MeSH News: a new classifier designed to annotate news with MeSH heading classes

Joao Pita Costa [1] [2], Luka Stopar [1] [2], Flavio Fuart [1] [2], Luis Rei [1], Marko Grobelnik [1] [2], Anthony Staines [4], Jarmo Pääkkönen [5], Jenni Konttila [5], Joseba Bidaurrazaga [6], Oihana Belar [6], Christine Henderson [7], Gorka Epelde [8][9], Mónica Arrúe [8][9], Paul Carlin [10], Jonathan Wallace [11], Dunja Mladenić [1] [2], Inna Novalija [1], Ricardo Mexia [3]

ABSTRACT

Motivation: In the age of 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 based on the MEDLINE/MeSH thesaurus, trained on the manual annotation of more than 26 million expert-annotated scientific abstracts. 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.

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

Conclusions: The proposed text classifier shows good results both in the evaluation of scientific text and on health related news. 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.

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 for 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.

Aiming to address this need in part, we have developed a text classifier trained on the existing hand annotations of MeSH heading classes given to biomedical research papers in MEDLINE. The classifier provides automated annotations of text, 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) Instituto Ricardo Jorge (INSA), Portugal, (4) Dublin City University, Ireland, (5) University of Oulu, Finland (6) BIOEF, Spain, (7) Northern Ireland Department of Health, UK, (8) Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Spain, (9) Biodonostia, Spain (10) South Eastern Health and Social Care Trust, (11) Ulster University, UK

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 [who] that motivated worldwide joint initiatives to help monitor the disease (as [2] and [21]), and understand it better [12], including the crowdsourcing initiative to build new machine learning methods based on biomedical knowledge [7]. In the context of these global initiatives, the proposed text classifier was designed to classify text, motivated by the potential for the classification of health reports and news articles on the 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 the diabetes [5] and the impact of the social distancing it is generating on mental health [23].

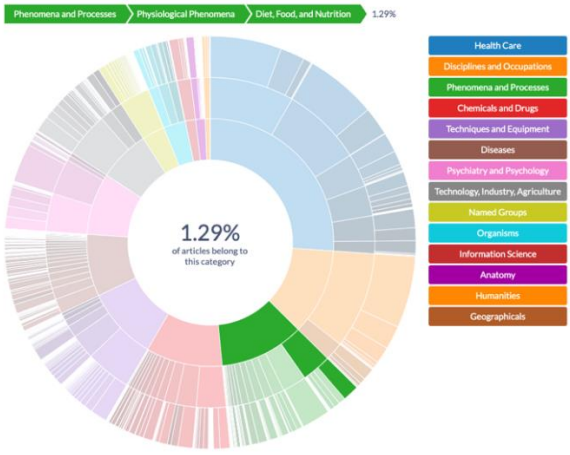

Figure 1 – A potential impact of the MeSH classifier applied to the classification of news: the percentage of news published in 2018/2019 that are annotated with the MeSH class "Public Health".

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 [9] 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 on 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 [19]. In the context of the meaningful integration and usage of data, the EU H2020 project MIDAS (Meaningful Integration of Data, Analytics and Services) [14] 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 [3].

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 [16] 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 [20]. More general approaches include: the Medical Text Indexer (MTIFL) [1], that provides MeSH indexing recommendations to support the human indexers of the NLM; the MeSH Now [11], including a learning-to-rank framework; the MeSH on Demand[1], that suggests MeSH vocabulary explicitly mentioned in the input text; and the Semantic MEDLINE [8], that aims for the semantic knowledge representation of MEDLINE itself. The latter two are very different in characteristics from the MeSH classifier proposed in this paper. The MTIFL and the MeSH Now are tools developed by the NLM designed to address the needs of the PubMed user. The Semantic MEDLINE initiative is different in the sense that it focus 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 trainned 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.

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. In particular, it includes 443 218 full-text 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 [17]. Trained NLH 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 NIH. 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.

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 includes also a significant volume of abstracts translated from other languages.

[1] https://meshb.nlm.nih.gov/MeSHonDemand

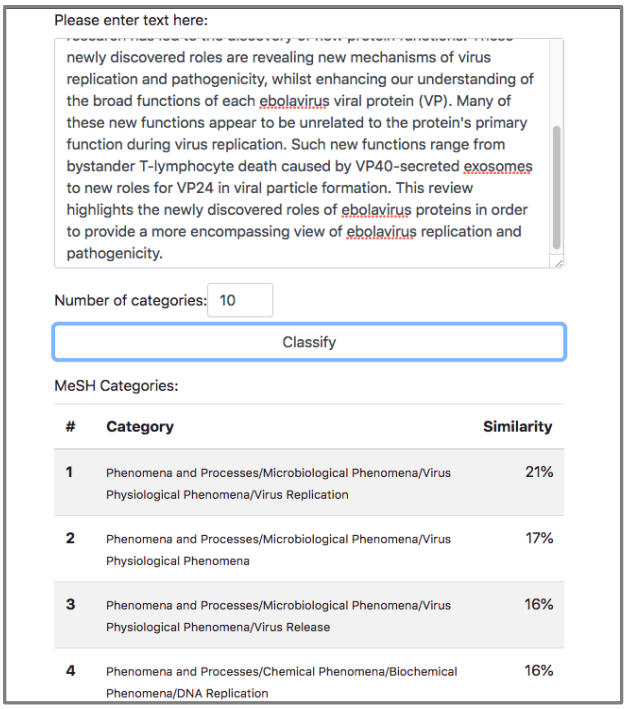

Figure 2 - An example of the MeSH classifier output for the automated MeSH annotation of a scientific article abstract extracted from PubMed (in the body of text above), the MeSH class rank (on the left), their MeSH tree path in the MeSH ontology-like structure (in the centre), and a measure of the similarity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

2.2. Classifier description

We have made available an automated classifier inspired by [12] and based on [6] 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 [14] 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. 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 [18], 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 showing 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 order 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[2] (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[3], 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..

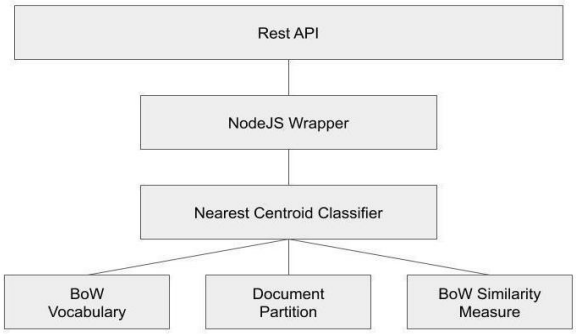

Figure 3 - A high-level diagram of the MeSH Classifier architecture

2.3. 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 (27 837 540 article abstracts), leaving the latest batch of annotated abstracts out. Then, the model was evaluated against new data from the MEDLINE 2019 dataset (325 128 articles), i.e., the hand-annotations of the articles that where not annotated in the 2018 version. 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 where 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.

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

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

3. EVALUATION

3.1. Evaluation methodology

3.1.1. Evaluation over research papers

The main goal was to determine the performance of the MeSH classifier for medical 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.

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 that in this case we have the domain experts annotating articles selected on 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 allowed us to evaluate the classifier in specific health topics as, e.g., diabetes or mental health.

3.1.3. Adopted evaluation approach

In our evaluation approach, we use the following measures: precision, recall, F1-score and F0.5 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. 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 Table 2.

			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

A combination of precision and recall is provided in the form of F1-score or F0.5 score. F1-score is the harmonic mean of precision and recall, while F0.5 score gives more weight to precision. According to the F_β formula (where β is chosen such that recall is considered β times as important as precision), we can express F1 measure for each article in terms of Type I and Type II errors as follows:

$$F_\beta = \frac{(1+\beta^2)\cdot TP}{(1+\beta^2)\cdot TP + \beta^2 \cdot FN + FP}$$

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. Optimisation approach

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 calculated 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 (news) articles, 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 tree 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 of major MeSH classes, both F1 and F0.5 measures yield similar results. 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 F0.5 = 0.17. Both F1 and F0.5 measure yield similar results also in the case of all MeSH classes. The performance at levels 1-2 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 F0.5 = 0.54; at level three the precision would be 0.6 and F0.5=0.43 (see Figure 5).

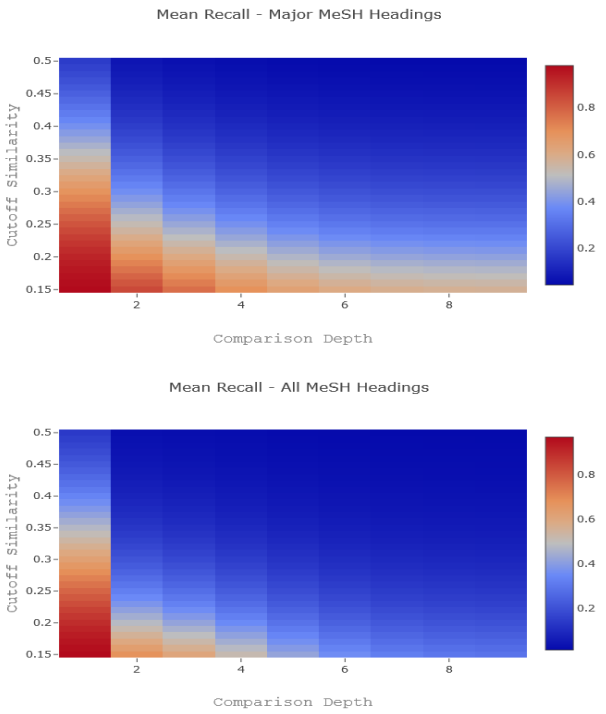

Figure 4 - Precision in the comparison between the MeSH tree depth and the cut-off based on similarity between major MeSH classes (above) and all MeSH classes (below).

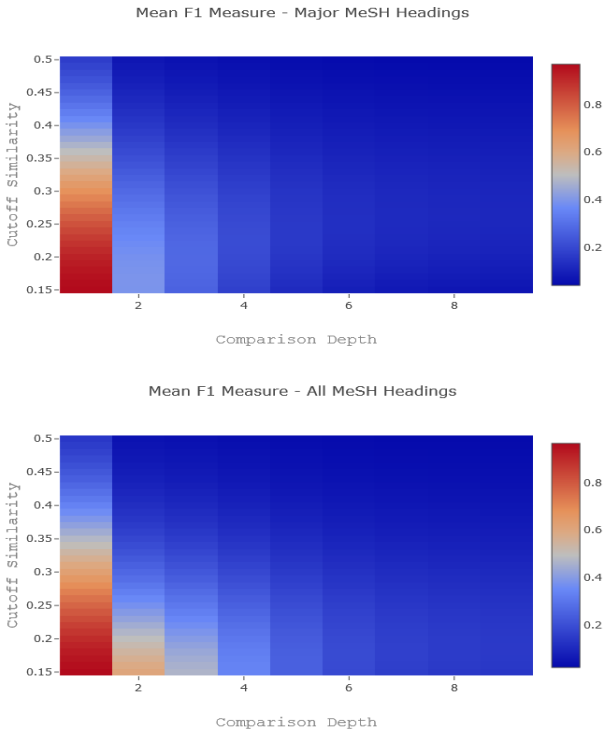

Figure 5 - F1 measures in the comparison between major MeSH classes (above) and all MeSH classes (below), distinguishing the MeSH tree depth and the cut-off based on similarity for major MeSH classes.

3.3.3. Evaluation conclusions

We evaluated several combinations of the three parameters: similarity cut-off [0.15,...,0.5], MeSH three depth [1,...,7] and major vs. all MeSH classes. For those combinations, we have calculated the average precision, recall, F1 and F0.5 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 three 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 items 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 F0.5=0.43.

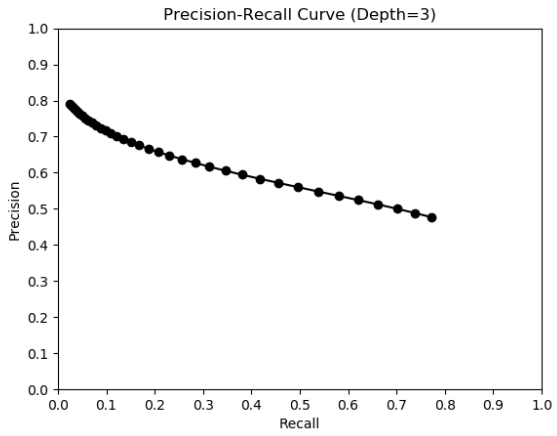

Figure 6 – Precision-recall curve contributing to the optimal choice of the appropriate cut-off at MeSH tree depth 3

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*[4] and *NIH MeSH Search*[5]. In that spreadsheet, each line was an article and the MeSH classes that are annotating it.

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

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

For MeSH-based manual annotation, we selected as news topics the five health domains that correspond to recent priorities of European public health authorities [4]: 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.

(MC1) The term *mental health*[6] 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 articles annotated with this MeSH class in the 2019 version of MEDLINE we use.

(MC2) The term *diabetes mellitus*[7] exists in MeSH with the unique ID D003920 since 1984. It is defined as a heterogeneous group of disorders characterized by hyperglycaemia 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 articles annotated with this MeSH.

(MC3) The term coronavirus[8] 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 deepness 6). There are 5976 articles 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, F1, and F0.5 measures).
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 and F0.5 above 50%. The case (MC2), over a cut-off 0.35 at level two, shows F1 above 0.5 and F0.5 around 0.4. Finally, in the case (MC3), a cut-off 0.35 at level three, would provide average precision 0.64 and F0.5 = 0.17.

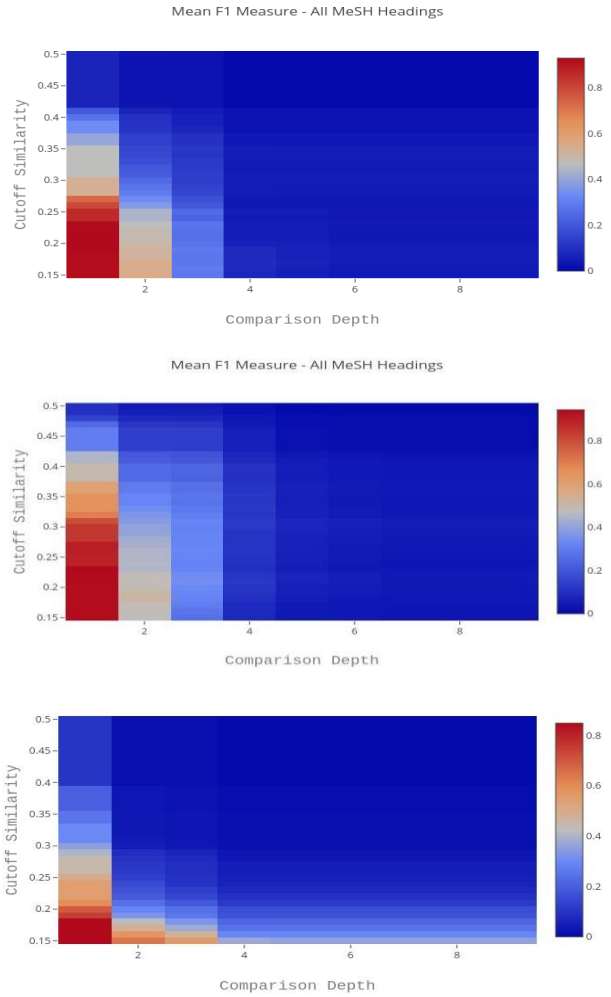

Figure 7 - F1 measure in the comparison between the evaluation of news articles with focus on mental health (above), diabetes mellitus (middle) and Coronavirus (below), considering the MeSH tree depth and the cut-off based on the similarity for major MeSH classes

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, F1 and F0.5 measures 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 both F1 and F0.5 measure yield similar results. Moreover, the performance at level 1 is good, and decreases significantly with greater tree depths. At that classification depth it is estimated

[6] https://www.ncbi.nlm.nih.gov/mesh/68008603

[7] https://www.ncbi.nlm.nih.gov/mesh/68003920

[8] https://www.ncbi.nlm.nih.gov/mesh/68017934

that the optimal similarity cut-off is around 0.2, providing precision 0.6 and F0.5=0.43. 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 the 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 challenging 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.

Health Domains	***MeSH Tree Depth***							
	0		1		2		3	
	cut-off	F1	cut-off	F1	cut-off	F1	cut-off	F1
Mental Health	0.27	0.73	0.19	0.52	0.25	0.25	0.19	0.13
Diabetes Mellitus	0.35	0.7	0.25	0.45	0.37	0.3	0.31	0.13
Coronavirus	0.25	0.5	0.24	0.19	0.21	0.19	0.18	0.17
Infectious Diseases	0.25	0.55	0.22	0.22	0.22	0.18	0.19	0.08
Childhood Obesity	0.31	0.85	0.38	0.54	0.31	0.58	0.31	0.14
Children in Care	0.38	0.66	0.24	0.2	0.34	0.08	0.25	0.03

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

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 promising results both in the classification of abstracts of scientific articles from MEDLINE, and on the annotation of news articles in English language. In the case of scientific articles, we confronted their automatic annotation with the hand annotations provided in the metadata of MEDLINE. For the news articles, we invited domain experts to annotate the news with MeSH classes and confronted those with the annotation of our classifier.

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, epidemics, childhood obesity, and children in care.

For the above combinations, we have calculated the average precision, recall, F1 and F0.5 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 the optimal similarity cut-off is around 0.2, providing precision 0.6 and F0.5=0.43. In the case of diabetes, the three depth 4 over cut-off values of approximately 0.37 show precision increase to over 60%, while below cut-off values of approximately 0.23 recall increases to over 50%. 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.

Taking into consideration the good 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).

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 results on the classification of news articles related to those health topics.

The specific challenges in the hand annotation of MEDLINE articles (where one can annotated 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 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 [11]). 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 to build from it a cross-lingual classifier. It can also contribute to the evaluation of the success of public health campaigns by allowing decision-makers to access what the news media's response 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* [15].

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 in: 20 March 2020.

[3] 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.*

[4] 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.

[5] Fang, Lei, George Karakiulakis, and Michael 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

[6] L. Henderson, Lachlan (2009). Automated text classification in the DMOZ hierarchy. TR.

[7] Kaggle (2020). COVID-19 Open Research Dataset Challenge - CORD-19
url: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge. Accessed in: 20 March 2020.

[8] 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.

[9] D. A. Lindberg (2000). Internet access to the National Library of Medicine. *Effective clinical practice: ECP*, *3*(5), 256.

[10] C. Manninget al (2008), "Introduction to Information Retrieval," Cambridge Univ. Press.

[11] 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.

[12] 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 in: 20 March 2020.

[13] 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.

[14] D. Mladenic and M. Grobelnik (2003). Feature selection on hierarchy of web documents. Journal: *Decision support systems*. vol. 35, pp. 45-87.

[15] J. Pita Costa et al (2019). Health News Bias and Epidemic Intelligence for Public Health. Proceedings of the SIKDD 2019.

[16] J. Pita Costa et al (2017). Text mining open datasets to support public health. In *Conf. Proceedings of* WITS 2017.

[17] 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.

[18] 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)*.

[19] P. Srinivasan and B. Libbus (2004). Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics*, *20*(Suppl 1), i290–i296.

[20] Y. Yan et al (2018). Biomedical literature classification with a CNNs-based hybrid learning network. *PloS one*, *13*(7), e0197933.

[21] 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.

[22] 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 in: 20 March 2020.

[23] 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.