Journal Pre-proof

NewsMeSH: a new classifier designed to annotate health news with MeSH headings

Joao Pita Costa, Luis Rei, Luka Stopar, Flavio Fuart, Marko Grobelnik, Dunja Mladenić, Inna Novalija, Anthony Staines, Jarmo Pääkkönen, Jenni Konttila, Joseba Bidaurrazaga, Oihana Belar, Christine Henderson, Gorka Epelde, Mónica Arrúe Gabaráin, Paul Carlin, Jonathan Wallace

PII: S0933-3657(21)00046-4

DOI: https://doi.org/10.1016/j.artmed.2021.102053

Reference: ARTMED 102053

To appear in: *Artificial Intelligence In Medicine*

Received Date: 6 April 2020

Revised Date: 21 January 2021

Accepted Date: 11 March 2021

Please cite this article as: Costa JP, Rei L, Stopar L, Fuart F, Grobelnik M, Mladenić D, Novalija I, Staines A, Pääkkönen J, Konttila J, Bidaurrazaga J, Belar O, Henderson C, Epelde G, Gabaráin MA, Carlin P, Wallace J, NewsMeSH: a new classifier designed to annotate health news with MeSH headings, *Artificial Intelligence In Medicine* (2021), doi: https://doi.org/10.1016/j.artmed.2021.102053

NewsMeSH: a new classifier designed to annotate health news with MeSH headings

Joao Pita Costa [(1) (2)], Luis Rei [(1)], Luka Stopar [(1) (2)], Flavio Fuart [(1) (2)], Marko Grobelnik [(1) (2)], Dunja Mladenić [(1) (2)], Inna Novalija [(1)], Anthony Staines [(3)], Jarmo Pääkkönen [(4)], Jenni Konttila [(4)], Joseba Bidaurrazaga [(5)], Oihana Belar [(5)], Christine Henderson [(6)], Gorka Epelde [(7)(8)], Mónica Arrúe Gabaráin[(7)(8)], Paul Carlin [(9)], Jonathan Wallace [(10)]

Highlights

- We present an automated text classifier based on the MEDLINE/MeSH thesaurus, with focus on the classification of health records and health news
- The proposed text classifier shows results both in the evaluation of scientific text (evaluated against the MEDLINE annotated articles) and on health-related news (evaluated against news articles annotated by health experts).
- This classifier was developed tailor-fit to the public health and health research domain experts, in the light of their specific challenges and needs.
- We have applied the proposed methodology on three specific health domains: the Coronavirus, Mental Health and Diabetes, considering the pertinence of the first, and the known relations with the other two health topics.

ABSTRACT

Motivation: In the age of big data, the amount of scientific information available online dwarfs the ability of current tools to support researchers in locating and securing access to the necessary materials. Well-structured open data and the smart systems that make the appropriate use of it are invaluable and can help health researchers and professionals to find the appropriate information by, e.g., configuring the monitoring of information or refining a specific query on a disease.

Methods: We present an automated text classifier approach based on the MEDLINE/MeSH thesaurus, trained on the manual annotation of more than 26 million expert-annotated scientific abstracts. The classifier was developed tailor-fit to the public health and health research domain experts, in the light of their specific challenges and needs. We have applied the proposed methodology on three specific health domains: the Coronavirus, Mental Health and Diabetes, considering the pertinence of the first, and the known relations with the other two health topics.

Results: A classifier is trained on the MEDLINE dataset that can automatically annotate text, such as scientific articles, news articles or medical reports with relevant concepts from the MeSH thesaurus.

Conclusions: The proposed text classifier shows promising results in the evaluation of health-related news. The application of the developed classifier enables the exploration of news and extraction of health-related insights, based on the MeSH thesaurus, through a similar workflow as in the usage of PubMed, with which most health researchers are familiar.

Keywords: Big Data; Semantic Technologies; Public Health; Healthcare; Text Mining; MeSH Headings; MEDLINE; PubMed; COVID-19; Diabetes; Mental Health.

1. INTRODUCTION

The day-to-day growth of knowledge to support public health and healthcare available online has reached a volume that is very hard to assimilate when researching specific health-related topics. Evidence of this abundance of information is the open scientific biomedical knowledge base – MEDLINE – and its comprehensive controlled vocabulary – the Medical Subject Headings (MeSH) thesaurus – which facilitates the correct refinement of a PubMed search based on the article metadata [30]. Aiming to address this need, we have: (i) developed a novel text classifier trained on the existing hand annotations of MeSH heading classes given to biomedical research papers in MEDLINE; (ii) proceeded with an extensive evaluation, both on new scientific articles, and on news articles over trending health topics; and (iii) provided a useful and easy-to-use web interface to access the MeSH classifier, as well as an API to allow its integration in existing systems.

The MeSH classifier approach discussed in this paper provides automated annotations of any text, and is thus useful to identify insight in health-related text as, e.g., reports, articles or news. To validate the proposed approach and the constructed classifier, we focus on news annotation. To that end, the integration of the MeSH classifier with a news engine allows the usage of MeSH classes on queries, and for visualisations to explore the health subtopics of interest (see Figure 1 for illustration). Moreover, this integration enables health professionals to use the MeSH controlled vocabulary with which many are familiar with, to enrich and extend their workflow when exploring and monitoring worldwide news.

Affiliations: (1) Jožef Stefan Institute, Slovenia, (2) Quintelligence, Slovenia, (3) Dublin City University, Ireland, (4) University of Oulu, Finland (5) BIOEF, Spain, (6) Northern Ireland Department of Health, UK, (7) Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Spain, (8) Biodonostia, Spain (9) Open University (10) Ulster University, UK

Figure 1 – A potential impact of the MeSH classifier applied to the classification of news: the percentage of news published in 2018/2019 that are annotated with the MeSH class "Public Health".

1.1. Motivation

With the accelerating use of big data, and with the analysis and visualisation of this information being used to positively affect the daily lives of people worldwide, health professionals need efficient and effective technologies to derive meaning and knowledge from information outputs, when planning and delivering care. The growth of online knowledge requires that the information sources utilised are complete, of high-quality and accessible. A particular example of this is the COVID-19 outbreak [39] that motivated worldwide joint initiatives to help monitor the disease (as [2] and [34]), and understand it better [20], including the crowdsourcing initiative to build new machine learning methods based on biomedical knowledge [13]. In the context of these global initiatives, the proposed text classifier was designed to automatically annotate text. Its development was motivated by the potential for the classification of health reports and news articles on Coronavirus, Mental Health and Diabetes, taking into consideration the pertinence of the further knowledge on the disease and the virus itself, but also the known relations with diabetes [9] and the impact of the social distancing it is generating on mental health [40].

A particular example of a well-established, useful and meaningful tool in the daily life of health professionals is the PubMed search engine, which allows access to state-of-the-art medical research. This tool is frequently used to gain an overview of a certain topic using several filters, tags and advanced search options. PubMed has been freely available since 1997, providing access to references and abstracts on life sciences and biomedical topics. MEDLINE is the underlying open database [16] served by the controlled vocabulary of the MeSH Headings, both of them maintained by the North American National Library of Medicine (NLM). The MeSH vocabulary is often used by health professionals to refine the search results provided by PubMed. This is done via the PubMed search engine directly, or via research assistant tools that integrate the access to this vocabulary such as Zotero.

The gain of automated knowledge discovery from MEDLINE/MeSH is transformative in medical research and can influence the progress of biomedical research [37]. In the context of the meaningful integration and usage of data, the EU H2020 project MIDAS (Meaningful Integration of Data, Analytics and Services) [29] is developing a big data platform that facilitates the utilisation of healthcare data beyond the existing isolated systems, making that data available for enrichment with open data. This data fusion approach thus enables evidence-based health policy decision making, and potentially may lead to significant improvements in healthcare and quality of life for all citizens [4].

The proposed classifier can also be easily integrated into a news search engine. There are several examples of such systems, and a range of news sources that can be annotated by the classifier to leverage its potential. The worldwide health monitoring potential of this tool was discussed in [27] in the context of Public Health decision-making support, though its application can extend to any domain where the automated annotation using terms from a health-related vocabulary such as that of the MeSH thesaurus could be useful.

1.2. Related work

There have been several well-accepted initiatives to use MeSH for the classification of text, often focusing on specific scientific problems [33]. More general approaches include the Medical Text Indexer (MTI) [1], that provides MeSH indexing recommendations to support the human indexers of the NLM using k-nearest neighbors (KNN), pattern matching and indexing rules, reaching a F measure of 56.37%; and the MeSH Now [19], including a learning-to-rank framework achieving a F measure of 61%. The latter two are very different in characteristics from the MeSH classifier proposed in this paper: the MTI and the MeSH Now are tools to address the needs of the PubMed user. The NLM also made available other useful tools like, e.g., the MeSH on Demand [32], that suggests MeSH vocabulary explicitly mentioned in the input text; and the Semantic MEDLINE [14], that aims for the semantic knowledge representation of MEDLINE itself. The Semantic MEDLINE initiative is different in the sense that it has focus on research on the semantic knowledge specific problems related to the summarization of MEDLINE citations. The MeSH Now technology is also aiming for the automatic MeSH indexing, with a tailor-fit methodology to the MEDLINE articles. The MeSH classifier we are proposing is also trained over the knowledge base provided by MEDLINE, integrating the concerns from the public health community. Though, it is a hierarchical labelling method,

designed to perform efficiently in text that is not tied to the formal jargon of the domain, allowing for the classification of, e.g., reports or news articles. Other alternatives include the MeSHLabeler [17], that combines predictions from MeSH classifiers, KNN and pattern matching, MTI and correlations between MeSH terms to achieve a F measure of 62.48%; and the DeepMeSH [24] that incorporates deep semantic information and the 'learning to rank' framework of MeSHLabeler to reach a better F measure of 63.23%. FullMeSH [7] is a recent alternative that is leveraging the increased availability of full text scientific articles in MEDLINE, combining semantic frameworks to get deeper insight from the paper's content in each section, achieving a F measure of 66.76%, higher than DeepMeSH and MeSHLabeler. Other MeSH-based classifiers to take into account are the end-to-end model AttentionMeSH [12], using deep learning and attention mechanism to index MeSH terms to biomedical text, to achieve a F measure of 68.4% with a high level of interpretability; the MeSHProbeNet [41], deep learning and self-attentive MeSH probes to index with achieving an F measure higher than DeepMesh and AttentionMeSH, reaching 67.89%; and the BERTMeSH [42], a pre-trained deep contextual representation, BERT (Bidirectional Encoder Representations from Transformers), capturing deep semantics of full text, reaching a F measure of 69.2%. The related work also includes other approaches: the BioNLP (biomedical natural language processing) [3], that is a biomedical text mining combining controlled vocabularies and ontologies available through the National Library of Medicine's Unified Medical Language System (UMLS) [5] and MeSH with the conventional one-vs.-rest (OVR) classification setup; the semantic biomedical question answering system SemBioNLQA for retrieving information based on natural language questions [31]; the usage of topic models to enrich the meta information provided by MeSH [23]; and the foundational work on the medical indexing expert system, MedIndEx, using knowledge-base frames to guide indexers in completing indexing frames [11].

2. BUILDING THE CLASSIFIER

2.1 Data and metadata description

The 2019 version of the MEDLINE dataset used to build the automated classifier proposed in this paper includes citations from more than 5,200 journals worldwide in approximately 40 languages (about 60 languages in older journals). It stores structured information on more than 27 million records dating from 1946 to the present. In most cases, the title and abstract are available but not the main body of the work. About 500 000 new records are added each year. 17.2 millions of these records are listed with their abstracts, and 16.9 million articles have links to full-text, of which 5.9 million articles have full-text available for free online use (see Table 1 for more information). In particular, it includes more than 43 thousand articles with the key-words string "public health".

Items	Public Health	All Domains
Number of abstracts	110023	27361292
Number of full-text articles	43844	17538890
Number of languages	42	58
Number of MeSH heading descriptors	10756	29256
Maximum depth of the MeSH tree	6	13
MeSH tree roots (major categories)	3	16

Table 1 – Dataset description based on the statistics for the open dataset MEDLINE and the MeSH headings

The comprehensive controlled vocabulary associated with the MEDLINE dataset – MeSH – delivers a functional system of indexing both journal articles and books in life sciences. It has proven very useful in the search of specific topics in medical research and is commonly used by medical researchers conducting initial literature reviews before engaging in particular research tasks [28]. Trained NLM librarians annotate the articles in MEDLINE with MeSH descriptors. These descriptors permit the user to explore a certain biomedical related topic, which relies on curated information made available by the NLM. MeSH is composed of 16 major categories (covering anatomy, diseases, drugs, etc.) that further subdivide from the most general to the most specific, with as many as 13 hierarchical depth levels.

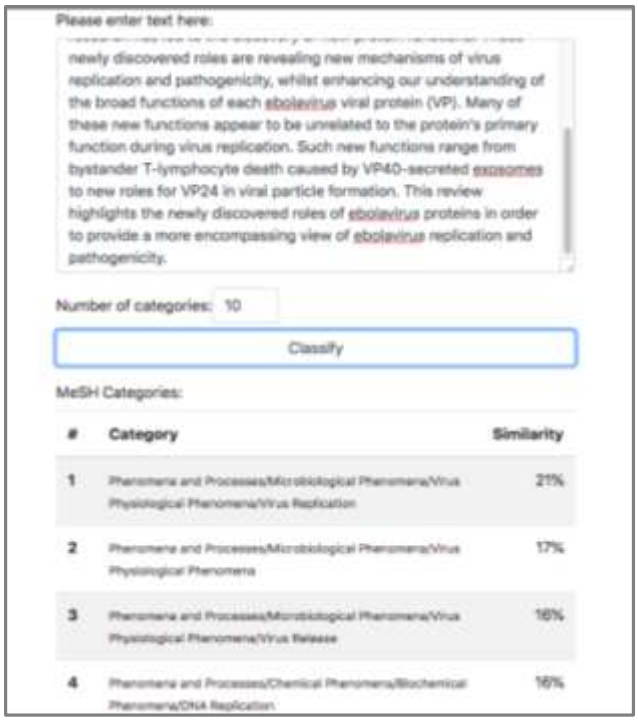


Figure 2 - An example of the MeSH classifier output for the automated MeSH annotation of a scientific article abstract extracted from PubMed (in the body of text above), the MeSH class rank (on the left), their MeSH tree path in the MeSH ontology-like structure (in the centre), and a measure of the similarity of each class to the text (on the right).

This rich data structure in the MEDLINE open set is manually annotated (although assisted by semi-automated NIH tools) and therefore is not available for the most recent citations. The MEDLINE dataset is mostly in English but also includes a significant volume of abstracts translated from other languages.

2.2. Nearest Centroid Classifier

We have made available an automated classifier inspired by [21] and based on [10] that is able to suggest MeSH categories for any health-related text. It is trained with the part of the MEDLINE dataset that is already annotated with MeSH and is able to suggest categories for submitted text snippets. These texts can be abstracts that do not yet include MeSH classification, medical summary records or even health related news articles.

To have an efficient automated annotation of text based on the already existing health-related categories provided by MeSH, we use the nearest centroid classifier [22] constructed from abstracts of the MEDLINE dataset and their associated MeSH annotations. Each document is embedded in a vector space as a feature vector (bag-of-words BoW) of Term Frequency-Inverse Document Frequency (TFIDF) weights [18]. The features are words and phrases and the weights reflect how important a word or a phrase is to a document within the collection. The classifier is trained in such a way that for each category, a centroid is computed by averaging the embeddings of all the documents in that category. For higher levels of the MeSH structure, as suggested in [27], we also include all the documents from descendant nodes when computing the centroid. To classify a document, the classifier first computes its embedding and then assigns the document to one or more categories whose centroids are most similar to the document's embedding. We measure the similarity using pairwise cosine distance (BoW Similarity Measure) between the embeddings. Figure 3 shows the architecture of the MeSH Classifier, where the BoW Vocabulary is calculated on the whole collection of abstracts during the process of training the classifier and is used for embedding documents in a vector space. BoW Similarity Measure is used during the process of classification, when documents are being annotated by MeSH categories.

The classifier checks all the categories and returns an ordered list of all the MeSH categories according to their relevance for annotating the provided text. The demonstrator version of the MeSH classifier available online through a web portal[1] (shown in Figure 2) provides the position number in the annotation and the percentage representing the weight of the MeSH term in the annotation (based on the cosine similarity). The classifier can be also used through a REST API[2], using a POST call, and taking a JSON input that includes the text to be classified. The availability of a well-defined API facilitates further integration with news aggregators (discussed in Section 4) and other systems. The MeSH classifier can suggest categories for any text documents including research papers, medical reports and news articles.

[1] https:// qmidas.quintelligence.com/classify-mesh-major/

[2] https://qmidas.quintelligence.com/classify-mesh-major/api/classify

2.3. Large Pretrained Transformer classifier

Parallel to the Nearest Centroid classifier we implemented a text classifier based on XLNet, a large pretrained transformer model inspired by BERT [8]. We used the XLNet-Base cased variant with 12-layers, hidden size of 768, 12 attention heads and totalling 110M parameters which matches the model parameters of the base BERT model. We chose XLNet over BERT simply because it has performed better on several text classification tasks. Our model adds a linear layer following the final hidden state corresponding to the [CLS] token (commonly used as the aggregate sequence representation). The model is then fine-tuned on our data using Binary Cross Entropy (BCE) loss. This setup is similar to the original BERT and XLNet papers except for the use of BCE loss, as in this work, the end task is framed as a multilabel task rather than simply multiclass. This is also different from the Nearest Centroid classifier which is trained as a fully hierarchical model. While the Nearest Centroid classifier was used for several experiments and explorations, the Large Pretrained Transformer classifier drew on some of those, such as a fixing label depth at 3, and is presented here only as a comparison to show the potential of the results that can be achieved when using state of the art text classification approaches.

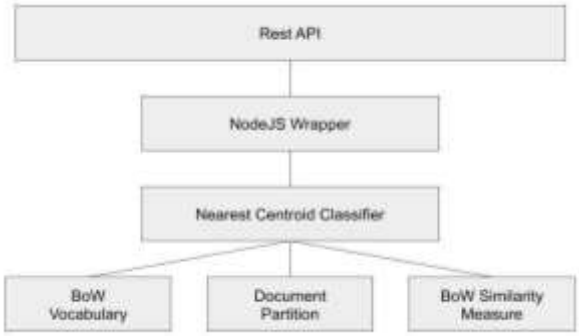


Figure 3 - A high-level diagram of the MeSH Classifier architecture

2.4. Learning approach

Research showed there is no existing database of curated and reliable data that can be used as the golden standard for testing automatic MeSH classifiers. For this purpose, we evaluated the system by leaving out one year of MEDLINE abstracts to be used as an evaluation dataset. For this reason, the classifier was trained on the MEDLINE 2018 dataset (27837540 article abstracts), leaving the latest batch of annotated abstracts out. Then, the model was evaluated against new data from the MEDLINE 2019 dataset (325128 articles), i.e., the hand-annotations of the articles that were not annotated in the 2018 version. Both of the MEDLINE dataset versions are accessible at [35], while the MeSH data including the MeSH tree is available at [36]. For the sake of the coherency of the evaluation, the new MeSH descriptors in the 2019 version that are not present in the 2018 version were not considered in the comparison between the automated annotation and the hand annotation. The complexity of the MeSH tree impacts the learning procedure in the time spent to this aim.

To improve the learning time, we reduced the complexity of the MeSH tree by deleting some of the branches that (i) are loosely related to the public health and healthcare topics (e.g., *Geographicals*); (ii) that can be better classified with other taxonomies (e.g., *Information Science*); or the refer to the bibliographical details (e.g., *Publication Characteristics*). Preliminary tests showed that the results of the automated classification of health-related text snippets were not impacted by this reduction of the MeSH classes considered. Though, the learning time was much improved with the mentioned reduction of the MeSH tree complexity.

3. EVALUATION

3.1. Evaluation methodology

3.1.1. Evaluation over research papers

The main goal was to determine the performance of the MeSH nearest centroid classifier for medical and news texts. Additionally, the evaluation results should provide an estimate for an optimal similarity cut-off, classification depth and a decision regarding the classification of all MeSH terms or major classes only. In the case of the Large Pretrained Transformer classifier, we don't do any hyper parameter tuning and instead use the parameters chosen in the original BERT paper for its evaluation on the GLUE benchmark tasks [38] with the initial learning rate of 3e-5.

3.1.2. Evaluation over news articles

To guarantee the coherence of the evaluation, we used an adaptation of the evaluation described in Section 3.1.1., but in this case we have the domain experts annotating articles selected in the context of the health domains corresponding to their areas of expertise. Each of the five experts provided between MeSH annotations to news articles on topics related to their expertise. This dataset of hand-annotated news articles, as well as the results in full of the evaluation of the MeSH classifier over these are available at [26]. The dataset was built by collecting news articles that relate to the five topics selected, and are distinguished by these, including the article title and body, as well as the assigned MeSH heading identifiers that can be used to consult more information on each of them through PubMed. This allowed us to evaluate our classifiers in these specific health topics (e.g., diabetes or mental health).

			Meaning and description
TP	True Positive	correctly identified	Number of detected MeSH classes matching the manual annotation.
FP	False Positive	incorrectly identified	Number of detected MeSH classes not matching the manual annotation.
TN	True Negative	correctly rejected	Number of not detected MeSH classes that are also not in manual annotation. This measure is not needed in the F1 calculation.
FN	False Negative	incorrectly rejected	Number of manual annotations that were not detected by the MeSH classifier.

Table 2 - Description of the variables used in MeSH classification

3.1.3. Adopted evaluation approach

In our evaluation approach, we use the following measures: precision, recall, and F1-score. In our case, precision is the fraction of detected MeSH terms that are relevant for a specific article. Recall is the fraction of the relevant MeSH terms that are successfully detected by the classifier, i.e., the number of correct classes divided by the number of classes that should have been returned. To compute these values in case of the centroid classifier, we choose a cutoff similarity and associate the document with all categories that have a higher similarity. The best performing cutoff similarities differ slightly in different domains, but are generally around 0.2. What we consider as "correct" annotations are the terms that were manually assigned by the experts. The precision and recall in terms for Type I and Type II errors can be expressed through the descriptions in the Table 2. A combination of precision and recall is provided in the form of F1-score. F1-score is the harmonic mean of precision and recall. Note that, to proceed with the evaluation of the system we left out one year of MEDLINE abstracts to be used as an evaluation dataset.

3.2. Nearest Centroid Classifier Tuning

The evaluation results provide an estimate of an optimal similarity cut-off, and classification depth. We start from the evaluation of the classifier against annotated scientific articles, and then elaborate on the evaluation of the classifier against annotated news articles.

3.2.1. Determine optimal similarity cut-off

The classifier provides a weight value for each label. We would like to determine what would be the optimal cut-off to be used for reporting the predicted labels. In principle, by decreasing the cut-off value we increase the precision and decrease the recall. In Figure 5 we show the value of F1 for different cut-off thresholds ranging from 0 to 0.9 with step 0.1.

3.2.2. Determine optimal tree depth for classification

We perform our evaluation by comparing matches at different tree depths. The reason is that the aim of the classifier is to assist experts in detecting broader topics, and not find exact MeSH term matches. As the system has been trained on abstracts only, we find that exact matches would be difficult to achieve. Another reason for aiming at broader topics is the intended usage of the classifier on non-medical articles (e.g., news), where detecting those broader topics is a pragmatic goal.

3.3. Evaluation results for setting the parameters on the annotation of MeSH classes over scientific papers

3.3.1. Evaluation Setting

We have used the experimental evaluation to determine whether only the major MeSH classes should be returned by the classification model or all the relevant MeSH classes. A major MeSH class is a MeSH descriptor, which is viewed as a focus of the paper, while minor MeSH classes are mentioned in the paper. For example, in a paper on survival following myocardial infarction in Ireland, myocardial infarction would be a major term, and Ireland a minor term, as the focus of the paper is on survival, and not on the locations where the patients lived.

3.3.2. Evaluation Results

The diagrams of Figure 4 compare the evaluation results based on the recall (colour code) over the similarity cut-off threshold (Y-axis) and MeSH tree depth (X-axis), for major MeSH classes with those for all MeSH classes. As expected, higher cut-off yields higher precision results and lower recall results. Lower depths yield both better precision and recall. Roughly, for the depth 3 in the MeSH tree, over cut-off values of approximately 0.35, precision increases to over 50% and below cut-off values of approximately 0.25, recall increases to over 50%. While the performance at level 1 is good, it decreases significantly with greater tree depths. Still, a cut-off 0.35 at level three, would provide average precision 0.64 and F1 = 0.16. The F1 measure then increases to 0.2 in for similarity cut-off 0.32, and to 0.3 when the cut-off is below 0.26. Overall, the performance at levels 1 and 2 is good, at level three acceptable, and it then decreases significantly for

depths greater than three. Here, a cut-off 0.2 at level two, would provide average precision 0.74 and F1 = 0.55; at level three the precision would be 0.6 and F1=0.44 (this variation is illustrated by Figure 5).

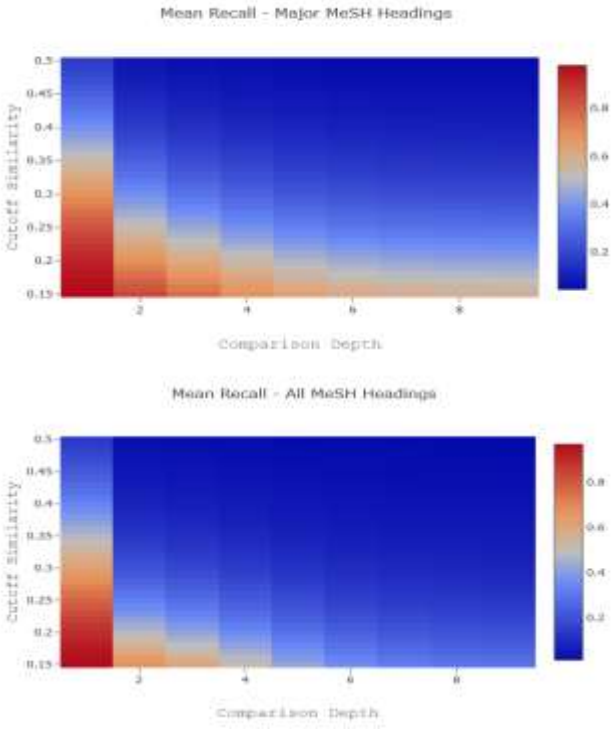
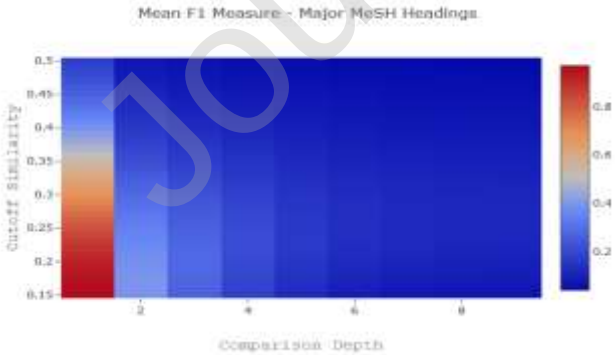


Figure 4 - Precision in the comparison between the MeSH tree depth and the cut-off based on similarity between major MeSH classes (above) and all MeSH classes (below).

3.3.3. Evaluation conclusions

We evaluated several combinations of the three parameters: similarity cut-off [0.15,...,0.5], MeSH tree depth [1,...,7] and major vs. all MeSH classes. For those combinations, we have calculated the average precision, recall, and F1 measures.
As expected, higher cut-offs yield higher precision results and lower recall results. Moreover, lower depths yield both better precision and recall. Compared to evaluation of major MeSH classes, the precision for this evaluation is much better, while recall results for major MeSH classes are slightly better than for this evaluation. Roughly, for tree depth 3, over cut-off values of approximately 0.25 precision increases to over 70% and below the same cut-off recall is around 30%. At tree depth three, which is the aim for news item annotation, results are of acceptable quality considering the aim to give emphasis to precision at level three. In conclusion, classification of all MeSH classes performs significantly better at the desired depth of classification, depth three. At that classification depth it is estimated that the optimal similarity cut-off is around 0.2, with precision 0.6 and F1=0.44. At that same depth (depth 3), the XLNet based Large Pretrained Transformer classifier achieves a superior F1=0.56, a very significant improvement when comparing with the results of the nearest centroid classifier.



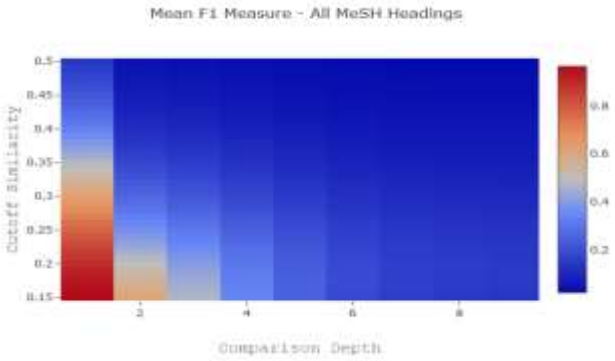


Figure 5 - F1 measures in the comparison between major MeSH classes (above) and all MeSH classes (below), distinguishing the MeSH tree depth and the cut-off based on similarity for major MeSH classes.

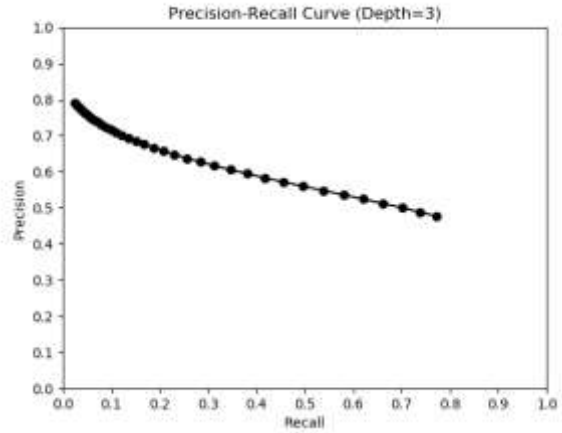


Figure 6 – Precision-recall curve contributing to the optimal choice of the appropriate cut-off at MeSH tree depth 3

3.4. Evaluation results for the annotation of MeSH classes over news articles

3.4.1. Evaluation Setting

In a second phase of this study, we evaluated the classifier in the context of news articles. For this purpose, we asked five experts (i.e., health professionals with experience in the usage of MeSH) to annotate news articles using the MeSH classes.

Based on the analysis of the prior evaluation over research articles, we considered that the annotation could go up to a fourth level of deepness in the MeSH tree. Thus, we proceeded with providing each of the five experts with a set of news articles and a spreadsheet where they should annotate with three to ten MeSH classes each of the articles. The appropriate MESH Id can be consulted and obtained at the *NIH MeSH Tree View*[3] and *NIH MeSH Search*[4]. In that spreadsheet, each line was an article and the MeSH classes that are annotating it. An example of this annotation, in the context of diabetes, is the annotation of the news article *"Obesity Stigma And Yo-Yo Dieting Are Behind Chronic Health Conditions, True or False?"* [5] with the following MeSH descriptors: Obesity D009765; Diet, Reducing D004038; Chronic Disease D002908; Statistics as Topic D013223; Social Stigma D057545; Causality D015984; Correlation of Data D000078331; Risk D012306; Complementary Therapies D000529; Epidemiologic Methods D004812.

For MeSH-based manual annotation, we selected as news topics the five health domains that correspond to recent priorities of European public health authorities [6]: mental health, diabetes mellitus, coronavirus, childhood obesity, and child in care. In the following paragraphs, we present the results of the evaluation for the first three health scenarios that we chose to analyse in depth. In the conclusions section, show the full range perspective covering the evaluation of the classifier over the five health topics. This will provide us with a range of different examples, which can lead to some conclusions regarding the annotation of health-related news through the MeSH classifier.

[3] https://meshb.nlm.nih.gov/treeView

[4] https://meshb.nlm.nih.gov/search

[5] https://www.edgemedianetwork.com/health_fitness/health//282003

(MC1) The term *mental health* (MeSH ID 68008603) exists in MeSH with the unique ID D008603 since 1967. It is defined as the emotional, psychological, and social well-being of an individual or group. It falls on two paths in the MeSH tree under the roots *Psychiatry and Psychology* MeSH category (at deepness 3) and *Health Care* MeSH category (at deepness 4). There are 81433 scientific articles hand-annotated with this MeSH class in the 2019 version of MEDLINE we use.

(MC2) The term *diabetes mellitus* (MeSH ID 68003920) exists in MeSH with the unique ID D003920 since 1984. It is defined as a heterogeneous group of disorders characterized by hyperglycemia and glucose intolerance. It falls on two paths in the MeSH tree under the root *Diseases Category* (at deepness 3 and 5). The 2019 version of MEDLINE we use here includes 315341 scientific articles hand-annotated with this MeSH class.

(MC3) The term coronavirus (MeSH ID 68017934) also exists in the MeSH tree with the unique ID D017934 since 1994. It is defined as being part of the CORONAVIRIDAE disease family, which causes respiratory or gastrointestinal disease in a variety of vertebrates. It falls on a unique path in the MeSH tree under the roots Coronaviridae (at depth 6). There are 5976 scientific articles hand-annotated with this MeSH class in the 2019 version of MEDLINE we use.

3.4.2. Evaluation Results

The diagrams in Figure 6 show the evaluation results where the X-axis is the tree depth, the Y-axis is the similarity cut-off threshold, and the colour code is the result of the evaluation (precision, recall, and F1 measures). Table 3 summarizes the most important results and Table 4 compares the results of both classifiers.

Higher cut-off thresholds yield higher precision results and lower recall results. Moreover, lower depths yield both better precision and recall. Roughly, in the case (MC1) for depth 3, over cut-off values of approximately 0.35 precision increases to over 50% and below cut-off values of approximately 0.25 recall increases to over 50%. In the case (MC2), for three depth 3, over cut-off values of approximately 0.27 precision is reaching 80% and below cut-off values of approximately 0.25 recall decreases to over 60%. Similarly, in the case (MC3), for three depth 3, over cut-off values of approximately 0.37 precision increases to over 60% and below cut-off values of approximately 0.23 recall increases to over 50%.

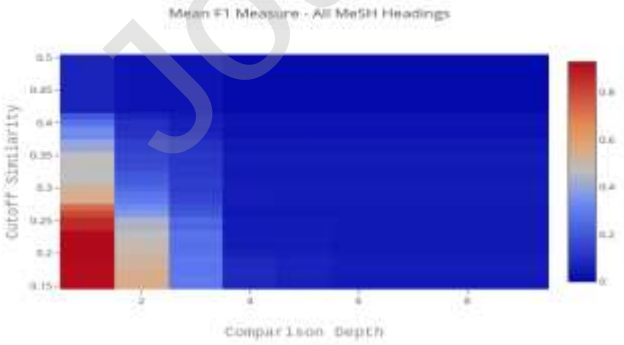
For the case (MC1), a cut-off around 0.3 at level two would provide F1 above 50%. The case (MC2), over a cut-off 0.35 at level two, shows F1 above 0.5. Finally, in the case (MC3), a cut-off of 0.35 at level three, would provide average precision 0.64 and F1 = 0.11.

3.4.3. Evaluation Conclusions

The study has evaluated combinations of the two parameters - similarity cut-off and MeSH three depth - for each of the five MIDAS use cases. We have calculated the average precision, recall, and F1 measure for each of these. In conclusion, the classification of MeSH classes over news articles performs significantly better at the desired depth of classification, depth three. Though the variation between the evaluation of the classification of news articles with different health topics vary over a small range up to deepness 3 much increasing after that.

The results are good up to tree depth three in all cases, where the F1 measure yields similar results. Moreover, the performance at level one is good, and decreases significantly with greater tree depths. At that classification depth it is estimated that the optimal similarity cut-off is around 0.2, providing on average a F1 measure above 40%. In all the cases analysed, at tree depth three, which is what the researchers aim for the annotation of news articles, the results vary significantly across the health topics in analysis.

In the Table 3 we present results for the five health domains studied providing the optimal threshold values (cut-off, F1) per a given depth of MeSH tree search. The poor results in the case *Children in Care* are mostly due to the limited number of news items that effectively discuss the topic, making it difficult to have a reasonable expert hand annotation of the related news articles. The case of the annotation of news on the topic of Coronavirus is also poor, mostly due to the challenging task of the annotation of the news article because the dimension of the topic in the media is not only on the health sphere, but affecting the lifestyle of the population, the economy, and many other aspects relevant for society In general.



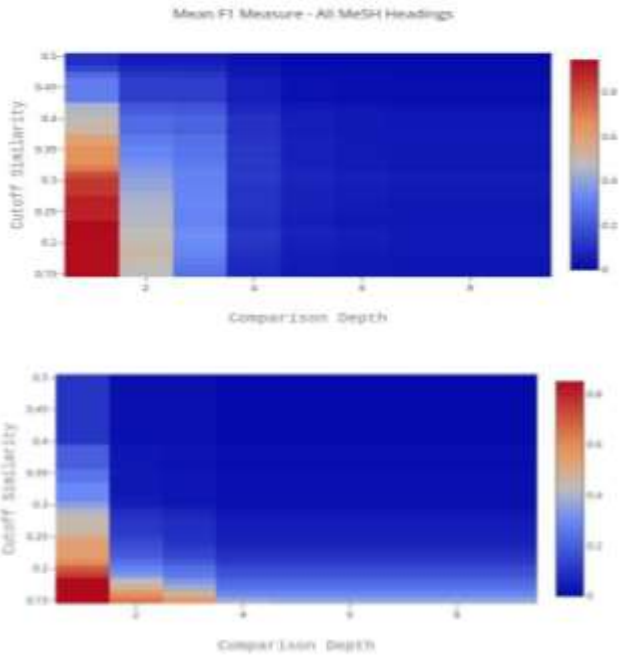


Figure 7 - F1 measure in the comparison between the evaluation of news articles with focus on mental health (above), diabetes mellitus (middle) and Coronavirus (below), considering the MeSH tree depth and the cut-off based on the similarity for major MeSH classes

Similarly, the topic mental health in news articles is poorly annotated. It is represented by a large variety of aspects of the disease in society and the way authors express these in their content. Although this is a topic of growing awareness and exposure, it is still not fully understood and accepted by the general public. This might be a substantial part of the challenge for this specific topic, in the sense that it is sometimes introducing ambiguity when comparing with other topics that are treated in a more precise manner by news outlets (e.g., diabetes).

4. CONCLUSIONS AND FUTURE WORK

We have proposed an innovative approach to advanced search of health-related news to support the workflow of health professionals. The proposed health classification enables the exploration of news and extraction of health-related insights, based on the MeSH vocabulary. One of the impactful applications of that annotation is the exploration of news events and articles in a similar way as in PubMed, with which most researchers in the health domain are familiar with.

The experimental evaluation shows that it is possible to leverage the annotations of scientific articles to automatically annotate news articles. For these, we invited domain experts to annotate the articles with MeSH classes and confronted those with the annotation of our classifier. The classification of health-related news articles is the main contribution of this paper. Text in scientific articles and news is very different, effectively a difference in domain. But there is also a marked difference across different health topics and how they are exposed in the media. One such example is the difference between the coverage of mental health and diabetes, mostly due to the fact that mental health is still a topic that fails to be considered in full, perhaps, due to social stigma or to its perception as a less important topic.

Table 4, which compares both classifiers on the news data, shows that the Nearest Centroid classifier outperformed the XLNet based classifier in 3 out of the 5 health domains. This is surprising given that the XLNet classifier significantly outperforms the Nearest Centroid classifier on the medical papers dataset.

Health Domains	***MeSH Tree Depth***			
	1	2	3	4
	F1	F1	F1	F1
Mental Health	0.930	0.5515	0.2905	0.0839
Diabetes Mellitus	0.9442	0.5066	0.3422	0.1196
Coronavirus	0.8873	0.3448	0.2790	0.0814
Childhood Obesity	1	0.7619	0.6355	0.2998
Children in Care	0.9883	0.5765	0.3315	0.1402

Table 3 - The results of the Nearest Centroid MeSH classifier evaluation on news throughout five different health domains, describing the optimal threshold values (cut-off and F1) per MeSH tree depth.

Health Domains	***Classifier***	
	Nearest Centroid	XLNet
Mental Health	0.2905	0.2743
Diabetes Mellitus	0.3422	0.4388
Coronavirus	0.2790	0.3692
Childhood Obesity	0.6355	0.4842
Children in Care	0.3315	0.2345

Table 4 - Comparison of both MeSH classifiers results (F1-score) on our news dataset at a fixed label depth of 3.

The automated evaluation in this paper considered several combinations of relevant parameters: a similarity cut-off; and the MeSH three depth. In the case of scientific articles, we compared the evaluation using only the major MeSH classes (annotated by domain experts) against all MeSH classes. In the case of news articles, we compared the evaluation of news relating to five public health policy domains: mental health, diabetes, coronavirus, childhood obesity, and children in care.

For the above combinations, we have calculated the average precision, recall, and F1 measures. With this analysis we conclude that the classification of scientific articles with all MeSH classes performs significantly better at the MeSH tree depth of classification 3. At that classification depth it is estimated that, for the Nearest Centroid classifier, the optimal similarity cut-off is around 0.2, with an average F1 measure around 0.4. In the case of diabetes, the three depth 4 over cut-off values of approximately 0.38 show precision increase to over 60%, while below cut-off values of approximately 0.36 recall increases to over 51%. In the classification of news articles, the health domain that it relates to and the frequency of news seems to have an impact in the evaluation. This study shows that news articles about diabetes get better evaluation results than those on the Coronavirus mostly because of the diversity in the scope of news reflecting the impact in several domains other than health, and the gap to that learned over scientific articles. Predictably, classifying health news at a depth greater than 3 is challenging. The additional specificity of labels easily leads to increases in sporadic correlations between news texts and scientific articles that a text classifier can pick up, leading to a decrease in precision. Simultaneously, even when details can allow a human expert to label at those greater depths, this knowledge is hard to capture given a classifier trained exclusively on scientific articles.

Taking into consideration the results obtained in the classification of health-related news, we will further explore the novelty presented by this MeSH classifier in that domain that entails its own challenges. Future work includes the improvements to the classifier itself, through a differentiated assignment of importance to the MeSH tree branches, refining those that are taken into consideration, and using weights to distinguish the relevance of the classes. Another envisaged improvement is the appropriate inclusion of the information obtained by the qualifiers (that, unlike the descriptors used as MeSH classes in the proposed classifier, provide complementary information). With regards to the Large Pretrained Transformer classifier, results should improve with domain adaptation. BioBERT [15] leveraged unsupervised pretraining to improve results in scientific article classification. BERTMeSH [42], showed further improvements via supervised pretraining (fine-tuning) on the specific dataset to be classified. Both unsupervised and supervised pretraining on health news data are possible. Unsupervised pretraining only requires the categorization of news articles as being health-related which is not difficult. Supervised pretraining would require clever tricks to match news articles to labeled scientific literature or a costly amount of expert hand labeling. An exploration of other domain adaptation techniques (between scientific articles and news articles) could also help push the results of on news further. Another avenue that should be explored in terms of news article classification is in identifying labels to prune: i.e. labels that may make sense in scientific literature but are unlikely to be meaningful in health news. How to perform such pruning automatically or semi-automatically is an open question.

Further research also includes the evaluation of news articles on a wider range of health-related topics (including, e.g., asthma) which will require a substantial increase of domain experts to annotate with the MeSH classes a selection of related articles. This will provide us with a larger perspective on the efficiency of the classifier across the public health and healthcare scope. It will also help us better understand where learning from the MEDLINE dataset (and from the narrower scope scientific articles it includes) is sufficient to have a satisfactory result on the classification of news articles related to those health topics.

The specific challenges in the hand annotation of MEDLINE articles (where one can annotate with a term and the reader can assume that related terms are represented) might impact the efficiency of the built classifier, which is learning from this labelled dataset. The precision to which the MeSH tree can reach in many health domains is reflected in the choice of MESH classes that are used by the MEDLINE experts to classify these scientific articles. We suspect that the human assumption both from the side of the expert providing the hand annotation and the human reader, make equivalent the annotation of two slightly different MeSH classes. The automated classification does not make these assumptions based on the multitude of parameters inherent to the context of the scientific article, but humans can. This can be partially solved by relying on higher nodes in the tree. Though, an automated classification that aims to provide MeSH classes deeper in the tree would need to tackle this problem.

Furthermore, the evaluation parameters obtained will be used to further optimise the classifier and evaluate the classifier improving its classification of news articles. The proposed classifier enables the user to follow a workflow similar to that of exploring scientific articles in *PubMed* when monitoring health news, and to extract further insights from the monitored news previously annotated (e.g., using the

MeSH headings in their search, as proposed in [27]). It also enables new functionalities that are based on the MeSH terminology (see Figure 1 for an example of a data visualization module allowing the user to account for the percentage of news articles that talk about a specific MeSH heading related to the search topic).

We aim to further explore the integration of this MeSH text classifier with the exploration and monitoring of local and worldwide news, as well as in social media. This is an important task in Public Health, impacted by the choice of the appropriate parameters that can express the defined priorities. The accurate monitoring of worldwide news contributes to a global perspective of world health, but also to the aspects of regional health where public health institutes can act. Though, this will require building it from a cross-lingual classifier. It can also contribute to the evaluation of the success of public health campaigns by allowing decision-makers to assess what the news media's response was to them, often reflecting the opinion of their communities. Moreover, the further exploration of health news articles can help health professionals to avoid news bias in the era of *fake news* [25].

Conflict of interest statement

There are no conflicts of interest

ACKNOWLEDGMENTS

This work was supported by the European Commission H2020 project MIDAS (G.A. nr. 727721).

REFERENCES

[1] A. Aronson et al (2004). The NLM indexing initiative's medical text indexer. Medinfo, vol. 89.
[2] ArcGis (2020). WHO's COVID-19 disease monitoring. url: https://experience.arcgis.com/experience/685d0ace521648f8a5beeeee1b9125cd. Accessed: 20 March 2020.
[3] S. Baker and A.L. Korhonen (2017). Initializing neural networks for hierarchical multi-label text classification. Association for Computational Linguistics. BioNLP 2017, Association for Computational Linguistics. pp. 307–315.
[4] M. Black et al (2019). Meaningful Integration of Data, Analytics and Services of Computer-Based Medical Systems: The MIDAS Touch. *32nd IEEE CBMS International Symposium on Computer-Based Medical Systems.*
[5] O. Bodenreider (2004). *The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl_1), D267-D270.*
[6] Boilson, A., Connolly, R., Staines, A., Davis, P., Connolly, J., Weston, D. (2019). Improving European Healthcare Systems through the Development of a Realist Evaluation Framework for a European Public Health Data Analytic Project. *Biomed Central* (BMC) Implementation Science Journal.
[7] S. Dai, R. You, Z. Lu, X. Huang, H. Mamitsuka and S. Zhu (2020). FullMeSH: improving large-scale MeSH indexing with full text. Bioinformatics, 36(5), pp. 1533-1541.
[8] J. Devlin, M-W. Chang, K. Lee and K.Toutanova (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171-4186,
[9] L. Fang, G. Karakiulakis, and M. Roth. (2020). Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection?. The Lancet Respiratory Medicine. DOI:10.1016/S2213-2600(20)30116-8
[10] L. Henderson, Lachlan (2009). Automated text classification in the DMOZ hierarchy. TR.
[11] S. M. Humphrey (1989). "MedIndEx system: medical indexing expert system." Information Processing & Management 25.1 (1989): 73-88.
[12] Q. Jin, B. Dhingra, W. Cohen and X. Lu (2018). AttentionMeSH: simple, effective and interpretable automatic MeSH indexer. In Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering. pp. 47-56.
[13] Kaggle (2020). COVID-19 Open Research Dataset Challenge - CORD-19. url: https://www.kaggle.com/allen-institute-for-ai/ CORD-19-research-challenge. Accessed: 20 March 2020.
[14] H. Kilicoglu et al (2008). Semantic MEDLINE: a web application for managing the results of PubMed Searches. In Proceedings of the third international symposium for semantic mining in biomedicine. Vol. 2008, pp. 69-76.
[15] J. Lee et al (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.
[16] D. A. Lindberg (2000). Internet access to the National Library of Medicine. *Effective clinical practice: ECP*, *3*(5), 256.
[17] K. Liu, S. Peng, J. Wu, C. Zhai, H. Mamitsuka and S. Zhu (2015). MeSHLabeler: improving the accuracy of large-scale MeSH indexing by integrating diverse evidence. Bioinformatics, 31(12), i339-i347.
[18] C. Manninget al (2008), "Introduction to Information Retrieval," Cambridge Univ. Press.
[19] Y. Mao and L. Zhiyong (2017) "MeSH Now: automatic MeSH indexing at PubMed scale via learning to rank." Journal of biomedical semantics 8.1: 15.
[20] Midas Project (2020). A MIDAS contribution to the global COVID-19 monitoring strategy. url: http://www.midasproject.eu/2020/03/13/ a-midas-contribution-to-the-global-covid-19-strategy/. Accessed: 20 March 2020.
[21] D. Mladenic (1998). Turning Yahoo into an automatic Web-page classifier. In: Prade, H. (ed.). Proceedings of *European Conference on Artificial Intelligence (ECAI)*. Chichester [etc.]: John Wiley & Sons, pp. 473-474.
[22] D. Mladenic and M. Grobelnik (2003). Feature selection on hierarchy of web documents. Journal: *Decision support systems*. vol. 35, pp. 45-87.

[23] D. Newman, S. Karimi, and L. Cavedon (2009). Using topic models to interpret MEDLINE's medical subject headings. In Australasian Joint Conference on Artificial Intelligence, pp. 270-279. Springer, Berlin, Heidelberg, 2009.
[24] S. Peng, R. You, H. Wang, C. Zhai, H. Mamitsuka and S. Zhu (2016). DeepMeSH: deep semantic representation for improving large-scale MeSH indexing. Bioinformatics, 32(12), i70-i79.
[25] J. Pita Costa et al (2019). Health News Bias and Epidemic Intelligence for Public Health. Proceedings of the SIKDD 2019.
[26] J. Pita Costa et al (2020). MIDAS hand-annotated news articles [dataset]. Zenodo. doi.org/10.5281/zenodo.4034281
[27] J. Pita Costa et al (2017). Text mining open datasets to support public health. In *Conf. Proceedings of* WITS 2017.
[28] J. Pita Costa et al (2019). The meaningfulness of open data in public health and healthcare. Proceedings of the 12th European Public Health Conference 2019.
[29] D. Rankin et al (2017). The MIDAS Platform: Facilitating the Utilisation of Healthcare Big Data in Northern Ireland and Beyond. In the *8th Annual Translational Medicine Conference. Clinical Translational Research and Innov. Centre (C-TRIC)*.
[30] F. B. Rogers (1963). Medical subject headings. Bulletin of the Medical Library Association, 51(1), 114-116.
[31] M. Sarrouti and Said Ouatik El Alaoui(2020). SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions. Artificial Intelligence in Medicine 102 (2020): 101767.
[32] P. Srinivasan and B. Libbus (2004). Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics*, *20*(Suppl 1), i290–i296.
[33] Y. Yan et al (2018). Biomedical literature classification with a CNNs-based hybrid learning network. *PloS one*, *13*(7), e0197933.
[34] UNESCO International Research Institute on Artificial Intelligence – IRCAI (2020). IRCAI's COVID-19 disease monitoring. url: http://coronaviruswatch.ircai.org/. Accessed in: 20 March 2020.
[35] U.S. National Library of Medicine – NLM (2020). MEDLINE dataset. url: https://www.nlm.nih.gov/databases/download/pubmed_medline.html. Accessed in: 1 September 2020.
[36] U.S. National Library of Medicine (2020). MeSH Data. url: https://www.nlm.nih.gov/databases/download/mesh.html. Accessed in: 1 September 2020.
[37] U.S. National Library of Medicine (2020). MeSH on Demand. url: https://meshb.nlm.nih.gov/MeSHonDemand. Accessed: 20 March 2020.
[38] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy and S. R. Bowman (2019). Glue: A multi-task benchmark and analysis platform for natural language understanding}. 7th International Conference on Learning Representations, ICLR 2019.
[39] World Health Organisation – WHO (2020). WHO Director-General's opening remarks at the media briefing on COVID-19—11, 11 March 2020.
url: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. Accessed: 20 March 2020.
[40] Yu-Tao Xiang, Yuan Yang, Wen Li, Ling Zhang, Qinge Zhang, Teris Cheung, and Chee H. Ng (2020). Timely mental health care for the 2019 novel coronavirus outbreak is urgently needed. The Lancet Psychiatry 7, no. 3: 228-229.
[41] G. Xun, K. Jha, Y. Yuan, Y. Wang and A. Zhang (2019). MeSHProbeNet: a self-attentive probe net for MeSH indexing. Bioinformatics, 35(19), pp. 3794-3802.
[42] R. You, Y. Liu, H. Mamitsuka and S. Zhu (2020). BERTMeSH: Deep Contextual Representation Learning for Large-scale High-performance MeSH Indexing with Full Text. bioRxiv.