MeSH2Matrix: Machine learning-driven biomedical relation classification based on the MeSH keywords of PubMed scholarly publications

Houcemeddine Turki[1*], Bonaventure F. P. Dossou[2*], Chris Chinenye Emezue[3*], Mohamed Ali Hadj Taieb[1], Mohamed Ben Aouicha[1], Hanen Ben Hassen[4] and Afif Masmoudi[4]

[1]*Data Engineering and Semantics Research Unit, Faculty of Sciences of Sfax, University of Sfax, Sfax, Tunisia*
[2]*Jacobs University Bremen, Germany*
[3]*Technical University of Munich, Germany*
[4]*Laboratory of Probability and Statistics, Faculty of Sciences of Sfax, University of Sfax, Sfax, Tunisia*
[*]*Equal Contribution*

Abstract

Biomedical relation classification has been significantly improved by the application of advanced machine learning techniques on the raw texts of scholarly publications. Despite this improvement, the reliance on large chunks of raw text makes these algorithms suffer in generalization, precision and reliability. However, the use of the distinctive characteristics of bibliographic metadata can prove effective in achieving a better performance for this challenging task. In this research paper, we introduce an approach for biomedical relation classification using the qualifiers of co-occurring Medical Subject Headings (MeSH). First of all, we introduce *MeSH2Matrix*, our dataset consisting of 46,469 biomedical relations curated from PubMed publications using our approach. Using *MeSH2Matrix*, we build and train three machine learning models (*SVM*, *D-Model* and *C-Net*) to evaluate the efficiency of our approach for biomedical relation classification. Our best model achieves an accuracy of 70.78% for 195 classes and 83.09% for five superclasses. Our results will hopefully shed light on developing better algorithms for biomedical ontology construction based on the MeSH keywords of PubMed publications. For reproducibility purposes, *MeSH2Matrix* as well as all our source codes are made publicly accessible at https://github.com/SisonkeBiotik-Africa/MeSH2Matrix.

Keywords

Biomedical Relation Classification, MeSH Keywords, PubMed Records, MeSH qualifiers, Machine Learning

BIR 2022: 12th International Workshop on Bibliometric-enhanced Information Retrieval at ECIR 2022, April 10, 2022, hybrid.

turkiabdelwaheb@hotmail.fr (H. Turki); femipancrace.dossou@gmail.com (B. F. P. Dossou); chris.emezue@gmail.com (C. C. Emezue); mohamedali.hajtaieb@fss.usf.tn (M. A. H. Taieb); mohamed.benaouicha@fss.usf.tn (M. B. Aouicha); hanenbenhassen@yahoo.fr (H. B. Hassen); afif.masmoudi@fss.usf.tn (A. Masmoudi)
0000-0003-3492-2014 (H. Turki); 0000-0002-0519-1761 (B. F. P. Dossou); 0000-0002-3533-6829 (C. C. Emezue); 0000-0002-2786-8913 (M. A. H. Taieb); 0000-0002-2277-5814 (M. B. Aouicha); 0000-0002-5163-8528 (H. B. Hassen); 0000-0003-1665-5354 (A. Masmoudi)

1. Introduction

Biomedical ontologies are currently having a growing place in driving intelligent systems and information retrieval for clinical decision support and natural language processing [1, 2]. Ontologies are constituted of concepts related to each other using statements in the form of triples: *Subject* (Concept), *Predicate* (Relation Type), and *Object* (Concept) [3]. As a result they can be easily enriched, processed, validated, and reused by machines [4]. Nevertheless, biomedical ontologies are mainly created through human curation by experts (e.g., physicians and ontologists), consortiums (e.g., *Open Biomedical Ontologies Consortium*), and institutions (e.g. *NIH National Center for Biomedical Ontology*) [5], while machines are delegated particular tasks in ontology engineering: biomedical relation extraction and classification [6], biomedical ontology matching and integration [7], and biomedical ontology evaluation and validation [8]. For these tasks, free-form text from large collections of scholarly publications are usually processed, while the metadata, particularly bibliographic metadata, of the analyzed publications are not leveraged [6, 9].

Although these systems achieve high accuracy rates (F1-Score of biomedical relation classification with raw texts from 63.5 to 84.3), the consideration of the characteristics of scholarly publications could help refine their results to achieve better outputs [9]. In addition, the development of customized algorithms driven by the structure and the logic behind facets of the extracted publications – such as controlled keywords – could bring more trustworthy results using fewer computer resources [10, 11].

In this context, *Medical Subject Headings* (MeSH) keywords used to annotate PubMed scholarly publications can be a very useful resource for biomedical relation extraction and classification [10]. Based on MeSH taxonomy, it always assigns the same concept to various publications in PubMed using the same term [10]. Consequently, the use of MeSH keywords as input for biomedical ontology engineering can be more efficient than the use of user-generated bibliometric metadata and raw texts of scholarly publications [10]. In this research paper, we propose an approach for biomedical relation classification using the associations of the MeSH keywords in PubMed records. For this, we will use the biomedical relations between MeSH terms as revealed by Wikidata, an open and collaborative knowledge graph available at https://www.wikidata.org [12], to construct the training dataset named MeSH2Matrix for biomedical relation classification using our method.

We will begin by giving an overview of MeSH Keywords and how they have contributed so far to enhancing tasks in Biomedical Informatics (Section 2.1). Then, we will describe Wikidata as a biomedical knowledge resource and explain how it can be used to extract biomedical relations between the MeSH terms (Section 2.3). After that, we will explain the principles of our proposed approach for the biomedical relation classification based on the MeSH keywords of PubMed scholarly publications and explicate our method for the creation of our training dataset from PubMed and Wikidata (Section 3.1). Later, we will provide a description of the MeSH2Matrix dataset and assess its quality by comparing its main features to previous research findings (Section 3.2). Subsequently, we will describe our experiments for the development of the biomedical relation classification machine-learning algorithms to be trained on MeSH2Matrix and outline our results for the biomedical relation classification based on MeSH2Matrix (Section 4). Finally, we will conclude our experiments and provide future directions for our research

paper (Section 5).

2. Overview

2.1. MeSH keywords as a valuable input

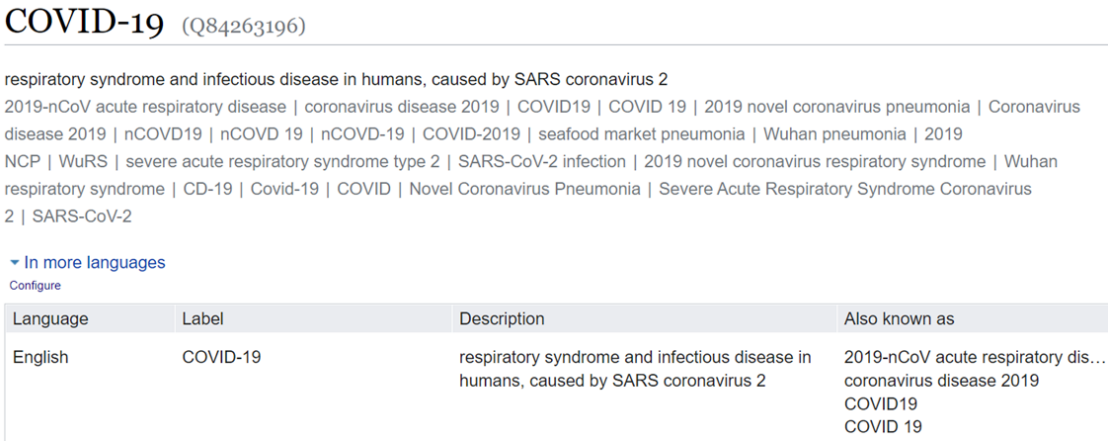
As a controlled vocabulary, MeSH supports sixteen types of biomedical concepts ranging from anatomical structures to symptoms, diseases, and drugs[1] [13]. This broad coverage of MeSH Terms makes them useful to represent the topics of all scholarly publications [13]. The value of these terms is further increased by the regular revision of MeSH to include new concepts such as COVID-19[2] and SARS-CoV-2[3] and to cover updates in the biomedical nomenclature [14]. That is why it has been used to annotate PubMed records for years by the human curators of the bibliographic database to enable consistent indexing of scholarly papers and intuitive data mining [13, 15]. Beyond its contribution to the enhanced topic granularity in PubMed, MeSH keywords have the ability to represent facets of a given topic through the use of predefined subheadings known as MeSH qualifiers providing more precision to the MeSH keywords of the PubMed records [9, 10]. Currently, there are 89 MeSH qualifiers representing all the characteristics and features of a biomedical entity as revealed at https://www.nlm.nih.gov/mesh/subhierarchy.html.

Table 1
Examples of relation types corresponding to the associations of two MeSH Keywords. MeSH Qualifiers contributing the relation type are indicated in bold

MeSH Keyword 1	MeSH Keyword 2	Relation Type
Sofosbuvir/**therapeutic use**	Hepatitis C/**drug therapy**	Medical Condition Treated
Asthma/**complications**	Dyspnea/**etiology**	Symptom
Retinopathy/**prevention & control**	Diabetes Mellitus/**complications**	Risk Factor

These interesting features of the MeSH keywords encouraged its usage beyond information seeking. MeSH keywords are gaining an increasing popularity in biomedical information retrieval as they allow better accuracy for relation extraction and classification from PubMed [15, 16]. This is due to the restriction of the considered publications to the ones that are most likely to include the required information [16] and the analysis of the co-occurrences of the MeSH Keywords [15]. The better output of clinical knowledge engineering driven by the MeSH terms of the PubMed publications has proven the value of MeSH Keywords, especially when assigned controlled qualifiers, to classify biomedical relations [9, 10]. Essentially, the qualifiers of two co-occurring MeSH keywords can inform us of the nature of the semantic relation between them as shown in Table 1. This is particularly motivated by the fact that biomedical publications usually have narrow research scope and do not consequently study multiple and unrelated facets of a given topic unless the publication type is an encyclopedic review [**?**]. The

[1] https://www.nlm.nih.gov/bsd/disted/meshtutorial/meshtreestructures/index.html
[2] https://www.ncbi.nlm.nih.gov/mesh/2052179
[3] https://www.ncbi.nlm.nih.gov/mesh/2052180



COVID-19 (Q84263196)

respiratory syndrome and infectious disease in humans, caused by SARS coronavirus 2

2019-nCoV acute respiratory disease | coronavirus disease 2019 | COVID19 | COVID 19 | 2019 novel coronavirus pneumonia | Coronavirus disease 2019 | nCOVD19 | nCOVD 19 | nCOVD-19 | COVID-2019 | seafood market pneumonia | Wuhan pneumonia | 2019 NCP | WuRS | severe acute respiratory syndrome type 2 | SARS-CoV-2 infection | 2019 novel coronavirus respiratory syndrome | Wuhan respiratory syndrome | CD-19 | Covid-19 | COVID | Novel Coronavirus Pneumonia | Severe Acute Respiratory Syndrome Coronavirus 2 | SARS-CoV-2

In more languages

Configure

Language	Label	Description	Also known as
English	COVID-19	respiratory syndrome and infectious disease in humans, caused by SARS coronavirus 2	2019-nCoV acute respiratory dis… coronavirus disease 2019 COVID19 COVID 19

Figure 1: English-language designations and language-independent identifier for *COVID-19* in Wikidata (Source: https://w.wiki/4gkc).

qualifiers of the MeSH keywords can be easily found as they are simply separated from their corresponding headings using a slash (/). The MeSH keywords of PubMed scholarly publications can be retrieved from the NCBI Entrez API using the Biopython Python Library [17, 18]. The structured format of MeSH keywords and their simple retrieval from the PubMed bibliographic database motivate their usage for the biomedical relation extraction and classification.

2.2. Wikidata as a biomedical semantic resource

Wikidata was created in October 2012 as a knowledge database to support structured data in Wikipedia such as interlanguage links and infoboxes [**?**]. However, it has grown over the past years to become one of the largest free and open knowledge graphs covering various range of fields, particularly biomedicine [12]. Its collaborative and crowdsourcing-based enrichment, regular updates according to recent advances in the major areas of interest, etc make it the most adequate knowledge base to support ever-changing scholarly evidences, mainly in the context of the COVID-19 pandemic [19]. Currently, Wikidata represents various types of medical entities as items, such as drugs, diseases, genes, proteins, organs, and symptoms [12]. These items are linked to their equivalent entities in external knowledge resources, mainly MeSH [19]. Every entity is assigned a language-independent identifier (so-called *Q-number*) as well as its main names (*labels*), glosses (*descriptions*), and alternative names (*aliases*) in a variety of natural languages, particularly English as shown in Fig. 1 [12, 19]. Furthermore, entities are related to other ones using semantic relations (*Subject – Predicate – Object*) where the relation type (*Predicate*) is also a Wikidata entity having its own identifier (*P-number*) and semantic description [12, 20].

There are two major kinds of relation types: taxonomic or non-taxonomic ones [12, 19]. Taxonomic relations are hierarchical ones that link an entity to its parent class or constituents (e.g., *instance of* [P31] – Fig. 2) [12]. Non-taxonomic relation types are non-hierarchical ones that are assigned to define specialized knowledge such as biomedical information (e.g., *signs*

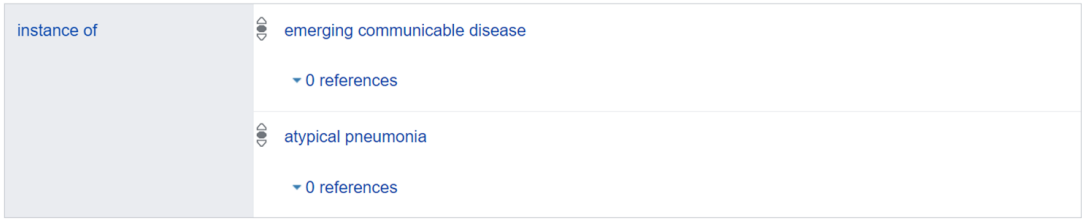


Figure 2: Examples of taxonomic relations for *COVID-19* in Wikidata (Source: https://w.wiki/4gkc).

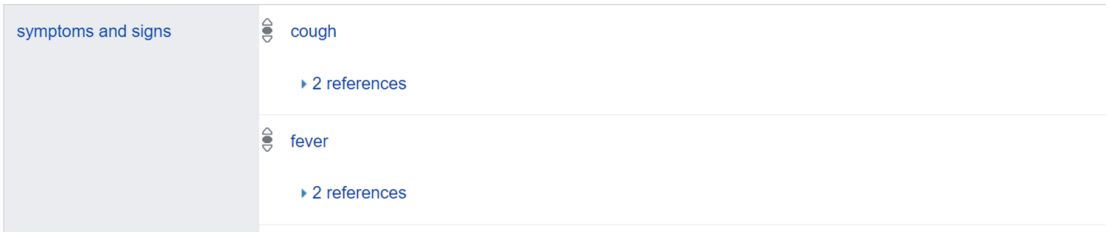


Figure 3: Examples of non-taxonomic relations for *COVID-19* in Wikidata (Source: https://w.wiki/4gkc).

or symptoms [P780] – Fig.3) [12]. A non-taxonomic relation can be symmetric: where the subject-object inversion would not affect the meaning of the statement. If this condition is not fulfilled, the relation in non-symmetric [20]. Such a classification of biomedical relation types can be easily inferred from the semantic information about Wikidata relation types, particularly the property constraints providing conditions for the definition of relational statements (Fig. 4) [20]. These conditions are not only useful for the validation of semantic information about the semantic relations but also to recognize the distinctive features of relation types. Moreover, Wikidata items are also matched to their equivalents in external resources using non-relational statements in the form of triples where the predicate reveals the aligned resource and the object is the identifier of the concept in the external database [20]. Particularly, *MeSH Descriptor ID* [P486] statements align between MeSH terms and Wikidata items as revealed by Fig. 5. As Wikidata knowledge graph is developed using the Resource Description Framework (RDF) format, it can be easily processed to get subsets of semantic information needed to drive knowledge-based applications using a variety of tools, particularly the MediaWiki API[4], the Wikidata SPARQL query service[5], and the Wikibase Integrator Python Library[6].

[4]https://www.wikidata.org/w/api.php
[5]https://query.wikidata.org
[6]https://github.com/LeMyst/WikibaseIntegrator

Figure 4: Examples of property constraints for *symptoms and signs* as a Wikidata relation type (Source: https://w.wiki/aeG).

Figure 5: MeSH descriptor ID for *COVID-19* in Wikidata as revealed by Wikidata (Source: https://w.wiki/4gkc).

3. MeSH2Matrix

3.1. Principles and dataset generation

We develop our approach upon the assumption that the qualifiers of two co-occurring MeSH terms can outline the type of semantic relation between them, as previously shown in Table 1 [9, 10]. Let t_1 and t_2 be two semantically related MeSH terms that are not assigned a relation type. Our method proposes to first search for the PubMed scholarly publications having both t_1 and t_2 as MeSH headings. Then for each of the found records, we extract the qualifiers q_1 and q_2 (e.g., *therapeutic use* for *Sofosbuvir/therapeutic use*) respectively corresponding to t_1 and t_2 (e.g., *Sofosbuvir* for *Sofosbuvir/therapeutic use*). This will enable the creation of (q_1, q_2) couples as shown in Fig. 6. When a term is assigned two or more qualifiers (e.g., t_2/*Z*/*U* for Paper 3 – Fig. 6), this means that a paper deals with a facet of a characteristic of the considered topic. In such a situation, we consider it as though the qualifiers were independently assigned to the MeSH term for the paper (e.g., t_2/*Z* and t_2/*U* for Paper 3 – Fig. 6). We restrict the number of considered publications to the 100 most relevant research papers according to PubMed *Best Match* search algorithm [21]. This will prevent matters related to the timeout limit of the NCBI PubMed API (*Error 429*). After the couples of MeSH qualifiers are retrieved, we draw a *matrix of correspondence* (M) – this is a square matrix of the qualifiers ($q_1, \ldots, q_{89}$)[7] where each element

[7]There are currently 89 pre-defined MeSH qualifiers as revealed at https://www.nlm.nih.gov/mesh/subhierarchy.html.

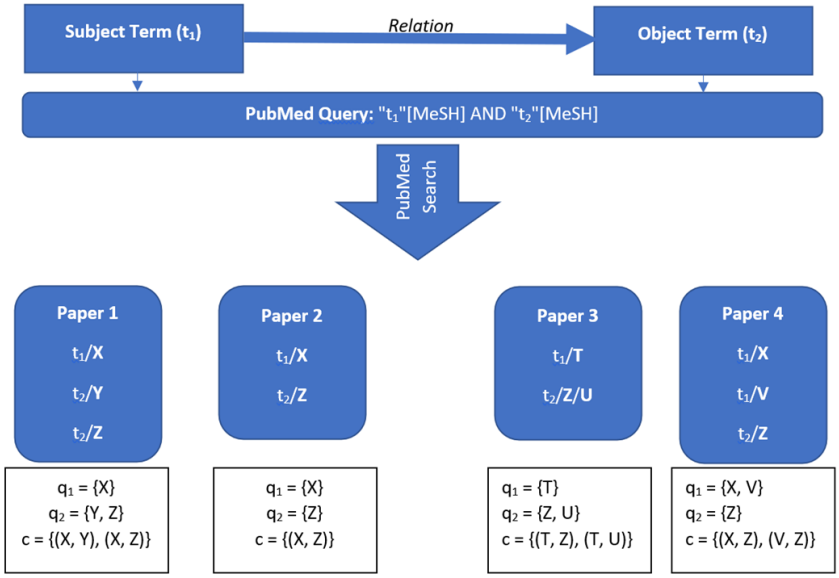


Figure 6: Process for the retrieval of the couples of MeSH qualifiers. t_1 is the subject MeSH term, t_2 is the object MeSH term, q_1 are the subject qualifiers, q_2 are the object qualifiers, and c is the set of the couples of the extracted MeSH qualifiers.

$m_{i,j}$ is the number of records featuring both t_1/q_i and t_2/q_j as MeSH keywords divided by the total number of records with the two MeSH terms t_1 and t_2 (Equation 1).

As a practical example, as of March 6, 2022, there are 32 PubMed records where *Hepatitis B* and *Sofosbuvir* are featured together as MeSH headings. From these 32 publications, there are 15 papers where *drug therapy* and *therapeutic use* are the respective qualifiers to *Hepatitis B* and *Sofosbuvir*. In this situation, the value that will be represented for the association between *drug therapy* and *therapeutic use* in the *Hepatitis B-Sofosbuvir* matrix is 15/32 = 0.469.

Floating-point representations (noted r) in this matrix will generally range between 0 and 1. This *matrix of correspondence* will be used as an input to find the nature of the semantic relation between t_1 and t_2.

$$r_{q_i,q_j} = \frac{N(t_1/q_i, t_2/q_j)}{N(t_1, t_2)}. \tag{1}$$

To construct our dataset, we first retrieve biomedical relations from Wikidata where the subject and the object are matched to their equivalent Medical Subject Headings (MeSH). This is accomplished with the SPARQL query featured in Fig. 7. The output of the query is saved as a tab-separated values (TSV) file to allow its automatic processing. After that, we use the Wikidata identifiers (also called P-numbers) of the extracted relation types as labels for the generated matrices. Then, we find the official names of the subjects and objects of the relations in MeSH based on their *MeSH Descriptor ID* [P486], which is then used to retrieve their associations in PubMed using the NCBI Entrez API. Finally, we extract the (q_1, q_2) couples and return the qualifiers' matrices for every subject-object association. Every matrix is assigned the relation type corresponding to the subject-object association as a label. For a better analysis of our

```
SELECT ?subject ?reltype ?object WITH {
  SELECT * WHERE {
    ?item wdt:P486 ?subject.
    }
  }
AS %item
WHERE {
  INCLUDE %item.
  ?item ?reltype ?item1.
  ?item1 wdt:P486 ?object.
}
LIMIT 81000
```

Figure 7: SPARQL query to extract 81,000 biomedical relations between MeSH terms in Wikidata (Live data: https://w.wiki/4h3g).

proposed approach, we extract the features of the considered Wikidata relation types and verify their names as well as if they are taxonomic, symmetric, or biomedical through the application of SPARQL queries on Wikidata using Wikibase Integrator coupled with human validation.

3.2. Results and Discussion

As of December 12, 2021, our SPARQL query (Fig. 7) has successfully retrieved 81,000 biomedical relations between MeSH Terms from Wikidata. This is a very significant amount of information as Wikidata only includes 99,208 semantic relations between MeSH concepts[8]. We have chosen 81,000 as the number of considered relations in order to simplify the performed computations for the analysis of our proposed approach. When analyzing the biomedical relations extracted for building our dataset, we found out that the supported relation types can be classified into five categories:

- *Non-Biomedical Non-Symmetric* (156 relation types, 17,758 relations),
- *Biomedical Non-Symmetric* (53 relation types, 27,429 relations),
- *Non-Biomedical Symmetric* (12 relation types, 9,000 relations),
- *Biomedical Symmetric* (3 relation types, 1,441 relations), and
- *Taxonomic* (3 relation types, 25,372 relations).

This goes in line with the coverage of various aspects of biomedical knowledge in Wikidata as a multidisciplinary knowledge graph [12, 19]. The extraction of the associations between the subject and object of every semantic relation in the MeSH keywords of PubMed publications has shown that most of the associations are likely to be found in a limited number of publications (Fig. 8A) and that commonly available MeSH associations in the PubMed records are rare (25,227 associations [42.7%] each available in 100 papers or more – Green dot in Fig. 8A). This is evident as scientific productivity follows Lotka's Law, an inverse power law that describes the uneven distribution of research outputs [22]. When seeing if the extraction of qualifiers describing the MeSH associations has been successful in generating matrices, we found that the probability of the creation of qualifiers' matrices tends to increase with the augmentation of the number of PubMed publications including the MeSH association before reaching a plateau near 1 at

[8]For a live update: https://w.wiki/4JN9

twenty publications (Fig. 8B). The existence of biomedical relations in Wikidata that cannot be found in PubMed records and that do not consequently return matrices of correspondence could be explained by the fact that Wikidata is subject to include irrelevant biomedical relations as it is collaboratively edited by human users without any restriction [20]. By contrast, it is important to reveal that the proportion of the generation of the matrices for the MeSH associations available in 100 publications or more is below the plateau with a rate of 73.3% (Red dots in Fig. 8B). To identify the reason behind such an unexpected behavior, we compute the quotient of the MeSH associations not generating matrices and available at least in 100 PubMed papers out of the overall number of the MeSH associations not having qualifiers' matrices for every class of Wikidata relation types. We found out that this rate is significantly higher for *taxonomic* (6,133 out of 13,411, 45.7%) and *non-biomedical non-symmetric* (1,712 out of 8,335, 20.5%) relations than for *biomedical non-symmetric* (1,137 out of 9,498, 12.0%), *non-biomedical symmetric* (189 out of 2,647, 7.1%), and *biomedical symmetric* (7 out of 640, 1.1%). This proves the ability of the MeSH qualifiers to better represent biomedical or symmetric relations than generic and non-symmetric ones.

The obtained dataset of qualifiers' matrices represented 46,469 relations covering the five classes of semantic relation types (195 supported relation types, 54.2% of the matrices based on 100 publications or more): 17,931 biomedical non-symmetric, 11,961 taxonomic, 9,423 non-biomedical non-symmetric, 6,353 non-biomedical symmetric, and 801 biomedical symmetric relations. Unsurprisingly, the most represented relation types in the dataset have been dominated by taxonomic (e.g., *subclass* of [P279] and *instance of [P31]*) or biomedical non-symmetric relation types (e.g., *drug and therapy used for treatment* [P2176]). This makes our dataset more inclusive than other available corpora for biomedical relation classification only covering a few relation types, particularly *drug interactions*, *drug adverse effects*, and *drug-disease relations* [6].

4. Biomedical relation classification using MeSH2Matrix

Machine learning-based approaches handle biomedical relation classification as a supervised learning classification task, where labelled data is used to train models. In this paper, we provide benchmark results on our dataset, using three machine learning models:

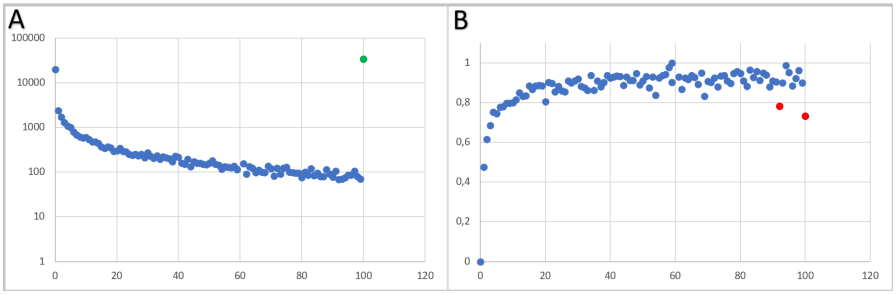


Figure 8: Variables in function of the number of PubMed publications about a given association: Number of semantic relations (A, Log-Scale), Rate of semantic relations returning matrices (B)

SVM: Support vector machines (SVMs) [23] are best suited for samples with many features because of their ability to learn is independent of the features space [24]. They have been used exensively in biomedical classification tasks [25, 26, 27, 28] due to their ability to generalize well with data consisting of sparse high-dimensional features. For our baseline, we trained a linear support vector machine. For this, we transformed each 89×89 matrix into a single 7921 feature vector.

D-Model: Neural networks (NNs) have produced state-of-the-art results in the area of relation classification [28, 29, 30, 31]. The major advantage of neural network based approaches lies in thier ability to directly learn the latent feature representation from the labeled training data without requiring experts to carefully craft them [6]. For our experiments with neural networks, we designed *D-Model*, a simple multi-layer perceptron with an input layer of output feature size of $3,960$, a hidden layer of $1,980$ and an output layer with an output feature size corresponding to the number of classes [rationale for the choice of the size of neurons: 1). we tested different sizes and this gave the best result, and 2). we followed [32, 33, 34] in keeping the hidden layer size between input layer size and output layer size]. ReLU activation function [35] was used between the input and hidden layers to introduce non-linearity. The output layer is connected to a softmax activation function which converts the model's output into a probability over the classes. Although NNs have shown great promise for relation classification, they are highly susceptible to overfitting [36] and require lots of hyperparameter tuning. Therefore, we experimented with regularization techniques (early stopping and dropout) the hyperparameters (learning rate, batch size, etc) in order to produce the best performing *D-Model*.

C-Net: Convolutional neural networks (CNNs) are a type of neural networks that can successfully capture the spatial and temporal dependencies in an image through the application of convolution operation and relevant filters. Their potential was first witnessed in computer vision around 2012 [37], and since then have been used extensively even in biomedical relation classification [38, 39, 30]. CNNs perform well on an image dataset better due to the reduction in the number of parameters involved and reusability of their weights - they are therefore best suited for image-type data. Furthermore, with CNNs we can work directly with the 2-dimensional matrix (compared to transforming it for *SVM* and *D-Model*). To explore the impact of CNNs on MeSH2Matrix, we decided to interprete our feature matrix as spatially correlated features and designed *C-Net*, a simple CNN-based architecture made up of four convolution layers (each layer consisting of a 2-dimensional convolution, batch normalization [40], a ReLU activation function [35] and max-pooling) and two fully connected layers. After passing through the fully connected layers, the final layer uses the softmax activation function which is used to get probabilities of the input matrix being in a particular class. CNN-based models, while being very promising, require practical knowledge to configure the model architecture with regard to the performance [41], and to set the hyperparameters for the best optimization [42]. Similarly, we conducted hyperparameter tuning and optimization in order to explore the reasonable ranges for the sensitive hyperparameters of the classification model.

4.1. Experiments

We performed two rounds of classification: one with all relation types (195), and another with 5 categories obtained after grouping the initial 195 relation types (see section 3.2 for more details on grouping). We split our dataset into training ($33,457$ samples), validation ($13,012$ samples) - for early stopping, regularization and hyperparameter tuning - and testing ($9,294$ samples) - for the final evaluation of the model. For *SVM* training, we merged the training and validation set, making a total of $46,469$ samples for training. For the training of *D-Model* and *C-Net*, we used the Adam optimizer [43]. The code for all our deep learning experiments was written using the PyTorch deep learning framework [44], while for *SVM* we implemented the training using the LinearSVC package.

4.2. Results

Table 2
Accuracy [and F1-Score] (in percentage) of the models used in our experiments. In both classes, *D-Model* performs best, followed by *C-Net* and lastly *SVM*. Also, all models performed better on 5 classes compared to 195 classes.

Models	195 classes	5 classes
SVM	66.43 [61.27]	78.74 [78.63]
D-Model	**70.78** [**66.90**]	**83.09** [**82.94**]
C-Net	70.49 [66.18]	82.78 [82.61]

Table 2 shows the results of the three benchmark models on the 195-classification and 5-classification tasks. The metric being used are *accuracy* and *multi-class F1-score* (which is a metric that combines the precision and recall of the model). It is clear that the three proposed models achieved acceptable accuracy measures that go in line with the recent advances in biomedical relation classification (*F1-Score* between 0.65 and 0.85) [6].

We see a notable improvement in the probabilistic methods (*D-Model* and *C-Net*) over *SVM*. *D-Model* performs the best, although outperforming *C-Net* by a small margin. Another observation is that for all the models, their performance on the 5-class takes was much better than the 195-class ones. This could be because the consideration of five generalized superclasses actually reduces the complexity of the task, making it easier for the model to learn [45]. For example, class generalization allowed us to be get rid of the closely related taxonomic relation types (i.e., *instance of* [P31], *subclass of* [P279], and *part of* [P361]) and eliminate the effect of the confusion between these three relation types on the accuracy of the models. Another possible reason could be that grouping increased the distribution of some minority classes (classes with a very few samples). On the one hand, due to the enrichment of Wikidata thanks to human efforts, important Wikidata statements can be mistakenly defined for minor relation types [12]. The effect of such deficient relations will become insignificant when the generalization occurs. On the other hand, as of December 12, 2021, 4,522 (4.4%) out of the 99,208 Wikidata relations between MeSH Terms are having the same subject and object as another supported

semantic relations between MeSH Concepts[9]. Statements having the same subjects and objects but different relation types are likely to be merged together due to the class generalization, allowing to reduce the confusion between slightly overlapping relation types.

5. Conclusion

In this research paper, we proposed a novel approach for the classification of biomedical relations based on the association between the qualifiers of two semantically related MeSH keywords of PubMed scholarly publications. We generated MeSH2Matrix as a training dataset (covering 195 relation types involved in five superclasses) to enable the MeSH-based biomedical relation classification and we trained three benchmarking machine-learning models (*SVM*, *D-Model* and *C-Net*) to evaluate the efficiency of our approach to classify various types of biomedical relations. We found an interesting efficiency of our approach in biomedical relation classification (*F1-Score* > 0.66) proving the promising value of using Bibliometric-Enhanced Information Retrieval towards the improvement of biomedical relation classification. For reproducibility purposes, our source code and dataset are currently available at https://github.com/SisonkeBiotik-Africa/MeSH2Matrix. As a future direction of this work, we propose to expand our approach into a method for converting subsets of the MeSH taxonomy into biomedical ontologies. Furthermore, we propose to further analyze the effect of the generalization of relation types to their respective superclasses on the accuracy rates of our models to study the behavior of our approach.

Acknowledgments

This work has been supported by the Tunisian Ministry of Higher Education and Scientific Research (MoHESR) within the framework of the Federated Research Project PRFCOV19-D1-P1 and by Craig Newmark Philanthropies, Facebook, and Microsoft through the WikiCred Grant Initiative. We thank the Sisonkebiotik Community, particularly Chris Fourie (University of the Witwatersrand, South Africa), for contributing to the development of the research collaboration that resulted in this research paper. We thank Dr. Ahmed Ben Abdelaziz (University of Sousse, Tunisia) for introducing the Medical Subject Headings to the first author of this research work in March 2016. We also thank Mr. Colin Leong (University of Dayton, United States of America) for his useful comments and discussion.

[9]Live data: https://w.wiki/4itd

References

[1] J. P. Bona, F. W. Prior, M. N. Zozus, M. Brochhausen, Enhancing clinical data and clinical research data with biomedical ontologies - insights from the knowledge representation perspective, Yearbook of Medical Informatics 28 (2019) 140–151. doi:10.1055/s-0039-1677912.

[2] M. Kulmanov, F. Z. Smaili, X. Gao, R. Hoehndorf, Semantic similarity and machine learning with ontologies, Briefings in Bioinformatics 22 (2021) bbaa199. doi:10.1093/bib/bbaa199.

[3] A. Bandrowski, R. Brinkman, M. Brochhausen, M. H. Brush, B. Bug, M. C. Chibucos, K. Clancy, M. Courtot, D. Derom, M. Dumontier, et al., The ontology for biomedical investigations, PLOS ONE 11 (2016) e0154556. doi:10.1371/journal.pone.0154556.

[4] R. Hoehndorf, M. Dumontier, G. V. Gkoutos, Evaluation of research in biomedical ontologies, Briefings in Bioinformatics 14 (2012) 696–712. doi:10.1093/bib/bbs053.

[5] B. M. Konopka, Biomedical ontologies—a review, Biocybernetics and Biomedical Engineering 35 (2015) 75–86. doi:10.1016/j.bbe.2014.06.002.

[6] Y. Zhang, H. Lin, Z. Yang, J. Wang, Y. Sun, B. Xu, Z. Zhao, Neural network-based approaches for biomedical relation classification: A review, Journal of Biomedical Informatics 99 (2019) 103294. doi:10.1016/j.jbi.2019.103294.

[7] D. Oliveira, C. Pesquita, Improving the interoperability of biomedical ontologies with compound alignments, Journal of Biomedical Semantics 9 (2018). doi:10.1186/s13326-017-0171-8.

[8] M. Amith, Z. He, J. Bian, J. A. Lossio-Ventura, C. Tao, Assessing the practice of biomedical ontology evaluation: Gaps and opportunities, Journal of Biomedical Informatics 80 (2018) 1–13. doi:10.1016/j.jbi.2018.02.010.

[9] H. Turki, M. A. Hadj Taieb, M. Ben Aouicha, G. Fraumann, C. Hauschke, L. Heller, Enhancing knowledge graph extraction and validation from scholarly publications using bibliographic metadata, Frontiers in Research Metrics and Analytics 6 (2021) 694307. doi:10.3389/frma.2021.694307.

[10] H. Turki, M. A. Hadj Taieb, M. Ben Aouicha, Mesh qualifiers, publication types and relation occurrence frequency are also useful for a better sentence-level extraction of biomedical relations, Journal of Biomedical Informatics 83 (2018) 217–218. doi:10.1016/j.jbi.2018.05.011.

[11] H. Turki, M. A. Hadj Taieb, M. Ben Aouicha, Enhancing filter-based parenthetic abbreviation extraction methods, Journal of the American Medical Informatics Association 28 (2021) 668–669. doi:10.1093/jamia/ocaa314.

[12] H. Turki, T. Shafee, M. A. Hadj Taieb, M. Ben Aouicha, D. Vrandečić, D. Das, H. Hamdi, Wikidata: A large-scale collaborative ontological medical database, Journal of Biomedical Informatics 99 (2019) 103292. doi:10.1016/j.jbi.2019.103292.

[13] N. Baumann, How to use the medical subject headings (mesh), International Journal of Clinical Practice 70 (2016) 171–174. doi:10.1111/ijcp.12767.

[14] L. Leydesdorff, J. A. Comins, A. A. Sorensen, L. Bornmann, I. Hellsten, Cited references and medical subject headings (mesh) as two different knowledge representations: Clustering and mappings at the paper level, Scientometrics 109 (2016) 2077–2091. doi:10.1007/

s11192-016-2119-7.
[15] Y. Lu, B. Figler, H. Huang, Y.-C. Tu, J. Wang, F. Cheng, Characterization of the mechanism of drug-drug interactions from pubmed using mesh terms, PLOS ONE 12 (2017) e0173548. doi:10.1371/journal.pone.0173548.
[16] T. Tran, R. Kavuluru, Distant supervision for treatment relation extraction by leveraging mesh subheadings, Artificial Intelligence in Medicine 98 (2019) 18–26. doi:10.1016/j.artmed.2019.06.002.
[17] B. Chapman, J. Chang, Biopython: Python tools for computational biology, ACM SIGBIO Newsletter 20 (2000) 15–19. doi:10.1145/360262.360268.
[18] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, et al., Biopython: Freely available python tools for computational molecular biology and bioinformatics, Bioinformatics 25 (2009) 1422–1423. doi:10.1093/bioinformatics/btp163.
[19] H. Turki, M. A. Hadj Taieb, T. Shafee, T. Lubiana, D. Jemielniak, M. Ben Aouicha, J. E. Labra Gayo, E. A. Youngstrom, M. Banat, D. Das, et al., Representing covid-19 information in collaborative knowledge graphs: The case of wikidata, Semantic Web 13 (2022) 233–264. doi:10.3233/sw-210444.
[20] H. Turki, D. Jemielniak, M. A. Hadj Taieb, J. E. Labra Gayo, M. Ben Aouicha, M. Banat, T. Shafee, E. Prud'Hommeaux, T. Lubiana, D. Das, D. Mietchen, Using logical constraints to validate statistical information about covid-19 in collaborative knowledge graphs: the case of wikidata, Zenodo (2021). doi:10.5281/zenodo.4008358.
[21] N. Fiorini, K. Canese, G. Starchenko, E. Kireev, W. Kim, V. Miller, M. Osipov, M. Kholodov, R. Ismagilov, S. Mohan, et al., Best match: New relevance search for pubmed, PLOS Biology 16 (2018) e2005343. doi:10.1371/journal.pbio.2005343.
[22] L. Egghe, R. Rousseau, Theory and practice of the shifted lotka function, Scientometrics 91 (2012) 295–301. doi:10.1007/s11192-011-0539-y.
[23] N. Cristianini, E. Ricci, Support Vector Machines, Springer US, Boston, MA, 2008, pp. 928–932. doi:10.1007/978-0-387-30162-4_415.
[24] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, in: C. Nédellec, C. Rouveirol (Eds.), Machine Learning: ECML-98, Springer Berlin Heidelberg, Berlin, Heidelberg, 1998, pp. 137–142.
[25] A. Ben Abacha, P. Zweigenbaum, A hybrid approach for the extraction of semantic relations from MEDLINE abstracts, in: Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg, 2011, pp. 139–150. doi:10.1007/978-3-642-19437-5_11.
[26] T. Mavropoulos, D. Liparas, S. Symeonidis, S. Vrochidis, I. Kompatsiaris, A hybrid approach for biomedical relation extraction using finite state automata and random forest-weighted fusion, in: A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing, Springer International Publishing, Cham, 2018, pp. 450–462.
[27] A. W. Muzaffar, F. Azam, U. Qamar, A relation extraction framework for biomedical text using hybrid feature set, Computational and Mathematical Methods in Medicine 2015 (2015) 1–12. doi:10.1155/2015/910423.
[28] D. Zeng, K. Liu, S. Lai, G. Zhou, J. Zhao, Relation classification via convolutional deep neural network, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin City University and Association for

Computational Linguistics, Dublin, Ireland, 2014, pp. 2335–2344. URL: https://aclanthology.org/C14-1220.

[29] C. dos Santos, B. Xiang, B. Zhou, Classifying relations by ranking with convolutional neural networks, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Beijing, China, 2015, pp. 626–634. doi:10.3115/v1/P15-1061.

[30] Y. Peng, Z. Lu, Deep learning for extracting protein-protein interactions from biomedical literature, in: BioNLP 2017, Association for Computational Linguistics, Vancouver, Canada,, 2017, pp. 29–38. doi:10.18653/v1/W17-2304.

[31] A. Rios, R. Kavuluru, Z. Lu, Generalizing biomedical relation classification with neural adversarial domain adaptation, Bioinformatics 34 (2018) 2973–2981. doi:10.1093/bioinformatics/bty190.

[32] J. Heaton, Introduction to Neural Networks for Java, 2nd Edition, 2nd ed., Heaton Research, Inc., 2008.

[33] D. Stathakis, How many hidden layers and nodes?, International Journal of Remote Sensing 30 (2009) 2133–2147. URL: https://doi.org/10.1080/01431160802549278. doi:10.1080/01431160802549278. arXiv:https://doi.org/10.1080/01431160802549278.

[34] H. B. Demuth, M. H. Beale, O. De Jess, M. T. Hagan, Neural Network Design, 2nd ed., Martin Hagan, Stillwater, OK, USA, 2014.

[35] A. F. Agarap, Deep learning using rectified linear units (relu), 2018. arXiv:1803.08375.

[36] L. Mou, Z. Meng, R. Yan, G. Li, Y. Xu, L. Zhang, Z. Jin, How transferable are neural networks in NLP applications?, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 479–489. doi:10.18653/v1/D16-1046.

[37] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 25, Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[38] S. Liu, B. Tang, Q. Chen, X. Wang, Drug-drug interaction extraction via convolutional neural networks, Computational and Mathematical Methods in Medicine 2016 (2016) 6918381. doi:10.1155/2016/6918381.

[39] C. Quan, L. Hua, X. Sun, W. Bai, Multichannel convolutional neural network for biological relation extraction, BioMed Research International 2016 (2016) 1850404. doi:10.1155/2016/1850404.

[40] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: F. Bach, D. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, volume 37 of *Proceedings of Machine Learning Research*, PMLR, Lille, France, 2015, pp. 448–456. URL: https://proceedings.mlr.press/v37/ioffe15.html.

[41] Y. Zhang, B. Wallace, A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification, in: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Asian Federation of Natural Language Processing, Taipei, Taiwan, 2017, pp. 253–263. URL:

https://aclanthology.org/I17-1026.

[42] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (2012) 281–305.

[43] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL: http://arxiv.org/abs/1412.6980.

[44] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

[45] H. Turki, M. A. Hadj Taieb, M. Ben Aouicha, How knowledge-driven class generalization affects classical machine learning algorithms for mono-label supervised classification, in: Proceedings of the 21st International Conference on Intelligent Systems Design and Applications, Springer, Cham, Online, 2021. doi:10.1007/978-3-030-96308-8_59.