tag{4}$$

where $P(c)$ is the fraction of the documents belonging to class, c, over the total number of documents, $P(f,c)$ is the fraction of the documents belonging to class, c, that contains a feature, f, over the total number of documents, and $P(f)$ represents the fraction of the documents that contain a feature, f, over the total number of documents.

For its part, the χ^2 measure is commonly used in mathematical statistics to evaluate the independence of any two given variables. In the proposed approach, the independence of a feature, f_k, with respect to a category, c_i, is measured by Equation 5, in which the greater the value of the $\chi^2(f_k,c_i)$ is, the more information provides the feature, f_k:

$$\chi^2(f_k,c_i) = \frac{D(a_{ki}d_{ki} - b_{ki}c_{ki})^2}{(a_{ki}+b_{ki})(a_{ki}+c_{ki})(b_{ki}+d_{ki})(c_{ki}+d_{ki})} \tag{5}$$

where D stands for the total number of documents, ak_i is the frequency of feature, f_k, in the category, c_i, b_{ki} is the frequency of feature, f_k, in all the existing categories except, c_i, c_{ki} is the frequency with which category, c_i, occurs without containing feature, f_k, and d_{ki} is the number of times neither c_i nor f_k occur.

Finally, the stability-correlation statistic evaluates the importance of any given variable based on its stability and correlation concerning a given class (i.e. variables with a high correlation and high stability achieve an importance nearest to 1). In the current study, the stability measure of a feature, f_k, over a class, c_i, corresponds to the percentage of documents that have similar values for the same feature, being defined as follows (Equation 6):

$$\mathrm{Stability}(f_k) = \frac{D_{(f_k,c_i)}}{D} \tag{6}$$

where D is the total number of documents, and $D_{(f_k,c_i)}$ stands for the frequency of feature, f_k, in class, c_i, for D.

In a complementary way, the correlation measure of a feature, f_k, over a class, c_i, is defined as follows (Equation 7):

$$\mathrm{Correlation}(f_k) = \frac{\sum_{j=1}^{D}(f_{kj}-\overline{f_k})(c_{ij}-\overline{c_i})}{\sqrt{[\sum_{j=1}^{D}(f_{kj}-\overline{f_k})^2][\sum_{j=1}^{n}(c_{ij}-\overline{c_i})^2]}} \tag{7}$$

where D is the total number of documents, $\overline{f_k}$ is the mean of feature, f_k, and $\overline{c_i}$ is the mean of f_k for the class c_i.

In the proposed approach, a combination of the three feature selection techniques was devised to select the top 300 features achieving the most significant average weight, normalized between 0 and 1.

5. Results and discussion

This section introduces the final gold standard dataset created by applying the suggested semi-automatic workflow to the gluten bibliome case study, giving relevant details about the corpus in terms of both relevance in classification and representativeness for the selected domain. After that, and with the goal of assessing the adequacy of the proposed document representation method (discussed in Sections 3.3.2 and 3.3.3), a well-known baseline (i.e. word embedding) is briefly described together with the introduction of the experimental setup and the definition of the selected performance measures. The results from six state-of-the-art classifiers are presented and analyzed in detail, evidencing the importance of having an accurate document representation for obtaining positive outcomes in the classification task. The section ends with a learned lessons discussion that summarizes the key findings resulting from applying the current workflow and the interaction with experts in the studied field.

5.1 Gluten-related gold standard

As previously commented, in order to create a comprehensive curated corpus following the proposed semi-automatic workflow, a total of 4,115 PubMed documents related to gluten bibliome (out of a total of 12,047 entries) were iteratively annotated and manually classified by experts with the help of the Marky platform. Table 2 describes the distribution of the final gold standard dataset regarding the relevance of documents.

Table 2: Gold standard dataset manually curated by experts.

	Curated documents
Relevant	1,871 (45.5%)
Irrelevant	2,244 (54.5%)
∑	4,115

The curation process (i.e. semi-automatic annotation and classification) of the documents reflected in Table 2 was carried out by experts through nine rounds with a non-regular number of documents to revise in each iteration. This round flexibility helped to adapt the revision process to the agenda of the curators and enabled the successive improvement of the different workflow algorithms.

In order to provide meaningful insights describing the existing knowledge in the newly generated gold standard dataset, a correlation analysis of the semantic categories was carried out using an association coefficient calculated as follows (Equation 8):

$$\phi_{D_i D_j} = \frac{D_{c_i \cap c_j} D_{c_{i'} \cap c_{j'}} - D_{c_{i'} \cap c_j} D_{c_i \cap c_{j'}}}{\sqrt{D_{c_i} D_{c_{i'}} D_{c_j} D_{c_{j'}}}} \quad (8)$$

where D_{c_i} represents the number of documents containing category c_i, $D_{c_{i'}}$ stands for the number of documents not containing category c_i, $D_{c_i \cap c_j}$ indicates the number of documents containing both categories, c_i and c_j, $D_{c_{i'} \cap c_{j'}}$ indicates the number of documents not containing both categories c_i and c_j, and $D_{c_i \cap c_{j'}}$ represents the number of documents containing category c_i, but not c_j. In this way, the association coefficient ranges between -1 to +1, representing to what extent documents tend to discuss one category but not the other, neither of them or both categories together.

This measure enabled the identification of statistical differences regarding the annotation of the different semantic types between the relevant and irrelevant documents that comprise the corpus. In this sense, the tables below summarize the number of annotations per semantic category and their co-occurrence for relevant (Table 3) and irrelevant (Table 4) documents.

Table 3: Correlation between different annotation types and the number of annotations per semantic category in relevant documents. Explicit mentions of general terms in the analyzed domain (e.g. "celiac", "gluten" or "protein") were not considered for the generation of the table. ANA stands for anatomy, CEL stands for cell types, COMP stands for compounds, DIS stands for diseases, FOOD stands for food or food products, ORG stands for organisms, PROT represents proteins, and SYMP stands for symptoms. In terms of summarization ∑Corr represents the summarization of all correlations and ∑Ann represents the summarization of all annotations

	ANA	CEL	COMP	DIET	DIS	FOOD	GENE	ORG	PROT	SYMP	∑Corr
ANA		0.739	0.749	0.076	0.930	0.550	0.513	0.399	1.001	0.801	**5.760**
CEL			0.244	-0.061	0.285	0.190	0.391	0.082	0.753	0.223	**2.846**
COMP				0.078	0.579	0.446	0.174	0.195	0.714	0.421	**1.640**
DIET					0.182	0.032	-0.060	-0.025	0.010	0.079	**0.286**
DIS						0.438	0.386	0.167	0.854	0.873	**2.057**
FOOD							0.168	0.332	0.615	0.177	**0.509**
GENE								0.083	0.674	0.219	**0.970**
ORG									0.366	0.086	**0.086**
PROT										0.631	**3.865**
SYMP											**0.086**
∑Ann	**3432**	**1074**	**2040**	**99**	**3693**	**1987**	**636**	**315**	**5975**	**2380**	

Table 4: Correlation between different annotation types and the number of annotations per semantic category in irrelevant documents. Explicit mentions of general terms in the analyzed domain (e.g. "celiac", "gluten" or "protein") were not considered for the generation of the table. ANA stands for anatomy, CEL stands for cell types, COMP stands for compounds, DIS stands for diseases, FOOD stands for food or food products, ORG stands for organisms, PROT represents proteins, and SYMP stands for symptoms. In terms of summarization ∑Corr represents the summarization of all correlations and ∑Ann represents the summarization of all annotations

	ANA	CEL	COMP	DIET	DIS	FOOD	GENE	ORG	PROT	SYMP	∑Corr
ANA		0.337	0.752	0.131	0.422	0.573	0.368	0.328	0.630	0.415	**3.957**
CEL			0.250	-0.001	0.179	0.167	0.130	0.035	0.399	0.189	**1.685**
COMP				0.184	0.362	1.079	0.535	0.707	0.876	0.392	**2.540**
DIET					0.128	0.130	-0.022	-0.016	0.033	0.079	**0.484**
DIS						0.300	0.070	-0.039	0.507	0.420	**1.043**
FOOD							0.588	0.764	0.806	0.243	**1.007**
GENE								0.328	0.455	0.062	**1.560**
ORG									0.350	0.070	**0.070**
PROT										0.406	**3.434**
SYMP											**0.070**

$\sum_{Ann}$	1961	335	5450	86	1149	5472	960	816	4200	686

From the annotated categories shown in Table 3 and Table 4, it can be observed that those relevant documents related to the gluten protein were mainly focused on the study of proteins, compounds, and foods that produce a body change in terms of diseases and symptoms. In contrast, irrelevant documents were essentially focused on the analysis of foods and organisms relating to compounds, showing less correlation with diseases and symptoms. In addition, with the goal of analyzing the incidence of some representative terms, Figure 5 presents the top most mentioned terms along with their associated semantic category.

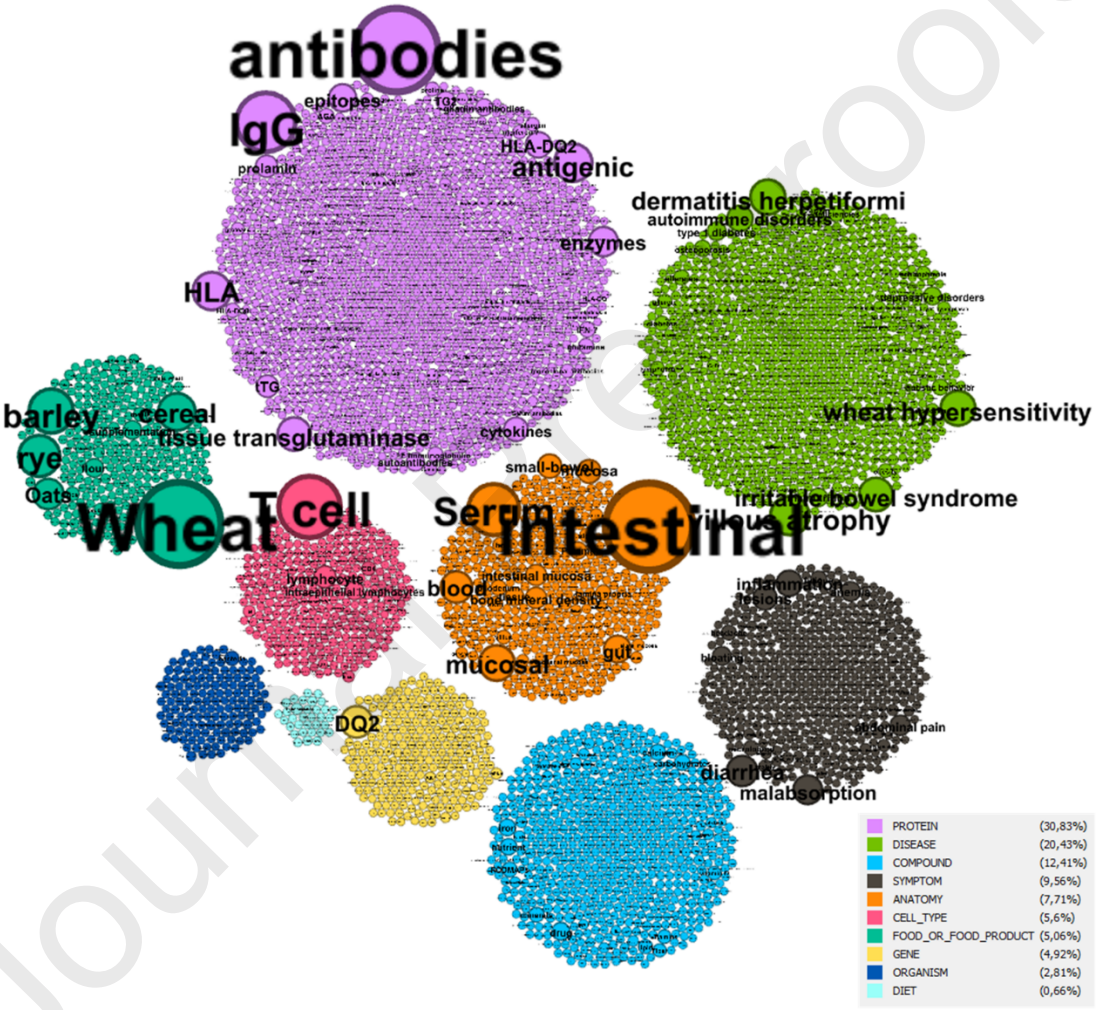


Figure 5. Topmost mentioned terms per semantic category. Terms are arranged taking into consideration the number of different documents that mention each term. Explicit mentions of general terms in the analyzed domain (e.g. "celiac", "gluten" or "protein") were not considered for the generation of the figure.

5.2 Experimental setup

In order to evaluate the proposed document representation method (explained in Sections 3.3.2 and 3.3.3) as part of the developed workflow (Figure 2 and Figure 3), we compared its performance against the use of word embedding [68], one of the most popular

representation technique for capturing the context of words in a given set of documents. To this end, we make use of several state-of-the-art classifiers, including Support Vector Machines (SVM) [69], Random Forest (RF) [70], Generalized Linear Model (GLM) [71], K-nearest neighbor (KNN) [72], Fast Large Margin (FLM) [73], and a Deep Learning (DL) multi-layer feed-forward artificial neural network (ANN) with a 2-2-2 layer configuration [73].

As previously commented, word embedding is a well-reputed unsupervised technique able to generate clusters of words based on their context in a given set of documents [74,75]. This method is generally used in computational linguistics to improve the performance of different ML algorithms [76–78] by taking advantage of the normalization of words with a similar meaning [79,80]. Prominent among other alternatives, Mikolov et al. developed word2vec [81] based on the hypothesis that words that occur in similar contexts tend to have similar meanings [82]. Therefore, word2vec uses a simple neural network to embed words into a continuous vector-space. In the particular case of this study, although there are available different word2vec models for the biomedical area [83], it was trained an in-house word2vec model using the overall gluten-related dataset (i.e. 12,047 documents initially obtained from PubMed, as commented in Section 3.1.1) with the goal of fairly comparing its results against the proposed approach, called NER+Ontology.

In order to obtain accurate results and a well-founded discussion, the six initially selected classifiers were evaluated following a 10-fold cross-validation strategy [84]. Standard measures of precision, recall, and F-score were calculated to assess the performance of the different classifiers, being computed as follows (Equations 9 to 11):

$$\text{Precision} = \frac{TP}{(TP + FP)} \tag{9}$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \tag{10}$$

$$F_{score} = 2 * \frac{TP}{(2 * TP + FP + FN)} \tag{11}$$

where TP is the number of true positives (i.e. relevant documents classified as relevant), FP is the number of false positives (i.e. non-relevant documents classified as relevant), FN is the number of false negatives (i.e. relevant documents classified as irrelevant), and TN stands for the number of true negatives (i.e. non-relevant documents classified as irrelevant).

5.3 Assessing the importance of document representation: word embedding vs. NER+Ontology

This section analyzes the performance obtained by both alternatives as document representation techniques when used to train different state-of-the-art classifiers to be further used in the proposed semi-automatic workflow.

In detail, the first analysis involved the establishment of a baseline to discover which alternative obtains good performance results at the beginning of the process. To carry out this experiment, a 10-fold cross-validation analysis was executed using the first set of 1,000 manually curated documents, which were the output of the first iteration round of the proposed workflow (see Figure 3, top). Table 5 summarizes the results obtained in terms of precision, recall and F-score.

Table 5: Performance comparison of the two document representation techniques under a 10-fold cross-validation scenario using the first set of 1,000 curated documents.

	Word embedding			**NER+Ontology**			**Gain**
	Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F-score*	*F-score*
SVM	0.717	0.653	0.683	0.766	0.861	0.810	+0.127
RF	0.702	0.694	**0.697**	0.826	0.815	**0.820**	+0.123
GLM	0.590	0.615	0.602	0.768	0.784	0.775	**+0.173**
KNN	0.620	0.559	0.587	0.657	0.856	0.743	**+0.156**
FLM	0.686	0.658	0.670	0.785	0.840	0.811	+0.141
DL(2-2-2)	0.672	0.737	0.686	0.797	0.780	0.787	+0.101
Average	0.663	0.650	**0.653**	0.765	0.822	**0.791**	**+0.135**

Regarding the baseline performance comparison shown in Table 5, RF has proven to be the best approach to establish a first recommended classifier as starting point for the semi-automatic curation workflow (F-score = 0.697 and F-score = 0.820). From another perspective, comparing both representation techniques, the proposed NER+Ontology algorithm obtained an average F-score of 0.791, whereas the word embedding alternative achieved an average F-score of 0.653. Considering the differences between the F-score values reported in Table 5, the GLM and KNN classifiers reached the most significant advantage using the proposed document representation technique (ΔF-score = +0.173 and ΔF-score = +0.156, respectively). As an initial conclusion from this first experiment, it seems that the NER+Ontology approach has succeeded in improving the performance of all the analyzed classifiers, regardless of their specific type.

In order to obtain conclusive results, Table 6 presents the final performance achieved in a subsequent experiment using the final gold standard dataset generated by the proposed workflow (see Figure 2). To properly compare the evolution and stability of the different classifiers, they were trained and tested using the same parameters as those adopted in the baseline evaluation.

Table 6: Performance comparison of the two document representation techniques under a 10-fold cross-validation scenario using the final gold standard dataset (4,115 documents).

	Word embedding			NER+Ontology			Gain
	Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F-score*	*F-score*
SVM	0.824	0.838	0.831	0.825	0.890	0.856	+0.025
RF	0.781	0.903	**0.838**	0.915	0.794	0.850	+0.012
GLM	0.803	0.821	0.812	0.869	0.846	0.857	**+0.045**
KNN	0.793	0.833	0.812	0.828	0.838	0.833	+0.021
FLM	0.807	0.836	0.821	0.880	0.841	**0.860**	**+0.039**
DL(2-2-2)	0.780	0.886	0.828	0.878	0.821	0.848	+0.020
Average	0.798	0.852	**0.824**	0.865	0.838	**0.851**	**+0.027**

Regarding the results summarized in Table 6, RF has proven to be the best classifier when using the word embedding technique for document representation (F-score = 0.838), whereas FLM obtained the best position through the use of the proposed NER+Ontology technique (F-score=0.860). As in the previous experiment, comparing the average F-score of both alternatives, the proposed NER+Ontology algorithm reached a better value (average F-score = 0.851) compared to the word embedding approach (average F-score = 0.825). This experiment makes it possible to conclude that the improvement demonstrated by the proposed NER+Ontology technique was stable, regardless of the size of the corpus or the specific classifier used.

In addition, to complement the study carried out, a grid search optimization was executed to evaluate the best performance that the selected classifiers could reach with the two document representation techniques plus a third combination of both. In this regard, Table 7 summarizes the performance measures obtained under a 10-fold cross-validation scenario over the final gold standard dataset.

Table 7. Performance comparison of the different document representation techniques using the best optimization parameters for the different classifiers being evaluated under a 10-fold cross-validation scenario over the final gold standard dataset (4,115 documents).

	Word embedding			NER+Ontology			Gain	Word embedding & NER+Ontology		
	Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F-score*	*F-score*	*Precision*	*Recall*	*F-score*
SVM	0.808	0.877	0.841	0.898	0.830	0.863	+0.022	0.891	0.842	**0.866**
RF	0.796	0.909	**0.849**	0.914	0.815	0.862	+0.013	0.915	0.815	0.862
GLM	0.819	0.844	0.831	0.880	0.850	**0.864**	**+0.033**	0.874	0.851	0.862
KNN	0.770	0.910	0.834	0.860	0.836	0.848	+0.014	0.915	0.789	0.847
FLM	0.795	0.873	0.832	0.878	0.846	0.861	+0.029	0.881	0.846	0.863
DL(2-2-2)	0.788	0.881	0.832	0.884	0.843	0.862	**+0.031**	0.882	0.830	0.854
Average	0.796	0.882	**0.836**	0.886	0.837	**0.860**	**+0.024**	0.893	0.828	**0.859**

From the results shown in Table 7 related to the performance of the two initial alternatives, it can be seen that the proposed NER+Ontology algorithm obtained a better average F-score value (0.860) than the one achieved by the word embedding counterpart (0.836). Considering the overall F-score values attained by the different classifiers, the GLM and RF algorithms reached the best classification performance using the proposed document representation technique (F-score = 0.864 and F-score = 0.849, respectively). Furthermore, the NER+Ontology approach always exceeded the performance obtained

by the word embedding alternative, as showed by the positive values present in the Gain F-score column, being the GLM and DL classifiers that benefit most.

From another interesting perspective, Table 7 also evidenced how the combination of the two document representation techniques (i.e. Word embedding & NER+Ontology) barely achieved a noticeable improvement in some specific cases. This behavior is because the unsupervised word embedding technique does not enhance the semantic normalization obtained using a domain ontology or specific domain NERs. In this sense, a more significant number of domain concepts were supported by the NER+Ontology domain normalization, and only a marginal set of remaining non-stop words was also considered by the standardization provided by the word embedding technique.

5.4 Learned lessons

With the goal of complementing the study carried out with useful insights, this section discusses certain expertise and some lessons learned from implementing the proposed workflow in terms of different design strategies and the manual curation of biomedical information.

In the first place, although the proposed semi-automatic workflow required more computational time and human effort to process (i.e. manually annotate and classify) all the documents comprising the final gold standard in several iterative rounds, human-in-the-loop (HITL) approaches provide better trade-offs guaranteeing an improvement of the accuracy in the majority of datasets while improving safety and precision. In this sense, even though all the automatically classified documents (i.e. relevant and irrelevant) were manually revised in order to correctly evaluate the proposed NER+Ontology technique, the following iterations of the implemented workflow will obtain more benefits since only the manual annotation of relevant documents, and a part of those automatically classified as irrelevant will be required. This strategy allows saving of manual classification efforts because it reduces the number of documents to be revised. In this way, in successive iterations of the proposed workflow, only documents automatically labeled as relevant and a random subset of documents classified as irrelevant (e.g. 20%), are going to be curated in order to recalibrate the internal classifier. In terms of global performance, the proposed semi-automatic workflow achieves a more significant advantage by supporting the overall annotation process.

From another perspective, mainly related to the analyzed case study and considering the most common annotated terms per semantic category identified in Figure 6, the topmost discussed concepts related to the topic of “anatomical parts” owned a pre-existing relationship to blood components and gastrointestinal organs due to the nature of the disease. In contrast, the term “bone mineral density” (BMD) stood out due to the high number of documents that relate untreated gluten diseases to a greater tendency to suffer from fractures and a density improvement on a gluten-free diet [85]. Consequently, associated with BMD, the terms “Osteoporosis” and “Osteopenia” (both diseases) were also widely mentioned and related with celiac disease (CD) and GFD, in the same way as BMD [86,87].

With regard to cell types, the most discussed concepts were related to T-cells with inflammatory and immune roles, namely, "$CD4^+$", "T-lymphocytes", and "Intraepithelial lymphocytes", derived from the autoimmune nature of gluten-related diseases [88,89]. Similarly, the most mentioned proteins, besides the different protein fractions that constitute gluten, were related to antibodies closely associated with developing distinct health issues. In this sense, several scientific documents referring to these proteins discussed their relationship in diagnosing different illnesses or the benefits of the GFD. An example of this case was the relationship of the "IgA" and "IgG" autoantibodies against "tissue transglutaminase" [90,91]. Another protein that deserved attention due to its substantial presence in the bibliome was casein, the collective term for a family of milk proteins [92]. This protein is positively associated with non-gastrointestinal diseases, notably autism spectrum disorder (ASD), and the casein/gluten elimination from the diet is encouraged to improved ASD behaviors in children who reported some gastrointestinal symptoms [93,94].

Considering the diet semantic category, most of the identified concepts were related to discussing the advantages of a gluten-free diet in treating different diseases and their relation with the most annotated discussed symptoms like diarrhea, malabsorption, inflammation of the small intestine, and abdominal pain. In this sense, the curated studies evaluated the effect that GFD and other counterparts diets could produce in humans with different health issues [95,96].

The most mentioned compounds were iron, calcium, and other general nutrients, being discussed in a large number of documents signaling the alimentary unbalances produced by GFD [97], how GFD places compounds within the normal range [98], and also the difficulty of their absorption into the digestive system in related diseases [99,100].

Concerning diseases, numerous documents discussed the damaged relationship that gluten can produce to health issues related to the digestive system, like "Irritable bowel syndrome" [101] and "Type 1 diabetes" [102], but also for other diseases with a less apparent relationship, such as skin diseases [103] and psychological disorders, like "Autism" and "Schizophrenia" [104–106].

With respect to the food or food product category, the relevant discussed terms were related to cereals and, above all, oats. In this sense, recent studies have questioned the suitability of oats in the diet of gluten-related patients, with some authors claiming that oats pose no risk to celiac [107] and others arguing that a subgroup of celiac patients may be intolerant to oats [108]. As occurring with diet therapy, mentioned organisms were oriented towards applying different bacteria to reduce or degrade toxic gluten peptides [109,110].

Finally, concerning the genes category, the most identified terms were related to genes responsible for CD development, namely "HLA-DQA1" or "HLA-DQB1". These genes are present in 30-40% of the general population, but only a few percentages of carriers develop gluten-related diseases [111,112]. Similarly, the "myosin IXB" gene was discussed as a potential risk factor in inflammatory conditions, having a role in intestinal barrier functions with evidence of its association with CD, dermatitis herpetiformis, inflammatory bowel disease, systemic lupus erythematosus, and rheumatoid arthritis risk [113]. It is noteworthy that apart from these genes, few studies related other genes to this

subject, which may be motivated because the genes that predispose individuals to gluten-related disorders are very well defined.

6. Conclusions and future work

This work presents a semi-automatic ML workflow able to reduce the manual curation cost (i.e. annotation and classification) of thousands of documents downloaded from PubMed with the goal of generating a gold standard corpus. As a fundamental part of the proposed approach, it is introduced the NER+Ontology document description technique for the automatic classification of the bibliome. The practical relevance of the implemented workflow was demonstrated in the manual curation of 4,115 gluten-related documents, while the proposed NER+Ontology technique showed satisfactory results compared to other state-of-the-art document representation techniques in three different scenarios using well-known classifiers.

Future work will be focused on applying the best-ranked classifier for the automatic classification of the remaining bibliome, adopting the proposed NER+Ontology description technique as the current baseline for identifying novel relation patterns in overall. Although the experimental results have demonstrated the proper operation of the proposed approaches, it would be interesting to consider the exploitation of the remaining ontology capabilities to obtain better classification results, for example, the semantic hierarchy inference provided by the different ontologies. Finally, a parallel objective is related to structuring and making available the curated knowledge through an online database to assist researchers in making decisions and developing new hypotheses based on the bibliome.

Acknowledgments

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 financed by the Portuguese Foundation for Science and Technology (FCT/MCTES) Ref. UID/QUI/50006/2020; the Portuguese Foundation for Science and Technology (FCT/MCTES) under the scope of the strategic funding of UIDB/04469/2020 unit and BioTecNorte operation funded by the European Regional Development Fund (ERDF) under the scope of Norte2020—Programa Operacional Regional do Norte. Ref. NORTE-01-0145-FEDER-000004; the 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 "Centro singular de investigación de Galicia" (accreditation 2019-2022) funded by the European Regional Development Fund (ERDF)-Ref. ED431G2019/06. The authors also acknowledge the postdoctoral fellowship [ED481B-2019-032] of Martín Pérez-Pérez, funded by Xunta de Galicia.

References

[1] W. Raghupathi, V. Raghupathi, Big data analytics in healthcare: promise and potential, Heal. Inf. Sci. Syst. 2 (2014). https://doi.org/10.1186/2047-2501-2-3.

[2] H.C. Lyson, G.M. Le, J. Zhang, N. Rivadeneira, C. Lyles, K. Radcliffe, R.J. Pasick, G. Sawaya, U. Sarkar, D. Centola, Social Media as a Tool to Promote Health Awareness: Results from an Online Cervical Cancer Prevention Study, J. Cancer Educ. 34 (2019) 819–822. https://doi.org/10.1007/s13187-018-1379-8.

[3] M. Song, W.C. Kim, D. Lee, G.E. Heo, K.Y. Kang, PKDE4J: Entity and relation extraction for public knowledge discovery, J. Biomed. Inform. 57 (2015) 320–332. https://doi.org/10.1016/j.jbi.2015.08.008.

[4] O. Balmau, R. Guerraoui, A.M. Kermarrec, A. Maurer, M. Pavlovic, W. Zwaenepoel, The fake news vaccine: A content-agnostic system for preventing fake news from becoming viral, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), Springer, 2019: pp. 347–364. https://doi.org/10.1007/978-3-030-31277-0_23.

[5] M. Househ, E. Borycki, A. Kushniruk, Empowering patients through social media: The benefits and challenges, Health Informatics J. 20 (2014) 50–58. https://doi.org/10.1177/1460458213476969.

[6] A. Holzinger, I. Jurisica, Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 8401 (2014) 1–18. https://doi.org/10.1007/978-3-662-43968-5_1.

[7] S.K. Vanga, A. Singh, B. Harish Vagadia, V. Raghavan, Global food allergy research trend: a bibliometric analysis, Scientometrics. 105 (2015) 203–213. https://doi.org/10.1007/s11192-015-1660-0.

[8] M. Pérez-Pérez, P. Jorge, G. Pérez Rodríguez, M.O. Pereira, A. Lourenço, Quorum sensing inhibition in Pseudomonas aeruginosa biofilms: new insights through network mining, Biofouling. 33 (2017) 128–142. https://doi.org/10.1080/08927014.2016.1272104.

[9] A. Singhal, R. Leaman, N. Catlett, T. Lemberger, J. McEntyre, S. Polson, I. Xenarios, C. Arighi, Z. Lu, Pressing needs of biomedical text mining in biocuration and beyond: Opportunities and challenges, Database. 2016 (2016). https://doi.org/10.1093/database/baw161.

[10] Q. Wang, S. S Abdul, L. Almeida, S. Ananiadou, Y.I. Balderas-Martínez, R. Batista-Navarro, D. Campos, L. Chilton, H.-J. Chou, G. Contreras, L. Cooper, H.-J. Dai, B. Ferrell, J. Fluck, S. Gama-Castro, N. George, G. Gkoutos, A.K. Irin, L.J. Jensen, S. Jimenez, T.R. Jue, I. Keseler, S. Madan, S. Matos, P. McQuilton, M. Milacic, M. Mort, J. Natarajan, E. Pafilis, E. Pereira, S. Rao, F. Rinaldi, K. Rothfels, D. Salgado, R.M. Silva, O. Singh, R. Stefancsik, C.-H. Su, S. Subramani, H.D. Tadepally, L. Tsaprouni, N. Vasilevsky, X. Wang, A. Chatr-Aryamontri, S.J.F. Laulederkind, S. Matis-Mitchell, J. McEntyre, S. Orchard, S. Pundir, R. Rodriguez-Esteban, K. Van Auken, Z. Lu, M. Schaeffer, C.H. Wu, L. Hirschman, C.N. Arighi, Overview of the interactive task in BioCreative V., Database (Oxford). 2016 (2016). https://doi.org/10.1093/database/baw119.

[11] CNIO Centro Nacional de Investigaciones Oncológicas., Coordination and edition Martin Krallinger & Alfonso Valencia, Proceedings of the BioCreative V.5 Challenge Evaluation Workshop, in: M.K.& A. Valencia (Ed.), Proc. BioCreative V.5 Chall. Eval. Work., Fundación CNIO Carlos III, 2017, 2017: pp. 8–27.

http://www.biocreative.org/media/store/files/2017/BioCreative_V.5_Proceedings.pdf (accessed May 9, 2018).

[12] S. Pyysalo, T. Ohta, R. Rak, A. Rowley, H.W. Chun, S.J. Jung, S.P. Choi, J. Tsujii, S. Ananiadou, Overview of the Cancer Genetics and Pathway Curation tasks of BioNLP Shared Task 2013, BMC Bioinformatics. 16 (2015). https://doi.org/10.1186/1471-2105-16-S10-S2.

[13] N. Collier, Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, (2004) 507505. http://www.genisis.ch/~natlang/NLPBA02/ (accessed November 4, 2020).

[14] P.D. Karp, Crowd-sourcing and author submission as alternatives to professional curation, Database. 2016 (2016) 149. https://doi.org/10.1093/database/baw149.

[15] S. Orchard, H. Hermjakob, Shared resources, shared costs-leveraging biocuration resources, Database. 2015 (2015) 9. https://doi.org/10.1093/database/bav009.

[16] W.A. Baumgartner, K.B. Cohen, L.M. Fox, G. Acquaah-Mensah, L. Hunter, Manual curation is not sufficient for annotation of genomic databases, in: Bioinformatics, 2007. https://doi.org/10.1093/bioinformatics/btm229.

[17] K.Z. Vardakas, G. Tsopanakis, A. Poulopoulou, M.E. Falagas, An analysis of factors contributing to PubMed's growth, J. Informetr. (2015). https://doi.org/10.1016/j.joi.2015.06.001.

[18] À. Bravo, J. Piñero, N. Queralt-Rosinach, M. Rautschka, L.I. Furlong, Extraction of relations between genes and diseases from text and large-scale data analysis: Implications for translational research, BMC Bioinformatics. 16 (2015) 55. https://doi.org/10.1186/s12859-015-0472-9.

[19] C.C. Huang, Z. Lu, Community challenges in biomedical text mining over 10 years: Success, failure and the future, Brief. Bioinform. 17 (2016) 132–144. https://doi.org/10.1093/bib/bbv024.

[20] C.N. Arighi, Z. Lu, M. Krallinger, K.B. Cohen, W.J. Wilbur, A. Valencia, L. Hirschman, C.H. Wu, Overview of the BioCreative III Workshop., BMC Bioinformatics. 12 Suppl 8 (2011) S1. https://doi.org/10.1186/1471-2105-12-S8-S1.

[21] P.D. Karp, Can we replace curation with information extraction software?, Database. 2016 (2016). https://doi.org/10.1093/database/baw150.

[22] C.-H. Wei, B.R. Harris, D. Li, T.Z. Berardini, E. Huala, H.-Y. Kao, Z. Lu, Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts, Database. 2012 (2012) bas041–bas041. https://doi.org/10.1093/database/bas041.

[23] M. Martinez-Alvarez, S. Yahyaei, T. Roelleke, Semi-automatic document classification: Exploiting document difficulty, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), Springer, Berlin, Heidelberg, 2012: pp. 468–471. https://doi.org/10.1007/978-3-642-28997-2_43.

[24] M.A.M. García, R.P. Rodríguez, L.E. Anido Rifón, Biomedical literature classification using encyclopedic knowledge: A Wikipedia-based bag-of-concepts approach, PeerJ. 2015 (2015). https://doi.org/10.7717/peerj.1279.

[25] Y. Chen, Y. Sun, B.Q. Han, Improving Classification of Protein Interaction Articles Using Context Similarity-Based Feature Selection, Biomed Res. Int. 2015 (2015). https://doi.org/10.1155/2015/751646.

[26] S. Matos, Improving document prioritization for protein-protein interaction extraction using shallow linguistics and word embeddings, in: Adv. Intell. Syst. Comput., Springer Verlag, 2017: pp. 43–49. https://doi.org/10.1007/978-3-319-

60816-7_6.

[27] P. Jorge, M. Perez-Perez, G.P. Rodriguez, F. Fdez-Riverola, M.O. Pereira, A. Lourenco, Construction of antimicrobial peptide-drug combination networks from scientific literature based on a semi-automated curation workflow, Database. 2016 (2016) 14310–1093. https://doi.org/10.1093/database/baw143.

[28] L. Luo, Z. Yang, L. Wang, Y. Zhang, H. Lin, J. Wang, L. Yang, K. Xu, Y. Zhang, Protein-Protein Interaction Article Classification: A Knowledge-enriched Self-Attention Convolutional Neural Network Approach, in: Proc. - 2018 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2018, Institute of Electrical and Electronics Engineers Inc., 2019: pp. 467–469. https://doi.org/10.1109/BIBM.2018.8621362.

[29] M. Kulmanov, F.Z. Smaili, X. Gao, R. Hoehndorf, Semantic similarity and machine learning with ontologies, Brief. Bioinform. 2020 (2020) 1–18. https://doi.org/10.1093/bib/bbaa199.

[30] H. Ding, I. Takigawa, H. Mamitsuka, S. Zhu, Similarity-based machine learning methods for predicting drug–target interactions: a brief review, Brief. Bioinform. 15 (2014) 734–747. https://doi.org/10.1093/bib/bbt056.

[31] N. Sanchez-Pi, L. Martí, A.C. Bicharra Garcia, Improving ontology-based text classification: An occupational health and security application, J. Appl. Log. 17 (2016) 48–58. https://doi.org/10.1016/j.jal.2015.09.008.

[32] D. Kwon, S. Kim, S.-Y. Shin, A. Chatr-aryamontri, W.J. Wilbur, Assisting manual literature curation for protein-protein interactions using BioQRator, Database. 2014 (2014) bau067–bau067. https://doi.org/10.1093/database/bau067.

[33] J. Szostak, S. Ansari, S. Madan, J. Fluck, M. Talikka, A. Iskandar, H. De Leon, M. Hofmann-Apitius, M.C. Peitsch, J. Hoeng, Construction of biological networks from unstructured information based on a semi-automated curation workflow, Database. 2015 (2015) 1–14. https://doi.org/10.1093/database/bav057.

[34] F. Rinaldi, O. Lithgow, S. Gama-Castro, H. Solano, A. López-Fuentes, L.J. Muñiz Rascado, C. Ishida-Gutiérrez, C.-F. Méndez-Cruz, J. Collado-Vides, Strategies towards digital and semi-automated curation in RegulonDB, Database. 2017 (2017) 1–11. https://doi.org/10.1093/database/bax012.

[35] R. Winnenburg, T. Wächter, C. Plake, A. Doms, M. Schroeder, Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies?, Brief. Bioinform. 9 (2008) 466–478. https://doi.org/10.1093/bib/bbn043.

[36] G. Zong, B. Lebwohl, F.B. Hu, L. Sampson, L.W. Dougherty, W.C. Willett, A.T. Chan, Q. Sun, Gluten intake and risk of type 2 diabetes in three large prospective cohort studies of US men and women, Diabetologia. 61 (2018) 2164–2173. https://doi.org/10.1007/s00125-018-4697-9.

[37] U. of B. CHEQ, The Economic Cost of Bad Actors on the Internet, (2019) 17. https://s3.amazonaws.com/media.mediapost.com/uploads/EconomicCostOfFakeNews.pdf (accessed November 10, 2020).

[38] K.M. Di Sebastiano, G. Murthy, K.L. Campbell, S. Desroches, R.A. Murphy, Nutrition and Cancer Prevention: Why is the Evidence Lost in Translation?, Adv. Nutr. 10 (2019) 410–418. https://doi.org/10.1093/advances/nmy089.

[39] M. Passali, K. Josefsen, J.L. Frederiksen, J.C. Antvorskov, Current evidence on the efficacy of gluten-free diets in multiple sclerosis, psoriasis, type 1 diabetes and autoimmune thyroid diseases, Nutrients. 12 (2020) 1–26. https://doi.org/10.3390/nu12082316.

[40] R. Krysiak, W. Szkróbka, B. Okopień, The Effect of Gluten-Free Diet on Thyroid Autoimmunity in Drug-Naïve Women with Hashimoto's Thyroiditis: A Pilot

Study, Exp. Clin. Endocrinol. Diabetes. 127 (2019) 417–422. https://doi.org/10.1055/a-0653-7108.

[41] B. Palmieri, M. Vadalà, C. Laurino, Gluten-free diet in non-celiac patients: Beliefs, truths, advantages and disadvantages, Minerva Gastroenterol. Dietol. 65 (2019) 153–162. https://doi.org/10.23736/S1121-421X.18.02519-9.

[42] R. Mesnage, M.N. Antoniou, Facts and Fallacies in the Debate on Glyphosate Toxicity, Front. Public Heal. 5 (2017). https://doi.org/10.3389/fpubh.2017.00316.

[43] G.A. Gaesser, S.S. Angadi, Navigating the gluten-free boom, J. Am. Acad. Physician Assist. 28 (2015). https://doi.org/10.1097/01.JAA.0000469434.67572.a4.

[44] C. Newberry, L. McKnight, M. Sarav, O. Pickett-Blakely, Going Gluten Free: the History and Nutritional Implications of Today's Most Popular Diet, Curr. Gastroenterol. Rep. 19 (2017) 1–8. https://doi.org/10.1007/s11894-017-0597-2.

[45] N.P. Chandrasekarasastry, K.M. Verspoor, Q. Chen, N.C. Panyam, A. Elangovan, M. Davis, K. Verspoor, Document Triage and Relation Extraction for Protein-Protein Interactions affected by Mutations, 2017. https://www.researchgate.net/publication/322852231 (accessed April 23, 2021).

[46] X. Jiang, M. Ringwald, J. Blake, H. Shatkay, Effective biomedical document classification for identifying publications relevant to the mouse Gene Expression Database (GXD), Database. 2017 (2017) 17. https://doi.org/10.1093/database/bax017.

[47] P. Jorge, M. Pérez-Pérez, G. Rodríguez, F. Fdez-Riverola, M. Pereira, A. Lourenço, Reconstruction of the network of experimentally validated AMP-drug combinations against Pseudomonas aeruginosa infections, Curr. Bioinform. 11 (2016). https://doi.org/10.2174/1574893611666160617093955.

[48] T. Barrett, J. Beck, D.A. Benson, C. Bollin, E. Bolton, D. Bourexis, J.R. Brister, S.H. Bryant, K. Canese, K. Clark, M. Dicuccio, I. Dondoshansky, S. Federhen, M. Feolo, K. Funk, L.Y. Geer, V. Gorelenkov, M. Hoeppner, B. Holmes, M. Johnson, V. Khotomlianski, A. Kimchi, M. Kimelman, P. Kitts, W. Klimke, S. Krasnov, A. Kuznetsov, M.J. Landrum, D. Landsman, J.M. Lee, D.J. Lipman, Z. Lu, T.L. Madden, T. Madej, A. Marchler-Bauer, I. Karsch-Mizrachi, T. Murphy, R. Orris, J. Ostell, C. O'Sullivan, A. Panchenko, L. Phan, D. Preuss, K.D. Pruitt, W. Rubinstein, E.W. Sayers, V. Schneider, G.D. Schuler, S.T. Sherry, K. Sirotkin, K. Siyan, D. Slotta, A. Soboleva, V. Soussov, G. Starchenko, T.A. Tatusova, B.W. Trawick, D. Vakatov, Y. Wang, M. Ward, W.J. Wilbur, E. Yaschenko, K. Zbicz., Database resources of the National Center for Biotechnology Information, Nucleic Acids Res. 43 (2015) D6–D17. https://doi.org/10.1093/nar/gku1130.

[49] D.M. Dooley, E.J. Griffiths, G.S. Gosal, P.L. Buttigieg, R. Hoehndorf, M.C. Lange, L.M. Schriml, F.S.L. Brinkman, W.W.L. Hsiao, Food on: A harmonized food ontology to increase global food traceability, quality control and data integration, Npj Sci. Food. 2 (2018) 1–10. https://doi.org/10.1038/s41538-018-0032-6.

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

[51] S.J. Nelson, W.D. Johnston, B.L. Humphreys, Relationships in Medical Subject Headings (MeSH), in: Springer, Dordrecht, 2001: pp. 171–184. https://doi.org/10.1007/978-94-015-9696-1_11.

[52] P. de Matos, R. Alcántara, A. Dekker, M. Ennis, J. Hastings, K. Haug, I. Spiteri, S. Turner, C. Steinbeck, Chemical entities of biological interest: An update,

Nucleic Acids Res. 38 (2009). https://doi.org/10.1093/nar/gkp886.

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

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

[55] W.A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, C.J. Mungall, J.X. Binder, J. Malone, D. Vasant, H. Parkinson, L.M. Schriml, Disease Ontology 2015 update: An expanded and updated database of Human diseases for linking biomedical knowledge through disease data, Nucleic Acids Res. 43 (2015) D1071–D1078. https://doi.org/10.1093/nar/gku1011.

[56] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. MacIejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, Di. Le, A. Pon, C. Knox, M. Wilson, DrugBank 5.0: A major update to the DrugBank database for 2018, Nucleic Acids Res. 46 (2018) D1074–D1082. https://doi.org/10.1093/nar/gkx1037.

[57] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: New perspectives on genomes, pathways, diseases and drugs, Nucleic Acids Res. 45 (2017) D353–D361. https://doi.org/10.1093/nar/gkw1092.

[58] C.F. Thorn, T.E. Klein, R.B. Altman, PharmGKB: The pharmacogenomics knowledge base, Methods Mol. Biol. 1015 (2013) 311–320. https://doi.org/10.1007/978-1-62703-435-7_20.

[59] A. Bateman, M.J. Martin, C. O'Donovan, M. Magrane, R. Apweiler, E. Alpi, R. Antunes, J. Arganiska, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, G. Chavali, E. Cibrian-Uhalte, A. Da Silva, M. De Giorgi, T. Dogan, F. Fazzini, P. Gane, L.G. Castro, P. Garmiri, E. Hatton-Ellis, R. Hieta, R. Huntley, D. Legge, W. Liu, J. Luo, A. Macdougall, P. Mutowo, A. Nightingale, S. Orchard, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Turner, V. Volynkin, T. Wardell, X. Watkins, H. Zellner, A. Cowley, L. Figueira, W. Li, H. McWilliam, R. Lopez, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. Argoud-Puy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M.C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. De Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A.L. Veuthey, C.H. Wu, C.N. Arighi, L. Arminski, C. Chen, Y. Chen, J.S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D.A. Natale, B.E. Suzek, C.R. Vinayaka, Q. Wang, Y. Wang, L.S. Yeh, M.S. Yerramalla, J. Zhang, UniProt: A hub for protein information, Nucleic Acids Res. 43 (2015) D204–D212. https://doi.org/10.1093/nar/gku989.

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

[61] M. Gerner, G. Nenadic, C.M. Bergman, LINNAEUS: A species name identification system for biomedical literature, BMC Bioinformatics. 11 (2010) 85. https://doi.org/10.1186/1471-2105-11-85.

[62] B. Settles, ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text, Bioinformatics. 21 (2005) 3191–3192. https://doi.org/10.1093/bioinformatics/bti475.

[63] D.M. Jessop, S.E. Adams, E.L. Willighagen, L. Hawizy, P. Murray-Rust, OSCAR4: A flexible architecture for chemical textmining, J. Cheminform. 3 (2011) 41. https://doi.org/10.1186/1758-2946-3-41.

[64] R. Leaman, C.H. Wei, Z. Lu, TmChem: A high performance approach for chemical named entity recognition and normalization, J. Cheminform. 7 (2015) S3. https://doi.org/10.1186/1758-2946-7-S1-S3.

[65] R. Leaman, R.I. Doğan, Z. Lu, DNorm: Disease name normalization with pairwise learning to rank, Bioinformatics. 29 (2013) 2909–2917. https://doi.org/10.1093/bioinformatics/btt474.

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

[67] W. Zhang, X. Tang, T. Yoshida, TESC: An approach to TExt classification using Semi-supervised Clustering, Knowledge-Based Syst. 75 (2015) 152–160. https://doi.org/10.1016/j.knosys.2014.11.028.

[68] B. Guo, C. Zhang, J. Liu, X. Ma, Improving text classification with weighted word embeddings via a multi-channel TextCNN model, Neurocomputing. 363 (2019) 366–374. https://doi.org/10.1016/j.neucom.2019.07.052.

[69] B. Schölkopf, SVMs - A practical consequence of learning theory, IEEE Intell. Syst. Their Appl. 13 (1998) 18–21. https://doi.org/10.1109/5254.708428.

[70] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32. https://doi.org/10.1023/A:1010933404324.

[71] C.E. McCulloch, Generalized Linear Models, J. Am. Stat. Assoc. 95 (2000) 1320–1324. https://doi.org/10.1080/01621459.2000.10474340.

[72] Probabilistic Networks and Expert Systems, Springer-Verlag, 1999. https://doi.org/10.1007/b97670.

[73] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLINEAR: A Library for Large Linear Classification, 2008. http://www.csie.ntu.edu.tw/ (accessed December 29, 2020).

[74] I. Inuwa-Dutse, M. Liptrott, I. Korkontzelos, Detection of spam-posting accounts on Twitter, Neurocomputing. 315 (2018) 496–511. https://doi.org/10.1016/j.neucom.2018.07.044.

[75] H.K. Kim, H. Kim, S. Cho, Bag-of-concepts: Comprehending document representation through clustering words in distributed representation, Neurocomputing. 266 (2017) 336–352. https://doi.org/10.1016/j.neucom.2017.05.046.

[76] E. Dynomant, R. Lelong, B. Dahamna, C. Massonnaud, G. Kerdelhué, J. Grosjean, S. Canu, S. Darmoni, Word embedding for French natural language in healthcare: A comparative study, in: Stud. Health Technol. Inform., IOS Press, 2019: pp. 118–122. https://doi.org/10.3233/SHTI190195.

[77] K. Jiang, S. Feng, Q. Song, R.A. Calix, M. Gupta, G.R. Bernard, Identifying tweets of personal health experience through word embedding and LSTM neural network, BMC Bioinformatics. 19 (2018). https://doi.org/10.1186/s12859-018-2198-y.

[78] M.N. Hamid, I. Friedberg, Identifying antimicrobial peptides using word

embedding with deep recurrent neural networks, Bioinformatics. 35 (2019) 2009–2016. https://doi.org/10.1093/bioinformatics/bty937.

[79] G. Rao, W. Huang, Z. Feng, Q. Cong, LSTM with sentence representations for document-level sentiment classification, Neurocomputing. 308 (2018) 49–57. https://doi.org/10.1016/j.neucom.2018.04.045.

[80] C. Wu, J. Su, Y. Chen, X. Shi, Boosting implicit discourse relation recognition with connective-based word embeddings, Neurocomputing. 369 (2019) 39–49. https://doi.org/10.1016/j.neucom.2019.08.081.

[81] T. Mikolov, K. Chen, G. Corrado, J. Dean, Distributed Representations of Words and Phrases and their Compositionality, 2013.

[82] P.D. Turney, P. Pantel, From frequency to meaning: Vector space models of semantics, J. Artif. Intell. Res. 37 (2010) 141–188. https://doi.org/10.1613/jair.2934.

[83] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, S. Ananiadou, Distributional Semantics Resources for Biomedical Text Processing, n.d. https://github.com/spyysalo/nxml2txt (accessed October 13, 2020).

[84] B.U. Ca, Y.G. Fr, No Unbiased Estimator of the Variance of K-Fold Cross-Validation Yoshua Bengio Yves Grandvalet, 2004.

[85] M.I. Pinto-Sanchez, J.C. Bai, Toward New Paradigms in the Follow Up of Adult Patients With Celiac Disease on a Gluten-Free Diet, Front. Nutr. 6 (2019). https://doi.org/10.3389/fnut.2019.00153.

[86] G. Valerio, R. Spadaro, D. Iafusco, F. Lombardi, A. Del Puente, A. Esposito, F. De Terlizzi, F. Prisco, R. Troncone, A. Franzese, The influence of gluten free diet on quantitative ultrasound of proximal phalanxes in children and adolescents with type 1 diabetes mellitus and celiac disease, Bone. 43 (2008) 322–326. https://doi.org/10.1016/j.bone.2008.04.004.

[87] N. McGough, J.H. Cummings, Coeliac disease: a diverse clinical syndrome caused by intolerance of wheat, barley and rye, Proc. Nutr. Soc. 64 (2005) 434–450. https://doi.org/10.1079/pns2005461.

[88] J. Huan, R. Meza-Romero, J.L. Mooney, A.A. Vandenbark, H. Offner, G.G. Burrows, Single-chain recombinant HLA-DQ2.5/peptide molecules block α2-gliadin-specific pathogenic CD4 T-cell proliferation and attenuate production of inflammatory cytokines: A potential therapy for celiac disease, Mucosal Immunol. 4 (2011) 112–120. https://doi.org/10.1038/mi.2010.44.

[89] J. V. Steenholt, C. Nielsen, L. Baudewijn, A. Staal, K.S. Rasmussen, H.J. Sabir, T. Barington, S. Husby, H. Toft-Hansen, The composition of T cell subtypes in duodenal biopsies are altered in coeliac disease patients, PLoS One. 12 (2017) 1–17. https://doi.org/10.1371/journal.pone.0170270.

[90] E. Tonutti, N. Bizzaro, Diagnosis and classification of celiac disease and gluten sensitivity, Autoimmun. Rev. 13 (2014) 472–476. https://doi.org/10.1016/j.autrev.2014.01.043.

[91] G. Byrne, C. Feighery, J. Jackson, J. Kelly, Coeliac disease autoantibodies mediate significant inhibition of tissue transglutaminase, Clin. Immunol. 136 (2010) 426–431. https://doi.org/10.1016/j.clim.2010.04.017.

[92] T.K. Głąb, J. Boratyński, Potential of Casein as a Carrier for Biologically Active Agents, Top. Curr. Chem. 375 (2017) 71. https://doi.org/10.1007/s41061-017-0158-z.

[93] C. Millward, M. Ferriter, S. Calver, G. Connell-Jones, Gluten- and casein-free diets for autistic spectrum disorder, in: Cochrane Database Syst. Rev., John Wiley & Sons, Ltd, 2004. https://doi.org/10.1002/14651858.cd003498.pub2.

[94] C.M. Pennesi, L.C. Klein, Effectiveness of the gluten-free, casein-free diet for children diagnosed with autism spectrum disorder: Based on parental report, Nutr. Neurosci. 15 (2012) 85–91. https://doi.org/10.1179/1476830512Y.0000000003.

[95] H. Li, U. Bose, S. Stockwell, C.A. Howitt, M. Colgrave, Assessing the utility of multiplexed liquid chromatography-mass spectrometry for gluten detection in Australian Breakfast food products, Molecules. 24 (2019) 1–14. https://doi.org/10.3390/molecules24203665.

[96] L.L.E. Koskinen, I.R. Korponay-Szabo, K. Viiri, K. Juuti-Uusitalo, K. Kaukinen, K. Lindfors, K. Mustalahti, K. Kurppa, R. Ádány, Z. Pocsai, G. Széles, E. Einarsdottir, C. Wijmenga, M. Mäki, J. Partanen, J. Kere, P. Saavalainen, Myosin IXB gene region and gluten intolerance: Linkage to coeliac disease and a putative dermatitis herpetiformis association, J. Med. Genet. 45 (2008) 222–227. https://doi.org/10.1136/jmg.2007.053991.

[97] E.G.D. Hopman, S. Le Cessie, B.M.E. Von Blomberg, M.L. Mearin, Nutritional management of the gluten-free diet in young people with celiac disease in The Netherlands, J. Pediatr. Gastroenterol. Nutr. 43 (2006) 102–108. https://doi.org/10.1097/01.mpg.0000228102.89454.eb.

[98] A. Vilppula, K. Kaukinen, L. Luostarinen, I. Krekelä, H. Patrikainen, R. Valve, M. Luostarinen, K. Laurila, M. Mäki, P. Collin, Clinical benefit of gluten-free diet in screen-detected older celiac disease patients, BMC Gastroenterol. 11 (2011) 136. https://doi.org/10.1186/1471-230X-11-136.

[99] M. Pazianas, G.P. Butcher, J.M. Subhani, P.J. Finch, L. Ang, C. Collins, R.P. Heaney, M. Zaidi, J.D. Maxwell, Calcium absorption and bone mineral density in celiacs after long term treatment with gluten-free diet and adequate calcium intake, Osteoporos. Int. 16 (2005) 56–63. https://doi.org/10.1007/s00198-004-1641-2.

[100] U. Krupa-Kozak, N. Drabińska, Calcium in Gluten-Free Life: Health-Related and Nutritional Implications, Foods. 5 (2016) 51. https://doi.org/10.3390/foods5030051.

[101] V.G. Zanwar, S. V. Pawar, P.A. Gambhire, S.S. Jain, R.G. Surude, V.B. Shah, Q.Q. Contractor, P.M. Rathi, Symptomatic improvement with gluten restriction in irritable bowel syndrome: A prospective, randomized, double blinded placebo controlled trial, Intest. Res. 14 (2016) 343–350. https://doi.org/10.5217/ir.2016.14.4.343.

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

[103] T.T. Salmi, K. Hervonen, K. Kurppa, P. Collin, K. Kaukinen, T. Reunala, Celiac disease evolving into dermatitis herpetiformis in patients adhering to normal or gluten-free diet, Scand. J. Gastroenterol. 50 (2015) 387–392. https://doi.org/10.3109/00365521.2014.974204.

[104] C. Millward, M. Ferriter, S.J. Calver, G.G. Connell-Jones, WITHDRAWN: Gluten- and casein-free diets for autistic spectrum disorder, Cochrane Database Syst. Rev. 4 (2019) CD003498. https://doi.org/10.1002/14651858.CD003498.pub4.

[105] C.M. Pennesi, L.C. Klein, Effectiveness of the gluten-free, casein-free diet for children diagnosed with autism spectrum disorder: Based on parental report, Nutr. Neurosci. 15 (2012) 85–91. https://doi.org/10.1179/1476830512Y.0000000003.

[106] A.E. Kalaydjian, W. Eaton, N. Cascella, A. Fasano, The gluten connection: The association between schizophrenia and celiac disease, Acta Psychiatr. Scand. 113 (2006) 82–90. https://doi.org/10.1111/j.1600-0447.2005.00687.x.

[107] M.I. Pinto-Sánchez, N. Causada-Calo, P. Bercik, A.C. Ford, J.A. Murray, D. Armstrong, C. Semrad, S.S. Kupfer, A. Alaedini, P. Moayyedi, D.A. Leffler, E.F. Verdú, P. Green, Safety of Adding Oats to a Gluten-Free Diet for Patients With Celiac Disease: Systematic Review and Meta-analysis of Clinical and Observational Studies, Gastroenterology. 153 (2017) 395-409.e3. https://doi.org/10.1053/j.gastro.2017.04.009.

[108] P. Fric, D. Gabrovska, J. Nevoral, Celiac disease, gluten-free diet, and oats, Nutr. Rev. 69 (2011) 107–115. https://doi.org/10.1111/j.1753-4887.2010.00368.x.

[109] M. Fernandez-Feo, G. Wei, G. Blumenkranz, F.E. Dewhirst, D. Schuppan, F.G. Oppenheim, E.J. Helmerhorst, The cultivable human oral gluten-degrading microbiome and its potential implications in coeliac disease and gluten sensitivity, Clin. Microbiol. Infect. 19 (2013) E386–E394. https://doi.org/10.1111/1469-0691.12249.

[110] G. Wei, N. Tian, R. Siezen, D. Schuppan, E.J. Helmerhorst, Identification of food-grade subtilisins as gluten-degrading enzymes to treat celiac disease, Am. J. Physiol. - Gastrointest. Liver Physiol. 311 (2016) G571–G580. https://doi.org/10.1152/ajpgi.00185.2016.

[111] Y. Kooy-Winkelaar, M. van Lummel, A.K. Moustakas, J. Schweizer, M.L. Mearin, C.J. Mulder, B.O. Roep, J.W. Drijfhout, G.K. Papadopoulos, J. van Bergen, F. Koning, Gluten-Specific T Cells Cross-React between HLA-DQ8 and the HLA-DQ2α/DQ8β Transdimer, J. Immunol. 187 (2011) 5123–5129. https://doi.org/10.4049/jimmunol.1101179.

[112] L.M. Sollid, S.W. Qiao, R.P. Anderson, C. Gianfrani, F. Koning, Nomenclature and listing of celiac disease relevant gluten T-cell epitopes restricted by HLA-DQ molecules, Immunogenetics. 64 (2012) 455–460. https://doi.org/10.1007/s00251-012-0599-z.

[113] L.L.E. Koskinen, I.R. Korponay-Szabo, K. Viiri, K. Juuti-Uusitalo, K. Kaukinen, K. Lindfors, K. Mustalahti, K. Kurppa, R. Ádány, Z. Pocsai, G. Széles, E. Einarsdottir, C. Wijmenga, M. Mäki, J. Partanen, J. Kere, P. Saavalainen, Myosin IXB gene region and gluten intolerance: Linkage to coeliac disease and a putative dermatitis herpetiformis association, J. Med. Genet. 45 (2008) 222–227. https://doi.org/10.1136/jmg.2007.053991.

Martín Pérez Pérez is a PhD in Computer Science of the University of Vigo. He is currently at the SING group, where her research is focused on the fields of text mining, social mining, machine learning and artificial intelligence, applied to biomedical areas.

Tânia Ferreira is a MS student of bioinformatics at UTAD and received the BS degree in Genetics and Biotechnology from UTAD.

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

Gilberto Igrejas has a degree in Biology-Geology, master in Genetic Resources and Improvement of Agricultural and Forest Species and PhD in Genetics and Biotechnology by UTAD. He is the coordinator of the Unit for Functional Genomics and Proteomics of the University of Trás-os-Montes and Alto Douro (UTAD), full member of the Research Group (Bio) Chem & OMICS, Associated Laboratory for Green Chemistry, LAQV-REQUIMTE, Faculty of Science and Technology from Universidade Nova de Lisboa.

Florentino Fdez-Riverola received the BS degree in Computer Science from University of Oviedo. He also received the MS and PhD degrees from University of Vigo, where he is Full Professor in the Department of Computer Science and Coordinator of the Next Generation Computer System group (http://sing-group.org/). F. Fdez-Riverola has published 130 research papers in JCR indexed journals, 78 of which correspond to the first quartile (representing 60% of the total) and 24 of which correspond to the first decile (representing 18% of the total). More importantly, some of these studies (particularly those from the last few years) have received various awards from editorial committees from JCR Q1 journals, and have been selected as Editor's Choice paper, Highly Accessed or Top 25 Hottest Articles.

6.1 Conflict of Interest

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

CRediT author statement

Martín Pérez-Pérez: Investigation; Software; Methodology; Writing – Original Draft. **Tânia Ferreira**: Investigation; Data curation; Visualization; Writing – Original Draft. **Anália Lourenço**: Formal analysis; Methodology; Writing – Original Draft. **Gilberto Igrejas**: Conceptualization; Formal analysis; Resources; Writing – Review & Editing. **Florentino Fdez-Riverola**: Conceptualization; Project administration; Supervision; Writing – Review & Editing.

6.2 Declaration of Interest Statement

On behalf of all the authors, Dr. Florentino Fdez-Riverola (corresponding author) declare no conflict of interest.

Table 1: Initial feature description vector for each document.

Document	**TF-IDF terms**			**#Normalised annotations**			**#Domain count**		
	celiac	**allergy**	**…**	**Glu-D1**	**DOID:9892**	**…**	**Diseases**	**Compounds**	**…**
Doc 1	0.6	0.30	…	2	5	…	2	8	…
Doc 2	0.32	0	…	0	2	…	5	3	…
…	…	…	…	…	…	…	…	…	…
Doc *n*	…	…	…	…	…	…	…	…	…

Table 2: Gold standard dataset manually curated by experts.

	Curated documents
Relevant	1,871 (45.5%)
Irrelevant	2,244 (54.5%)
∑	4,115

Table 3: Correlation between different annotation types and the number of annotations per semantic category in relevant documents. Explicit mentions of general terms in the analyzed domain (e.g. “celiac”, “gluten” or “protein”) were not considered for the generation of the table. ANA stands for anatomy, CEL

stands for cell types, COMP stands for compounds, DIS stands for diseases, FOOD stands for food or food products, ORG stands for organisms, PROT represents proteins, and SYMP stands for symptoms. In terms of summarization ∑Corr represents the summarization of all correlations and ∑Ann represents the summarization of all annotations.

	ANA	CEL	COMP	DIET	DIS	FOOD	GENE	ORG	PROT	SYMP	**∑Corr**
ANA		0.739	0.749	0.076	0.930	0.550	0.513	0.399	1.001	0.801	**5.760**
CEL			0.244	-0.061	0.285	0.190	0.391	0.082	0.753	0.223	**2.846**
COMP				0.078	0.579	0.446	0.174	0.195	0.714	0.421	**1.640**
DIET					0.182	0.032	-0.060	-0.025	0.010	0.079	**0.286**
DIS						0.438	0.386	0.167	0.854	0.873	**2.057**
FOOD							0.168	0.332	0.615	0.177	**0.509**
GENE								0.083	0.674	0.219	**0.970**
ORG									0.366	0.086	**0.086**
PROT										0.631	**3.865**
SYMP											**0.086**
∑Ann	**3432**	**1074**	**2040**	**99**	**3693**	**1987**	**636**	**315**	**5975**	**2380**	

Table 4: Correlation between different annotation types and the number of annotations per semantic category in irrelevant documents. Explicit mentions of general terms in the analyzed domain (e.g. "celiac", "gluten" or "protein") were not considered for the generation of the table. ANA stands for anatomy, CEL stands for cell types, COMP stands for compounds, DIS stands for diseases, FOOD stands for food or food products, ORG stands for organisms, PROT represents proteins, and SYMP stands for symptoms. In terms of summarization ∑Corr represents the summarization of all correlations and ∑Ann represents the summarization of all annotations.

	ANA	CEL	COMP	DIET	DIS	FOOD	GENE	ORG	PROT	SYMP	**∑Corr**
ANA		0.337	0.752	0.131	0.422	0.573	0.368	0.328	0.630	0.415	**3.957**
CEL			0.250	-0.001	0.179	0.167	0.130	0.035	0.399	0.189	**1.685**
COMP				0.184	0.362	1.079	0.535	0.707	0.876	0.392	**2.540**
DIET					0.128	0.130	-0.022	-0.016	0.033	0.079	**0.484**
DIS						0.300	0.070	-0.039	0.507	0.420	**1.043**
FOOD							0.588	0.764	0.806	0.243	**1.007**
GENE								0.328	0.455	0.062	**1.560**
ORG									0.350	0.070	**0.070**
PROT										0.406	**3.434**
SYMP											**0.070**
∑Ann	**1961**	**335**	**5450**	**86**	**1149**	**5472**	**960**	**816**	**4200**	**686**	

Table 5: Performance comparison of the two document representation techniques under a 10-fold cross-validation scenario using the first set of 1,000 curated documents.

	Word embedding			**NER+Ontology**			**Gain**
	Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F-score*	*F-score*
SVM	0.717	0.653	0.683	0.766	0.861	0.810	+0.127
RF	0.702	0.694	**0.697**	0.826	0.815	**0.820**	+0.123
GLM	0.590	0.615	0.602	0.768	0.784	0.775	**+0.173**
KNN	0.620	0.559	0.587	0.657	0.856	0.743	**+0.156**
FLM	0.686	0.658	0.670	0.785	0.840	0.811	+0.141
DL(2-2-2)	0.672	0.737	0.686	0.797	0.780	0.787	+0.101

Average	0.663	0.650	**0.653**	0.765	0.822	**0.791**	**+0.135**

Table 6: Performance comparison of the two document representation techniques under a 10-fold cross-validation scenario using the final gold standard dataset (4,115 documents).

	Word embedding			**NER+Ontology**			**Gain**
	Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F-score*	*F-score*
SVM	0.824	0.838	0.831	0.825	0.890	0.856	+0.025
RF	0.781	0.903	**0.838**	0.915	0.794	0.850	+0.012
GLM	0.803	0.821	0.812	0.869	0.846	0.857	**+0.045**
KNN	0.793	0.833	0.812	0.828	0.838	0.833	+0.021
FLM	0.807	0.836	0.821	0.880	0.841	**0.860**	**+0.039**
DL(2-2-2)	0.780	0.886	0.828	0.878	0.821	0.848	+0.020
Average	0.798	0.852	**0.824**	0.865	0.838	**0.851**	**+0.027**

Table 7. Performance comparison of the different document representation techniques using the best optimization parameters for the different classifiers being evaluated under a 10-fold cross-validation scenario over the final gold standard dataset (4,115 documents).

	Word embedding			**NER+Ontology**			**Gain**	**Word embedding & NER+Ontology**		
	Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F-score*	*F-score*	*Precision*	*Recall*	*F-score*
SVM	0.808	0.877	0.841	0.898	0.830	0.863	+0.022	0.891	0.842	**0.866**
RF	0.796	0.909	**0.849**	0.914	0.815	0.862	+0.013	0.915	0.815	0.862
GLM	0.819	0.844	0.831	0.880	0.850	**0.864**	**+0.033**	0.874	0.851	0.862
KNN	0.770	0.910	0.834	0.860	0.836	0.848	+0.014	0.915	0.789	0.847
FLM	0.795	0.873	0.832	0.878	0.846	0.861	+0.029	0.881	0.846	0.863
DL(2-2-2)	0.788	0.881	0.832	0.884	0.843	0.862	**+0.031**	0.882	0.830	0.854
Average	0.796	0.882	**0.836**	0.886	0.837	**0.860**	**+0.024**	0.893	0.828	**0.859**