Text mining MEDLINE to support public health

João Pita Costa, Luka Stopar, Flavio Fuart, Marko Grobelnik
Jožef Stefan Institute, Ljubljana
Quintelligence, Ljubljana, Slovenia

Raghu Santanam, Chenlu Sun
Arizona State University, USA

Paul Carlin
South Eastern Health and Social Care Trust, UK

Michaela Black, Jonathan Wallace
Ulster University, UK

ABSTRACT

Today's society is data rich and information driven, with access to numerous data sources available that have the potential to provide new insights into areas such as disease prevention, personalised medicine and data driven policy decisions. This paper describes and demonstrates the use of text mining tools developed to support public health institutions to complement their data with other accessible open data sources, optimize analysis and gain insight when examining policy. In particular we focus on the exploration of MEDLINE, the biggest structured open dataset of biomedical knowledge. In MEDLINE we utilize its terminology for indexing and cataloguing biomedical information – MeSH – to maximize the efficacy of the dataset.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

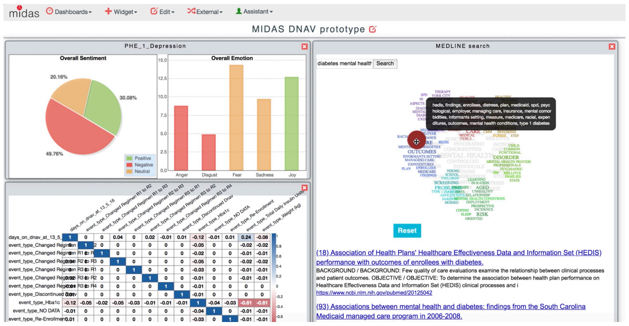
Measurement, Performance, Health.

Keywords

Big Data, Public Health, Healthcare, Text Mining, Machine Learning, MEDLINE, MeSH Headings.

1. MEANINGFUL BIG DATA TOOLS TO SUPPORT PUBLIC HEALTH

The Meaningful Integration of Data, Analytics and Service [MIDAS], Horizon 2020 (H2020) project [1] is developing a big data platform that facilitates the utilisation of healthcare data beyond existing isolated systems, making that data amenable to enrichment with open and social data. This solution aims to enable evidence-based health policy decision-making, leading to significant improvements in healthcare and quality of life for all citizens. Policy makers will have the capability to perform data-driven evaluations of the efficiency and effectiveness of proposed policies in terms of expenditure, delivery, wellbeing, and health and socio-economic inequalities, thus improving current policy risk stratification, formulation, implementation and evaluation. MIDAS enables the integration of heterogeneous data sources, provides privacy-preserving analytics, forecasting tools and visualisation modules of actionable information (see the dashboard of the first prototype in Figure 1). The integration of open data is fundamental to the participatory nature of the project and core ideology, that heterogeneity brings insight and value to analysis. This will democratize, to some extent, the contribution to the results of MIDAS. Moreover, it enables the MIDAS user to profit from the often powerful information that exists in these open datasets. A point in case is MEDLINE, the scientific biomedical knowledge base, made publicly available through PubMed. The set of tools described in this demonstration paper focuses on this large open dataset, and the exploration of its structured data.



	days_on_dnav_at_13_5_16	event_type_Changed Regimen R1 to R2	event_type_Changed Regimen R1 to R3	event_type_Changed Regimen R2 to R3	event_type_Changed Regimen R3 to R1	event_type_Changed Regimen R3 to R2	event_type_Changed Regimen R3 to R4	event_type_Discontinued Dnav	event_type_Hba1c	event_type_NO DATA	event_type_Re-Enrollment	event_type_Total Daily Insulin Units	event_type_Weight (kg)
days_on_dnav_at_13_5_	1	0	0	0.04	0	0.02	-0.01	0.01	-0.12	-0.01	0.01	0.24	-0.08
event_type_Changed Regimen R1 to R2	0	1	0	0	0	0	0	0	-0.05	0	0	-0.02	-0.02
event_type_Changed Regimen R1 to R3	0	0	1	0	0	0	0	0	-0.02	0	0	-0.01	-0.01
event_type_Changed Regimen R2 to R3	0.04	0	0	1	0	0	0	0	-0.05	0	0	-0.02	-0.02
event_type_Changed Regimen R3 to R1	0	0	0	0	1	0	0	0	-0.03	0	0	-0.01	-0.01
event_type_Changed Regimen R3 to R2	0.02	0	0	0	0	1	0	0	-0.01	0	0	-0.01	-0.01
event_type_Changed Regimen R3 to R4	-0.01	0	0	0	0	0	1	0	-0.01	0	0	0	0
event_type_Discontinued Dnav	0.01	0	0	0	0	0	0	1	-0.01	0	0	0	0
event_type_Hba1c	-0.12	-0.05	-0.02	-0.05	-0.03	-0.01	-0.01	-0.01	1	-0.04	-0.03	-0.61	-0.61
event_type_NO DATA	-0.01	0	0	0	0	0	0	0	-0.04	1	0	-0.01	-0.01
event_type_Re-Enrollment	0.01	0	0	0	0	0	0	0	-0.03	0	1	-0.01	-0.01

diabetes mental health

hedis, findings, enrollees, distress, plan, medicaid, spd, psychological, employer, managing care, insurance, mental comorbidities, informants setting, measure, medicare, racial, expenditures, outcomes, mental health conditions, type 1 diabetes

(18) Association of Health Plans' Healthcare Effectiveness Data and Information Set (HEDIS) performance with outcomes of enrollees with diabetes.

BACKGROUND / BACKGROUND: Few quality of care evaluations examine the relationship between clinical processes and patient outcomes. OBJECTIVE / OBJECTIVE: To determine the association between health plan performance on Healthcare Effectiveness Data and Information Set (HEDIS) clinical processes and i

https://www.ncbi.nlm.nih.gov/pubmed/20125042

(93) Associations between mental health and diabetes: findings from the South Carolina Medicaid managed care program in 2006-2008.

Figure 1. MIDAS platform dashboard, composed of visualisation modules customized to the public health data sourced in each governamental institution, and combined with open data.

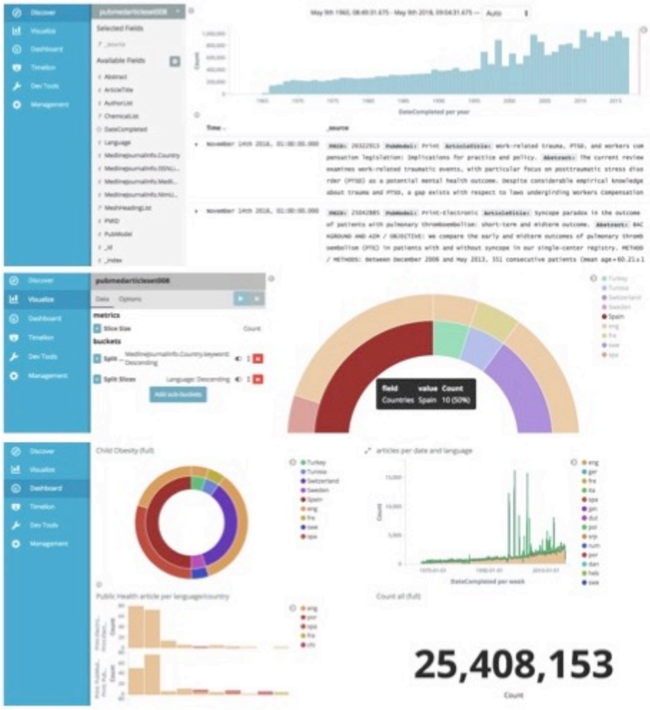
2. THE MEDLINE BIOMEDICAL OPEN DATA SET AND IT'S CHALLENGES.

2.1. MEDLINE DATASET.

With the accelerating use of big data, and the analytics and visualization of this information being used to positively affect the daily life of people worldwide, health professionals require more and more efficient and effective technologies to bring added value to the information outputs when planning and delivering care. The day-to-day growth of online knowledge requires that the high quality information sources are complete, high quality and accessible. A particular example of this is the PubMed system, which allows access to the state-of-the-art in 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 [7], maintained by the United States National Library of Medicine (NLM) at the National Institutes of Health (NIH). It includes citations from more than 5200 journals worldwide journals 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. About 500,000 new records are added each year. 17.2 million 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".

2.2. MEDLINE STRUCTURE.

The MEDLINE dataset includes a comprehensive controlled vocabulary – the *Medical Subject Headings* (MeSH) – that delivers a functional system of indexing journal articles and books in the life sciences. It has proven very useful in the search of specific topics in medical research, which is particularly useful for researchers conducting initial literature reviews before engaging in particular research tasks. Humans annotate most of the articles in MEDLINE with MeSH Heading 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 anatomical terms, diseases, drugs, etc) that further subdivide from the most general to the most specific in up to 13 hierarchical depth levels.



November 14th 2018, 01:00:00.000 PMID: 26322913 PubModel: Print ArticleTitle: Work-related trauma, PTSD, and workers compensation legislation: Implications for practice and policy. Abstract: The current review examines work-related traumatic events, with particular focus on posttraumatic stress disorder (PTSD) as a potential mental health outcome. Despite considerable empirical knowledge about trauma and PTSD, a gap exists with respect to laws undergirding workers compensation

November 14th 2018, 01:00:00.000 PMID: 25042885 PubModel: Print-Electronic ArticleTitle: Syncope paradox in the outcome of patients with pulmonary thromboembolism: short-term and midterm outcome. Abstract: BACKGROUND AND AIM / OBJECTIVE: We compare the early and midterm outcomes of pulmonary thromboembolism (PTE) in patients with and without syncope in our single-center registry. METHODS / METHODS: Between December 2006 and May 2013, 351 consecutive patients (mean age=60.21±1

Figure 2. *MEDLINE data visualisation tool enabling exploration of that open dataset in its full potential, based on data representations easy to understand and to communicate. It provides an interactive public instance that can be managed at the dashboard management tool (below) for which the visualisation modules are constructed (in the center) based on the queries made to the MEDLINE dataset (above).*

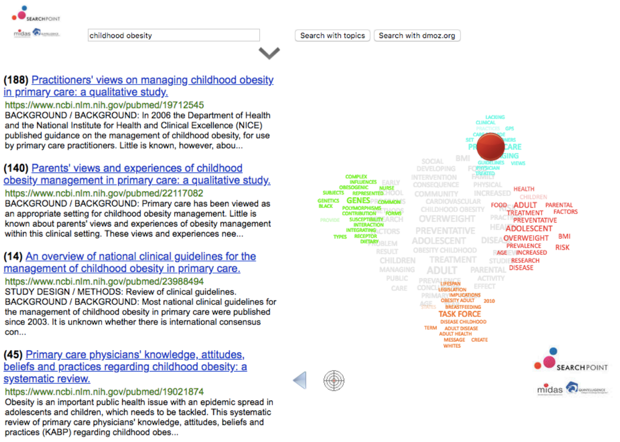
2.3. MEDLINE INDEX.

This paper demonstrates the interactive data visualisation text-mining tools that enable the user to extract meaningful information from MEDLINE. To do that we are using the underlying ontology-like structure MeSH. MEDLINE data, together with the MeSH annotation, that is indexed with ElasticSearch and made available to data analytics and visualisation tools. This will be discussed in more detail in the next section.

The manipulation and visualization of such a complete data source brings challenges, particularly in the efficient search, review and presentation when choosing appropriate scientific knowledge. The manipulation and visualisation of complex text data is an important step in extracting meaningful information from a dataset such as MEDLINE. Although powerful, the online search engine provided by the NLM does not provide suitable tools for in-depth analysis and the emergence of scientific information. As one of the main goals of MIDAS is to experiment with advanced visualisation techniques in support of public health policy making, a suitable MIDAS PubMed repository had to be developed. This repository has to allow exploration of a wide range of different visualisation techniques in order to evaluate their applicability to policy-making tasks within the policy cycle. Therefore, there was a need for a selection of a powerful, semi-structured text index, that would allow free text searches, but also allow the creation of complex queries based on available metadata. An obvious choice is elasticSearch, which combines features provided by NoSQL databases with standard full text indexes, as it is based on the Apache Lucene Index. The main design challenge when choosing this particular toolset was that querying based on arrays or parent-child relations are not supported, meaning that for complex use-cases different indexes based on the same source dataset have to be created. Nevertheless, excellent results, particularly with regards to the area of performance have been obtained.

2.4. MEDLINE DASHBOARD.

One of the identified needs motivating this work is assuring the availability of a dynamic dashboard that permits the user to explore data visualisation modules, representing the queries to the MEDLINE dataset through pie charts, bar charts, etc [5]. The dashboard that we made available (in Figure 2) feeds on that dataset through the elasticSearch index earlier discussed. It is composed of several interactive visualisation modules that utilises the mouse hover when interacting and provide information through pop-up messages on several aspects of the data based on particular queries of interest (e.g. a pie chart representing the "public health" citations that talk about "childhood obesity" during a selected period of time; or a bar chart showing different concepts included in the articles related to "mental health" in Finnish scientific journals). The MEDLINE dataset is mostly in the English language but includes a significant volume of translated abstracts of scientific articles that were written in several other languages. The open source data visualisation Kibana is a plugin to elasticSearch that supports the described dashboard. Thus, it is the data visualisation dashboard of choice for elasticSearch-based indexes, such as the one we present here. It is used in the context of MIDAS for fast prototyping and support of part of MIDAS use-cases. While the dashboard itself serves the less technical user to explore the data available (over a subset of the data generated by a topic of interest), other options are available that permit more control of the data by the data scientists at a more operational level. These are: (i) the management dashboard, where the technical user can perform the appropriate subsampling based on the topics of interest and the required advanced options over the available data features; and (ii) the visual modules creator permitting the technical user to easily create new interactive visualisation modules. Moreover, this tool enables one to query large datasets and produce different types of visualisation modules that can be later integrated into customized dashboards. The flexibility of such dashboards permits the user to profit from data visualisations that feed on his/her preferences, previously set up as filters to the dataset. The MIDAS data visualisation tools permit the user to explore the potential of the MEDLINE dataset, based on pie charts and other representations that are easy to comprehend, interact with, and to communicate. It also enables a public instance based on a particular query to the dataset, which includes different types of data visualisation modules that can later integrate a customised dashboard, designed in agreement with the workflows and preferences of the end-user. This live dashboard can easily be integrated through an iframe in any website, not showing the customization settings but maintaining the interaction capability and the real-time update. *It permits a complete base solution to further explorer the MEDLNE index and the associated dataset [6].*



(188) Practitioners' views on managing childhood obesity in primary care: a qualitative study.
https://www.ncbi.nlm.nih.gov/pubmed/19712545
BACKGROUND / BACKGROUND: In 2006 the Department of Health and the National Institute for Health and Clinical Excellence (NICE) published guidance on the management of childhood obesity, for use by primary care practitioners. Little is known, however, abou...

(140) Parents' views and experiences of childhood obesity management in primary care: a qualitative study.
https://www.ncbi.nlm.nih.gov/pubmed/22117082
BACKGROUND / BACKGROUND: Primary care has been viewed as an appropriate setting for childhood obesity management. Little is known about parents' views and experiences of obesity management within this clinical setting. These views and experiences nee...

(14) An overview of national clinical guidelines for the management of childhood obesity in primary care.
https://www.ncbi.nlm.nih.gov/pubmed/23988494
STUDY DESIGN / METHODS: Review of clinical guidelines. BACKGROUND / BACKGROUND: Most national clinical guidelines for the management of childhood obesity in primary care were published since 2003. It is unknown whether there is international consensus con...

(45) Primary care physicians' knowledge, attitudes, beliefs and practices regarding childhood obesity: a systematic review.
https://www.ncbi.nlm.nih.gov/pubmed/19021874
Obesity is an important public health issue with an epidemic spread in adolescents and children, which needs to be tackled. This systematic review of primary care physicians' knowledge, attitudes, beliefs and practices (KABP) regarding childhood obes...

Figure 4. A screenshot of the MEDLINE SearchPoint, with groups of keywords (on the right) extracted from the search results, represented by different colors, and on the left the reindexed search results themselves with the number that they appear in the original index *[6]*.

4. MEDLINE SEARCHPOINT.

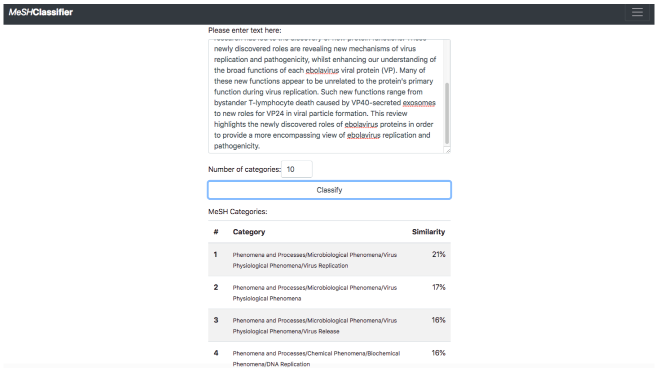
The efficient visualisation of complex data is today an important step in obtaining the research questions that describe the problem that is extracted from that data. The MEDLINE SearchPoint is an interactive exploratory tool refocused from the proprietary open source technology *SearchPoint* [8] (available at searchpoint.ijs.si) to support health professionals in the search for appropriate biomedical knowledge. It exhibits the clustered keywords of a query, after searching for a topic. When we use indexing services (such as standard search engines) to search for information across a huge amount of text documents – the MEDLINE index described in Section 2 being an example – we usually receive the answer as a list sorted by a relevance criteria defined by the search engine. The answer we get is biased by definition. Even by refining the query further, a time consuming process, we can never be confident about the quality of the result. This interactive visual tool helps us in identifying the information we are looking for by reordering the positioning of the search results according to subtopics extracted from the results of the original search by the user. For example, when we enter a search term ‘childhood obesity, the system performs an elasticSearch search over the MEDLINE dataset, extracts groups of keywords that best describe different subgroups of results (these are most relevant, and not most frequent terms). This tool gives us an overview of the content of the retrieved documents (e.g. we see groups of results about prevention, pregnancy, treatments, etc.) represented by: (i) a numbered list of 10 MEDLINE articles with a short description extracted from the first part of the abstract; (ii) a word-cloud representing the k-means clusters of topics in the articles that include the searched keywords; (iii) a pointer that can be moved through the word-cloud and that will change the priority of the listed articles. The word-cloud in (ii) is done by taking a set of MEDLINE documents S and transforming them into vectors using TF-IDF, where each dimension represents the "frequency" of one particular word. For example, lets say that we have document D_1: "psoriasis is bad" and document D_2: "psoriasis is good". This could be transformed as $D_1 = (1, 1, 1, 0)$ and $D_2 = (1, 1, 0, 1)$. Then the documents are clustered into k groups $G_1, G_2, \ldots, G_k$ using the k-means algorithm. For each group we compute the "average" document (centroid), which is the representative of the group. The most frequent words in the "average" document are drawn in the word cloud - the central grey word cloud is the "average" of all the documents in S. We can calculate how similar a particular document d is to a group G_i by calculating the cosine of the angle between the vector representation of d and the "average" document (centroid) of the group G_i. By dragging the red SearchPoint ball over the word-groups, we provide the relevance criteria to the search result, thus bringing to the top results the articles we are most interested in (see Figure 4). When that ball is moved, for each document, we calculate the similarity to each of the word-groups and combine it with the distance between the ball and the group. The result is used as the ranking weight where the document with the highest cumulative weight is ranked first. When having the mouse over the word-clouds we get a combination of the most frequent words shown in the tag clouds that change based on how close the ball is to a particular group. After getting to a position with the SearchPoint over the word cloud highlighting “primary care”, a qualitative study in primary care on childhood obesity that occupied the position 188 is now in the first position. The user can read its title and first lines of abstract, and when clicking on it, the system opens the article in the browser at its PubMed URL location.

3. MeSH CLASSIFIER

This rich data structure in the MEDLINE open set is annotated by human hand (although assisted by semi-automated NIH tools) and therefore is not available in the most recent citations. However, in the context of MIDAS we made available an automated classifier based on [2] that is able to suggest the categories of any health related free text. It learns over the part of the MEDLINE dataset that is already annotated with MeSH, and is be able to suggest categories to the submitted text snippets. These snippets can be abstracts that do not yet include MESH classification, medical summary records or even health related news articles. To do that we use a nearest centroid classifier [3] constructed from the abstracts from the MEDLINE dataset and their associated MeSH headings. Each document is embedded in a vector space as a feature vector of TF-IDF weights. 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, 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 as the cosine of the angle between the embeddings. Preliminary analysis shows promising results. For instance when classifying the first paragraph of the Wikipedia page for “childhood obesity”, excluding the keyword “childhood obesity” from the text, the classifier returns the following MeSH headings:

> *"Diseases/Pathological Conditions, Signs and Symptoms/Signs and Symptoms/Body Weight/Overweight",*
> *"Diseases/Pathological Conditions, Signs and Symptoms/Signs and Symptoms/Body Weight/Overweight/Obesity".*

The demonstrator version of the MeSH classifier is already available through a web app, as well as through a REST API using a POST call, and is at the moment under qualitative evaluation. This is being done together with health professionals with years of practical experience in using MeSH themselves through PubMed. In addition, we aim to further explore the potential of the developed classifier in several public health related contexts including non classified scientific articles of three types: (i) review articles; (ii) clinical studies; and (iii) standard medical articles. The potential impact of this technology will also include electronic health records and the monitoring health related news sources. We believe that his approach will address an identified recurrent need of health departments to enhance the biomedical knowledge, and motivate a step change in health monitoring.



Please enter text here:

newly discovered roles are revealing new mechanisms of virus replication and pathogenicity, whilst enhancing our understanding of the broad functions of each ebolavirus viral protein (VP). Many of these new functions appear to be unrelated to the protein's primary function during virus replication. Such new functions range from bystander T-lymphocyte death caused by VP40-secreted exosomes to new roles for VP24 in viral particle formation. This review highlights the newly discovered roles of ebolavirus proteins in order to provide a more encompassing view of ebolavirus replication and pathogenicity.

Number of categories: 10

MeSH Categories:

#	Category	Similarity
1	Phenomena and Processes/Microbiological Phenomena/Virus Physiological Phenomena/Virus Replication	21%
2	Phenomena and Processes/Microbiological Phenomena/Virus Physiological Phenomena	17%
3	Phenomena and Processes/Microbiological Phenomena/Virus Physiological Phenomena/Virus Release	16%
4	Phenomena and Processes/Chemical Phenomena/Biochemical Phenomena/DNA Replication	16%

Figure 3. *A screenshot of the web app to the MEDLINE classifier, when requesting the automated MeSH annotation of a scientific review article abstract extracted from PubMed (in the body of text above) and the results as MeSH headings descriptors including their tree path in the MeSH ontology-like structure (center), their similarity measure (right) and their positioning in the classification (left).*

5. CONCLUSION AND FUTURE WORK

To further extend the usability of the MEDLINE SearchPoint, we are developing a data visualisation tool that permits the user to plot the top results mostly related with a topic of interest, as explored with SearchPoint. Based on the choice of a time window and a certain topic, such as “mental health”, the user is able to view the clustered MEDLINE documents, rolled over the plot or click to view the plotted points, each of which represents an article in PubMed. This will be done through multidimensional scaling, plotting the articles in the subsample using cosine text similarity. The difficulties to plot large datasets using these methods, and the lack of potential in the outcomes of that heavy computation, provided a focus for the team to only plot the first hundred results of the explorations done within MEDLINE SearchPoint. With this extended tool the healthcare professional will be able to: (i) explore a certain area of research with the aim of a more accessible scientific review, in identifying the evidence base for a medical study or a diagnostic decision; (ii) identify areas of dense scientific research corresponding to searchable topics (e.g. the evaluation of the coverage of certain rare diseases that need more biomedical research, or the identification of alternative research paths to overpopulated but inefficient research); and (iii) exploration of the research topic over time windows that enable filtering to avoid known unreliable results.

In line with this work we have been developing research to contribute to the smart automation of the production of biomedical review articles. This collaborative research lead by Raghu T. Santanam at Arizona State University, aims to provide a wide knowledge over a restricted topic over the wider knowledge available at MEDLINE. We utilize the deep learning algorithm Doc2vec [4] to create similarity measures between articles in our corpus. In that we built a balanced test dataset and three different representations of the corpus, and compared the performance between them. The implementation currently builds a matrix of similarity scores for each article in the corpus. In the next steps, we will compare similarity of documents from our implementation against the baseline for a randomly chosen set of articles in the corpus.

The further development of the MeSH classifier will consider the feedback of the usability of health professionals working in partner institutions, profiting of their years of experience with MEDLINE and MeSH itself, to tune the system to ensure the best usability in the domain. Furthermore, we will use the outcomes of the final version of this classifier to label health related news with the MeSH Headings descriptors, potentiating a new approach on the processing and monitoring of population health, population awareness of certain diseases, and the general public acceptance of public health decisions through news.

ACKNOWLEDGMENTS

We thank the support of the European Commission on the H2020 project MIDAS (G.A. nr. 727721).

REFERENCES

[1] B. Cleland et al (2018). Insights into Antidepressant Prescribing Using Open Health Data, Big Data Research, doi.org/10.1016/j.bdr.2018.02.002
[2] L. Henderson, Lachlan (2009). Automated text classification in the dmoz hierarchy. TR.
[3] C. Manninget al (2008), “Introduction to Information Retrieval,” Cambridge Univ. Press, 2008, pp. 269-273.
[4] T. Mikolov et al (2013). Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781.
[5] J. Pita Costa et al (2017). Text mining open datasets to support public health. In *Conf. Proceedings of* WITS 2017.
[6] J. Pita Costa et al (2018). MIDAS MEDLINE Toolset Demo. http://midas.quintelligence.com (accessed in 28-8-2018).
[7] F. B. Rogers, (1963). Medical subject headings. *Bull Med Libr Assoc*. **51**: 114–6.
[8] L. Stopar, B. Fortuna and M. Grobelnik (2012). Newssearch: Search and dynamic re-ranking over news corpora. In Conf. Proceedings of SiKDD2012.