Wolff *et al. Journal of Biomedical Semantics* (2020) 11:12
https://doi.org/10.1186/s13326-020-00226-w

Journal of Biomedical Semantics

RESEARCH Open Access

Methodologically grounded semantic analysis of large volume of chilean medical literature data applied to the analysis of medical research funding efficiency in Chile

Patricio Wolff[1], Sebastián Ríos[1], David Clavijo[1], Manuel Graña[2*] and Miguel Carrasco[3]

Abstract

Background: Medical knowledge is accumulated in scientific research papers along time. In order to exploit this knowledge by automated systems, there is a growing interest in developing text mining methodologies to extract, structure, and analyze in the shortest time possible the knowledge encoded in the large volume of medical literature. In this paper, we use the Latent Dirichlet Allocation approach to analyze the correlation between funding efforts and actually published research results in order to provide the policy makers with a systematic and rigorous tool to assess the efficiency of funding programs in the medical area.

Results: We have tested our methodology in the Revista Médica de Chile, years 2012-2015. 50 relevant semantic topics were identified within 643 medical scientific research papers. Relationships between the identified semantic topics were uncovered using visualization methods. We have also been able to analyze the funding patterns of scientific research underlying these publications. We found that only 29% of the publications declare funding sources, and we identified five topic clusters that concentrate 86% of the declared funds.

Conclusions: Our methodology allows analyzing and interpreting the current state of medical research at a national level. The funding source analysis may be useful at the policy making level in order to assess the impact of actual funding policies, and to design new policies.

Keywords: Data science, Machine learning, Latent Dirichlet allocation, Healthcare management, Strategy

Background

Due to the speed of innovation and change of research trends in the medical community, research topic taxonomies published by governmental agencies for funding calls often diverge from the reality of the research practice. Our working hypothesis is that semantic topic analysis provides an unbiased and accurate portrait of the actual research topics that are generating published results. In this paper we exploit the information from a national medical publication, described below, to identify the areas of active research, correlating them with the acknowledged funding sources, and non-funded personal effort backing these scientific results. This analysis provides the policymaker with a systematic, unbiased, and automated tool for the evaluation of the results of funding programs, allowing to assess the coherence of the national research funding policies with the actual research outcomes.

Methodology background

The growth of number of PubMed references (from 363 in 2009 to 1820 in 2019 in a search with the terms "biomedical literature analysis") demonstrates the

*Correspondence: manuel.grana@ehu.es
[2]Computational Intelligence Group, University of Basque Country, P. Manuel Lardizabal 1, 20018 San Sebastián, Spain
Full list of author information is available at the end of the article

increasing research efforts devoted to the automatic process of the biomedical literature, starting with the clustering of documents dealing with the same issues [1] based on the identification of semantic relevant terms, which can either be defined by some preexisting ontology or provided by a human expert. Document clustering is fundamental for online indexing of biomedical publications allowing easy search based on supervised machine learning approaches [2] in order to decrease human processing costs and increase availability of semantically indexed documents to the research community. Natural language processing techniques are useful for the extraction of information for further processing, such as ranking of terms in order to discover new concepts like phenotypic disease characterization [3]. Recent works use deep learning techniques to anchor a specific semantic ontology in the relevant literature [4]. A very promising application of medical literature is the discovery of new relations between concepts that may lead to breakthrough treatments [5].

The definition of the semantic domain is the first step in any attempt to automatic biomedical literature processing. The identification of topics for document semantic indexing can be done by humans that carry out the manual annotation of documents. Another approach is the so called topic modelling, i.e. the automated induction of semantic topics from the document data, under the assumption that these topics are defined in a latent space which can be uncovered by analytical means. Topic modeling alleviates the cost in human resources and time of the semantic domain definition, but the discovered topics are not guided by any human medical expert meaning, hence they require post-hoc human validation and interpretation. Latent Dirichlet Allocation (LDA) is the foremost topic modeling approach. It has been applied on different types of documents and their corresponding knowledge disciplines (regardless of the format in which the information is found as longs as it is text), such as work place and personal e-mails, abstracts in scientific documents and newspapers [6]. LDA has allowed pattern discovery in words and documents in the medical field, where it has been used to link diagnostic groups, medicines and publications [7–16]. After the topic modelling achieved by LDA and post-hoc analysis of the discovered topics, some meta-analysis can be carried out over the topic segmentation of the semantic domain. In our study, we perform a descriptive statistical analysis of the declared funding sources, which allows to assess the impact of funding agencies in the research actually reported in the literature.

Case study background: medical scientific research in Chile

We showcase our approach on the analysis of medical scientific production in Chile, using as the main information source for this task the *Revista Médica de Chile* (RevMed). RevMed is a national and international reference in terms of dissemination of knowledge in the medical area. It was founded in 1872 as a result of the creation of the *Sociedad Médica* (Medical Society) in 1869. It's the third oldest periodical publication of Chile, it's the oldest medical journal in South America and second oldest in Spanish language in the world [17], and still continues to be relevant.

RevMed has covered in depth the technological and knowledge progress in each of the main medical research areas, such as clinical research, public health, ethics, medical education, and medical history. Consequently, it has a very important role in educational tasks, learning and scientific knowledge development in the country. For these reasons it is a faithful record of medical research in Chile.

Materials and methods

Corpus

Our corpus is composed of 643 research papers from RevMed published between the years 2012 and 2015. Prior to 2012, there is no access to online documents that could be used in our experiments. Besides, our processing capabilities allowed us to process only until year 2015. The categories of the papers included in the corpus were: Research Papers, Review Papers, Special Paper and Clinical Cases. Public Health articles and Letters to the Editor sections were excluded. The distribution of the number of papers *per* issue is shown in Table 1.

Methodological steps

Figure 1 visualizes the methodological steps followed by our research funding analysis based on the semantic analysis of medical literature. The steps of the methodology are the following:

1 Publications: We carried out the textual data preprocessing, which consisted in cleaning and

Table 1 Monthly and annual distribution of the Research Articles downloaded from RevMed (2012-2015)

	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec	total
2012	16	15	16	15	17	15	14	12	10	16	13	11	**170**
2013	14	15	16	12	13	13	12	13	14	14	14	15	**165**
2014	13	15	13	14	14	13	8	13	14	13	14	14	**158**
2015	13	12	11	13	14	13	14	14	12	12	12	10	**150**
total	**56**	**57**	**56**	**54**	**58**	**54**	**48**	**52**	**50**	**55**	**53**	**50**	**643**

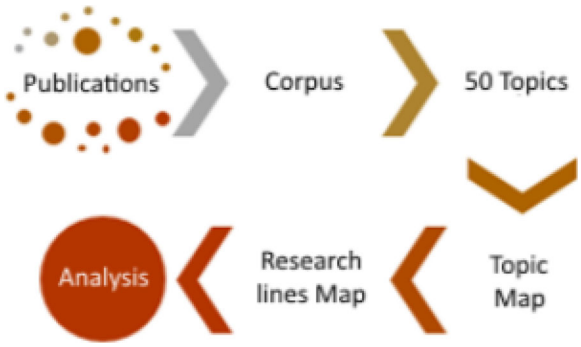

Fig. 1 Methodological steps for the analysis of medical research funding from the literature

removing less relevant content in order to select meaningful words in the medical context that allow to provide interpretations to the topics identified by LDA.

(a) Every character of the articles was converted to lowercase.
(b) Non-alpha numeric characters were removed.
(c) Double spaces between words were removed.
(d) Numbers were removed. Even if some entity names may contain numbers, this has no impact on our methodology as far as we are looking for high level semantic topics.
(e) Words or terms irrelevant for our analysis were removed, such as prepositions, conjunctions, articles, etc. Additionally, terms with very high frequency were removed (rev, med, Chile, etc).

Preprocessing of documents was done on the programming language R (version 3.3.0), using the tm library (version 0.6-2) [18, 19], which provides a very convenient text mining framework.

2 Corpus: We extract the latent topics from the preprocessed corpus through LDA probabilistic modeling using the R package topicmodels (version 0.3-2)[20]. We use three different metrics, i.e. Griffiths2004 [14]; CaoJuan2009 [21]; and Arun2010 [22], to evaluate the quality of the topic modeling in order to determine the optimal number of topics for the ensuing semantic analysis. Each topic is represented by its 30 more meaningful words.
3 Topics: We give an interpretation and name to each of the LDA identified topics using the topic visualization tool LDAvis (version 0.3.2) [23][1] running in R, D3,[2] and the qualitative analysis of a team of medical experts. The team was composed of three medical staff from the local hospital with extensive experience publishing in the journal, and research experience to carry out the topic semantic identification. Funding sources were not considered in this process. Additionally, we achieve an information size reduction going from the number of papers to the number of topics.
4 Topic Map: Using the visualization tool LDAvis, we create a 2D map of the LDA identified topics, where we were able to identify groups of topics by the judgment of experts, achieving a further information dimensionality reduction from the number of topics to the number of groups groups. Additionally, the axes of the topic map were interpreted accordingly to the topic grouping.
5 Research lines map: Publications or research papers where assigned to each of the research lines to observe its scientific production.
6 Funding Analysis: We use the funding acknowledgements in each paper to compute statistics of funding *per* research topic and research line.

[1] https://pyldavis.readthedocs.io
[2] https://d3js.org

Latent Dirichlet allocation

Relevant definitions

- A document is a sequence of N words denoted by $w = (w_1, w_2, \ldots, w_N)$, where w_n is the $n - th$ word of the sequence
- A corpus is the collection of M documents denoted by $D = \{d_1, d_2, \ldots, d_M\}$

Latent Dirichlet Allocation (LDA) is a generative probabilistic model of a corpus, where every document of the corpus is represented as a mixture of latent topics, and each topic is characterized by a probability distribution of words [24–26]. Specifically, in LDA this probability distribution is a Dirichlet distribution [27].

The model is represented with plate notation in Fig. 2. The parameters α and β are the priors of the Dirichlet distributions of topics per document and word per document, respectively. The inner and outer rectangular plates represent the word positions in a document and the documents, respectively. Each word position is associated with a topic choice $z_{ij} \in \{0, 1\}$. Each document d_i is described by a distribution of topics θ_i. Additionally, each topic k is modelled by a distribution of words φ_k, where we have a total of K topics. Equation (1) summarizes the model.

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)\, p(w_n|z_n, \beta) \quad (1)$$

The generative process underlying LDA for each document of the corpus is as follows:

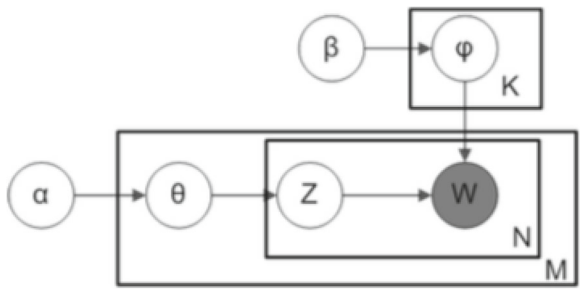

Fig. 2 LDA's model graphic representation

1 Choose a N $\sim$ Poisson(ξ) :
2 Choose a $\theta \sim$ Dir (α)
3 For each of the words w_n of N:

(a) Choose a topic $z_n \sim$ Multinomial (θ)
(b) Choose a word w_n according to $\varphi\,(w_n|z_n, \beta)$, a conditional probability on the topic z_n.

Where N follows a Poisson distribution with mean ξ, and θ and φ follow a Dirichlet distribution with parameters α and β, respectively .

Topic model visualization systems

There are many systems and applications for the visualization of the results of topic modeling, (Termite [28], MALLET[3] [29], ThemeRiver [30], and FacetAtlas [31] to name a few). Most of them try to link documents, topics and words for deeper post-hoc analysis of the obtained topic modelling results. The representations used by these systems are several: a list of words that belong to topics, limited bar graphs associated with the frequency of these words, clouds of relevant words describing a topic, pie charts representing the probability of each topic in a document, and many others.

For our work we have selected LDAvis [23], a visualization tools that allows quick and easy understanding of the modeling results. It carries out multidimensional scale analysis, achieving a distribution in a bidimensional space of the topics each represented by a circle. The size of a topic circle represents the relevance of the topic within the entire corpus, and each topic is associated to a list of relevant words describing it. The distance in the bidimensional projection space between the circle centers is a measure of the similarity of the topics: more similar topics have their circles placed at shorter distances. This tool allows to describe the meaning of each topic; to determine the prevalence of each topic in the corpus, and to infer the similarity link between each of the obtained topics.

Results

Topic modeling implementation

To estimate the optimal quantity of topics, we explored the results of topic modeling on the processed corpus carrying out a grid search over the number of topics from 5 up to 55. Figure 3 shows the results of this exploration. The optimal number of topics corresponds to the minimal values of Griffiths2004 and CaoJuan2009 metrics, and the maximal value of the Arun2010 metric. According to the plots in Fig. 3, it can be inferred that the optimal number of topics is K=50. After finding out the optimal number of topics, we apply two topic modeling approaches: Latent semantic analysis (LSA) and LDA. We apply Gibbs's sampling [32] to estimate the parameters and inference. We used the gensim Python libraries,[4] and the R implementations of LDA and LDAvis [23] for our purposes.

Comparing topic modeling results of the LSA and LDA approaches, LDA achieved better results according to the distribution of the number of documents *per* topic shown in Fig. 4. One of LSA topics accumulated 42% of the total research articles. Such concentration hinders the analysis and doesn't allow to make meaningful interpretations. For this reason, we selected LDA results for deeper analysis and interpretation.

For LDA, the number Gibbs's sampling scans over the whole corpus was set as 5000. The scalar value of Dirichlet distribution hyperparameter for the word distribution *per* topic, and for the topic distribution *per* document was set to 0.02.

LDA results visualization using LDAvis places topics discovered by LDA in a 2D space spanned by the principal components found by multidimensional scaling. In this visualization the Euclidean distance between the centers of the topic circle representation is a measure of the similarity between topics. This visualization also allows a quick inspection of the association between words and topics for a qualitative assessment of each topic meaning. The visualization of the topic spatial distribution can be observed on Fig. 5. Each topic is represented by a circle whose center is determined a multidimensional scaling process [33, 34] computed over the distances between topic word distributions. The prevalence of each topic is visualized via the proportional size of the circle diameters. The axes of the bidimensional map are constructed from the main components that come from the multidimensional scaling reduction of dimensionality process.

Topic modeling term results

The corpus vocabulary is composed of 12,328 different terms, after removing repetitions, we got a total of 307883

[3] https://wp.nyu.edu/exceltextanalysis/visualize-mallet-topics/

[4] https://radimrehurek.com/gensim/

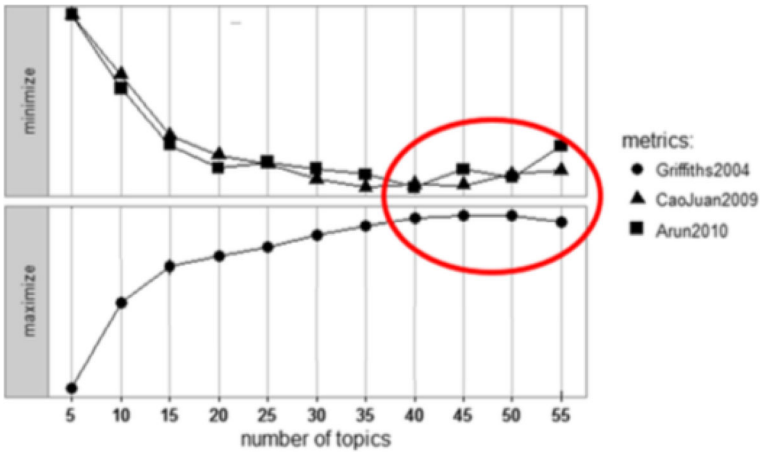

Fig. 3 Plot of the metrics for the identification of the optimal number of topics for the analysis. The upper part corresponds to the metrics that are minimized, the lower part for the metric that is maximized

unique terms. We discarded terms with less than 5 repetitions. The most frequent terms of the entire corpus can be observed in Fig. 6. It can be observed that terms such as treatment, cancer, woman and cell, are found more than 2,000 times in the corpus

Expert post-hoc analysis of the results

The judgment of experts was used to give a semantic interpretation clustering the 50 topics into groups with similar semantics. The result of this grouping was 11 clusters of topics, each one identified with its name as shown on Fig. 7. The group name identifies its general medical area, according to the meaning of the common terms found in the aggregated topics. It is possible to find interesting topics in the bidimensional space that are far away from the others, such as topic 32, which is located on the lower left side of the map (encompassing 1.5% of corpus terms) but that nevertheless semantically belong to a group. This topic is too specific when taking into account its representative terms, but they allow to aggregate it into the oncology group because it is related to cancer genetic studies. The oncology cluster is elongated in the representation space, due to the specificity of its belonging topics. Some clusters are single topics, like topic 27 containing genetics research. Topic 4 located in the upper right zone (encompassing 3.8% of corpus terms), almost defines a single topic cluster specifically devoted to physical activity and healthy life research.

The grouping of the topics allows us to give a semantic interpretation to the bidimensional space in which the topics are located. The horizontal axis corresponds to the size of the population under study in the reported medical research, ranging from research on individual subjects up to the entire population (going from the left to the right of the map). The vertical axis corresponds to the stage of the research according to the management of the disease, ranging from evaluating hypotheses about its cause and describing diagnostic instruments up to its prognosis or management (going from the bottom to the top of the

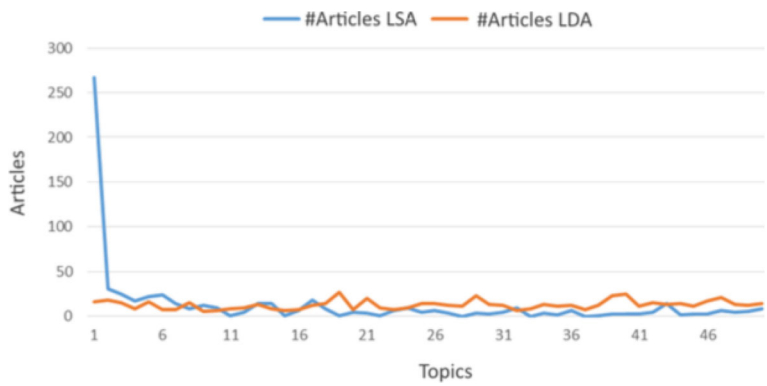

Fig. 4 Distribution of the number of documents *per* topic for LDA and LSA algorithms

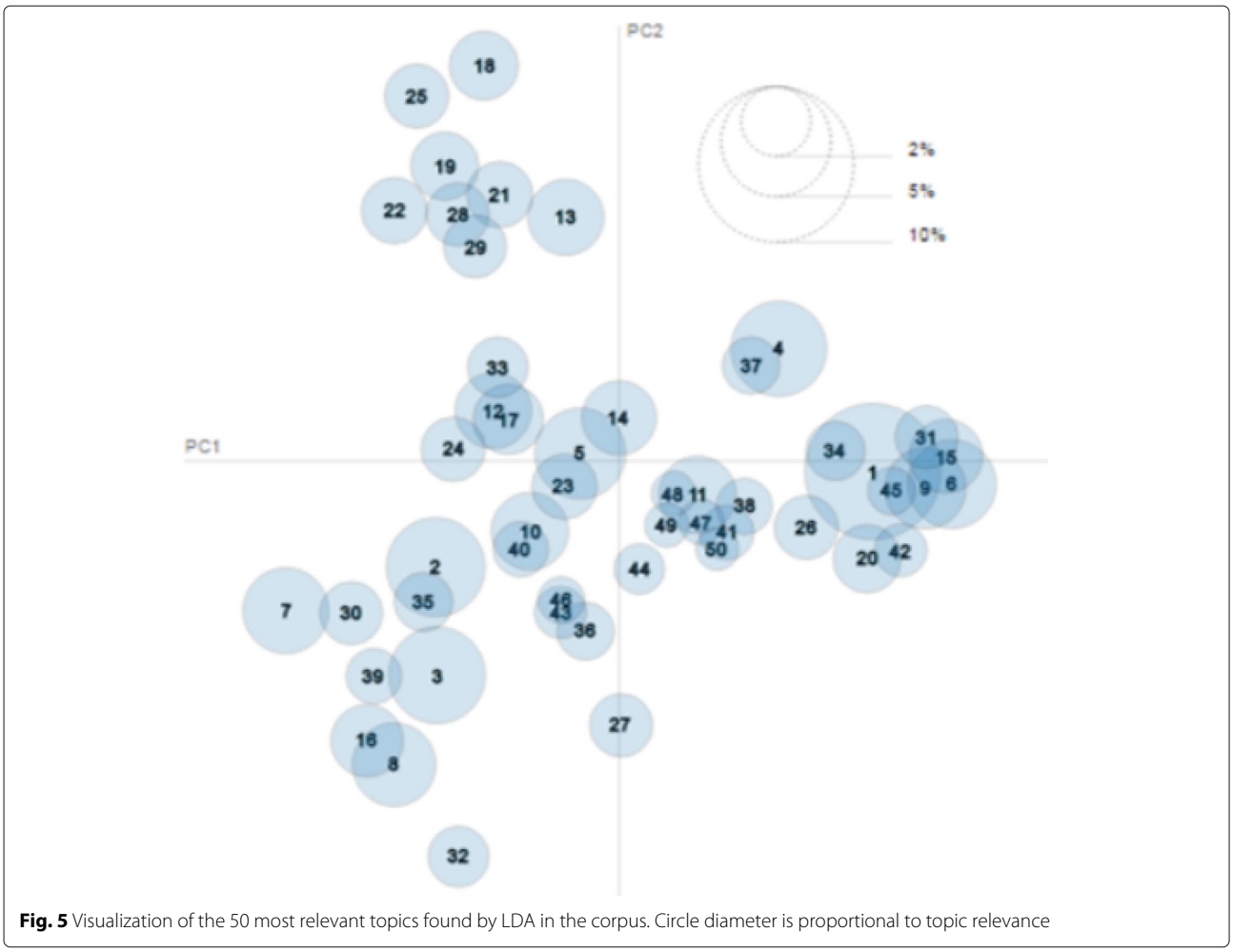

Fig. 5 Visualization of the 50 most relevant topics found by LDA in the corpus. Circle diameter is proportional to topic relevance

map). The map origin is the location of studies associated with treatments and patient care.

Topic modeling application to the analysis of research funding sources

The papers in the corpus describe research works that were developed with or without direct funding. Funding sources could be public, private or mixed. We found that only 29% of the researches published in RevMed in the period 2012-2015 declared direct funding sources, where the nature of the declared funding source is distributed as follows: 62% public, 32% private , and 6% mixed public and private. Some groups of topics concentrate the highest amount of funds: 87% of the funds were distributed into five groups: Treatment (22%), Oncology (20%), Inpatient research (18%), Population Studies (16%) and in Women's Health and Pregnancy (11%). These groups cover a large fraction (over 85%) of the published papers in this period.

We visualize the funding sources in the topic distribution bidimensional space, as shown in Fig. 8. Each pie represents the funding source distribution in each group of topics. The size of each circle represents the quantity of research works reported in each group of topics.

Discussion

Regarding the techniques used

The document analysis methodology presented in this paper is structured in a sequence of stages which achieve a progressive dimensionality reduction of the data arriving to a visual representation that allows diverse semantic-drive analysis by experts. We use well known techniques and software tools allowing the analysis of large volumes of data with off-the-shelf computing resources.

Regarding the number of topics

We have carried out a systematic search for the optimal number of topics in the sense of three metrics that are well known in the literature. The search was carried out incrementally. In our computational experience, the effect of small increments of the number of topics is difficult to

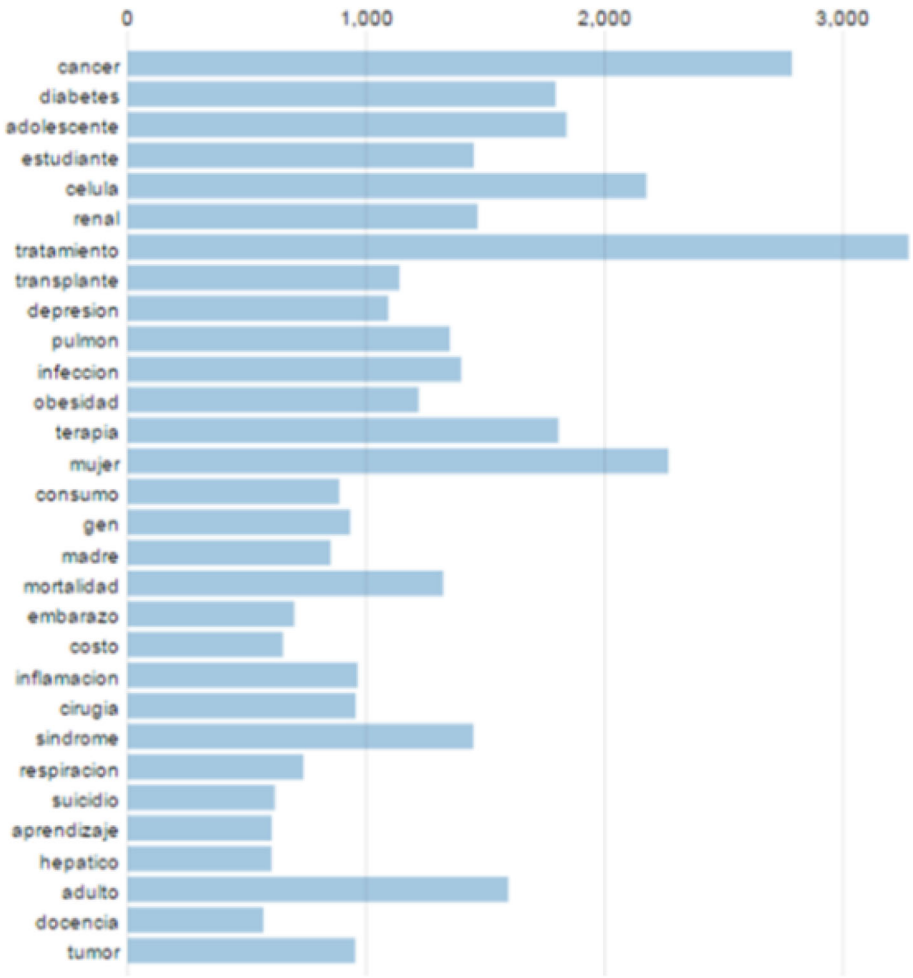

Fig. 6 The 30 most frequent terms of the corpus

ascertain, hence we have carried out jumps of size 5 in the exploration. The determination of the optimal number is carried out detecting when the three measures start to degrade, which in our study happens after 50 topics. Though it may be argued that a more fine exploration would be advisable, we note that the ensuing semantic interpretation gives the meaning to the topics and determines their usefulness for the desired analysis. Each topic is described in such a specific manner by its representative words that they were easy to identify by experts. Out of the 50 topics, medical experts found difficulties in giving a semantic interpretation only for one of the topics (topic 48). Moreover, we are looking for big research categories, hence we want to avoid over-segmentation of the semantic space into small categories that would clog the analysis.

Regarding processing time

The data processing time, from the topic modeling algorithm application up to the visualization of topics in the bidimensional space, was approximately 22 minutes using a laptop with an Intel Core i7 processor. Though this processing time is affordable for the considered corpus, scalating the methodology to extremely big corpuses like PubMed would require much bigger computing resources and some rewriting of the code to allow for parallel execution of several threads. One of the time critical tasks is the preprocessing of the corpus, which can be easily formulated as a trivially parallel task.

Regarding the funding efficiency

A very salient conclusion is that most of the research in the medical area that achieves publication is done by researchers not involved in directly funded projects, overall only 29% of the papers report funding sources). This percentage is quite homogeneous, so it seems that funding is not a driver of published research. We report the relative percentages of the events 'funded' versus 'non-funded', but we do not have at this time information about

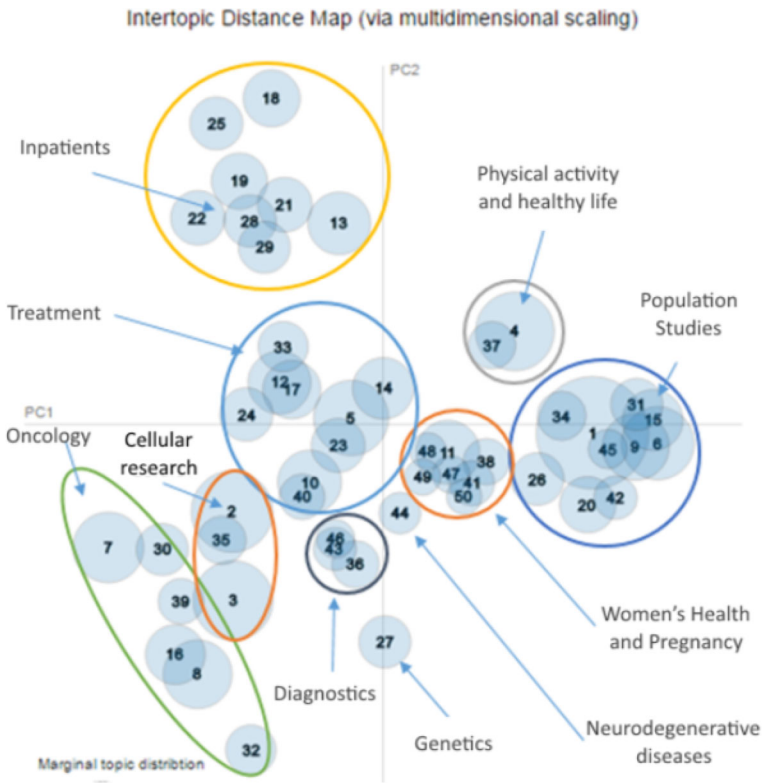

Fig. 7 Visualization of topic grouping done by the team of experts based on semantic grounds

the amounts of the acknowledged projects. This information would allow to compute research efficiency measures like the investment of money leading to each publication, and their differences among medical specialities. Another issue of interest is the relation between the prevalence and cost of the diseases and the cost of each publication. Unfortunately, at the time of writing we do not have this information, but it is an avenue of research worth pursuing.

Regarding funded research, we found that almost all groups of topics, except for groups with low scientific production (10 or less papers), receive similar funding measured in number of funded paper publications. On the other hand, there are strong differences in the publishing productivity among topics.

Regarding the generalization of the methodological approach

The basic methodological approach can be applied to diverse fields of science, and even documented industrial activity. However, our intermediate results can not be translated directly. In several of the methodological steps, a team of experts is required, i.e. to do the semantic identification of the LDA topics, and to provide interpretations of the visualization map axes and groupings of topics. Therefore, these intermediate results are only meaningful in the framework of the current study.

Conclusions

We have analyzed 643 scientific papers (2012-2015) of the chilean medical journal RevMed using topic modelling followed by topic visualization that reflects the topics in which medical research is being carried out in the country and their degree of funding. This analysis allowed to reduce gradually the quantity of information from the orginal 643 scientific documents, to 50 topics described by some 30 words each, down to the aggregation into 10 groups of topics located in the visualization space. Finally we are able to identify the meaning of the axes of the bidimensional space in which the topics where located as the size of the population under study (horizonatal axis) and the stage of the research carried out (vertical axis). A team of experts was able to interpret each topic representative words in order to assign them a specific semantic. The same was done for each group of topics and the dimensions of the bidimensional visualization space.

This study shows the application of text mining techniques in medical knowlege areas whose results can

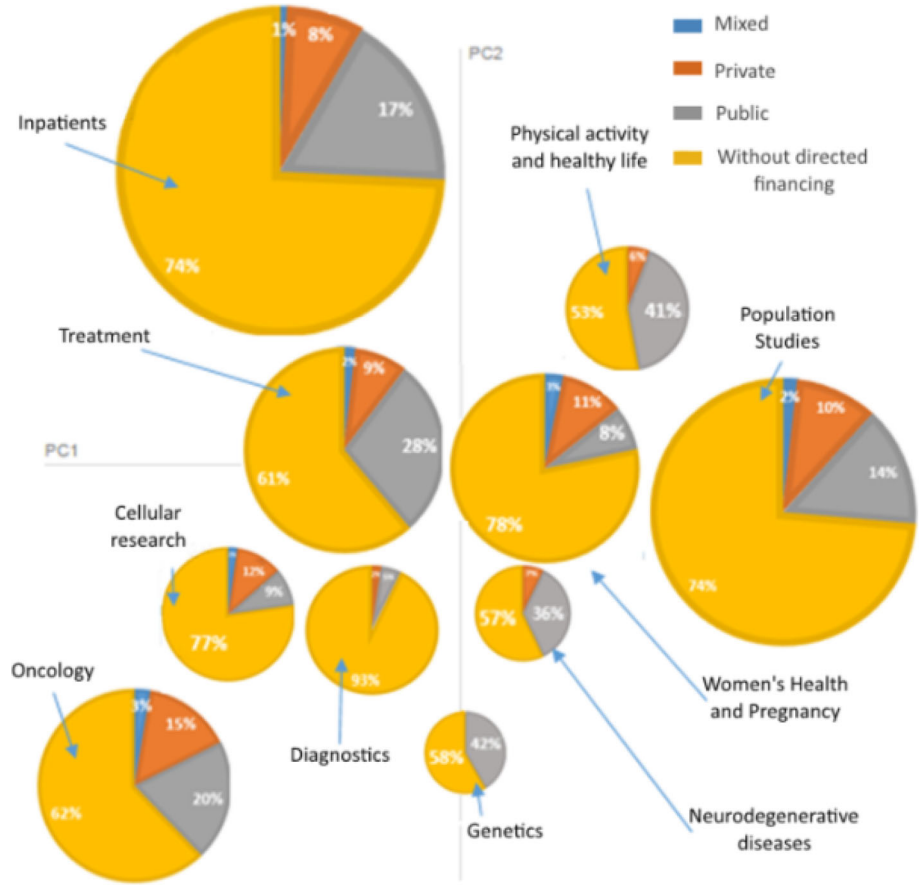

Fig. 8 Nature of funding sources for each group of topics

be utilized for socio-economical analysis of the research activity in the medical area, specifically we demonstrated that it provides tools to evaluate the impact of funding policies on the published research. Future work will be addressed to gather additional information in order to assess the specific funding resources backing each publication and its correspondence with the real needs of the population in terms of disease prevalence and estimated cost.

Availability of supporting data

The datasets during and/or analysed during the current study available from the corresponding author on reasonable request.

Abbreviations

D3: data driven document visualization; LDA: Latent Dirichlet Allocation; LDAvis: LDA visualization tool; LSA: Latent Semantic Analysis; Python: Programming language; R: The R Project for Statistical Computing; RevMed: Revista Medica de Chile

Acknowledgments

Authors would like to thank Ma. Begoña Yarza, M.D. for the support and good suggestion to enhance this work.

Authors' contributions

PW and DC were responsible for the development of the algorithms and the map, including all definitions where needed, and wrote and edited the manuscript. SR, MC, and MG provided scientific direction and contributed to the development of the map. All author(s) critically reviewed and edited the manuscript. All authors read and approved the final manuscript.

Funding

This research was partially funded by CONICYT, Programa de Formación de Capital Humano avanzado (CONICYT-PCHA / Doctorado Nacional / 2015-21150115). MG work in this paper has been partially supported by FEDER funds for the MINECO project TIN2017-85827-P, and projects KK-2018/00071 and KK-2018/00082 of the Elkartek 2018 funding program. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 777720. No role has been played by funding bodies in the design of the study and collection, analysis, or interpretation of data or in writing the manuscript.

Ethics approval and consent to participate

Does not apply.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Business Intelligence Research Center, Universidad de Chile, Beauchef 851, Santiago, 8370459 Santiago, Chile. [2] Computational Intelligence Group, University of Basque Country, P. Manuel Lardizabal 1, 20018 San Sebastián, Spain. [3] Universidad Adolfo Ibañez, , Santiago, Chile.

Received: 14 May 2019 Accepted: 2 August 2020
Published online: 29 September 2020

References

1. Kim S, Wilbur WJ. Thematic clustering of text documents using an em-based approach. J Biomed Semant. 2012;3(3):6. https://doi.org/10.1186/2041-1480-3-S3-S6.
2. Papanikolaou Y, Tsoumakas G, Laliotis M, Markantonatos N, Vlahavas I. Large-scale online semantic indexing of biomedical articles via an ensemble of multi-label classification models. J Biomed Semant. 2017;8(1):43. https://doi.org/10.1186/s13326-017-0150-0.
3. Collier N, Oellrich A, Groza T. Concept selection for phenotypes and diseases using learn to rank. J Biomed Semant. 2015;6(1):24. https://doi.org/10.1186/s13326-015-0019-z.
4. Arguello Casteleiro M, Demetriou G, Read W, Fernandez Prieto MJ, Maroto N, Maseda Fernandez D, Nenadic G, Klein J, Keane J, Stevens R. Deep learning meets ontologies: experiments to anchor the cardiovascular disease ontology in the biomedical literature. J Biomed Semant. 2018;9(1):13. https://doi.org/10.1186/s13326-018-0181-1.
5. Weissenborn D, Schroeder M, Tsatsaronis G. Discovering relations between indirectly connected biomedical concepts. J Biomed Semant. 2015;6(1):28. https://doi.org/10.1186/s13326-015-0021-5.
6. Zhao W, Zou W, Chen JJ. Topic modeling for cluster analysis of large biological and medical datasets. BMC Bioinformatics. 2014;15(11):11. https://doi.org/10.1186/1471-2105-15-S11-S11.
7. Wu X, Guo H, Cai K, Zhang L, Su Z. Linkthemall mining hybrid semantic associations from medical publications. In: 23rd International Conference of the European Federation for Medical Informatics. Oslo: University of Oslo; 2011.
8. Li DC, Thermeau T, Chute C, Liu H. Discovering associations among diagnosis groups using topic modeling. AMIA Jt Summits Transl Sci Proc. 2014;2014:43–49.
9. Crain SP, Yang S-H, Zha H, Jiao Y. Dialect topic modeling for improved consumer medical search. AMIA Annu Symp Proc. 2010;2010:132–6.
10. Shatkay H, Feldman R. Mining the biomedical literature in the genomic era: An overview. J Comput Biol. 2003;10(6):821–55. https://doi.org/10.1089/106652703322756104. PMID: 14980013.
11. Wang H, Ding Y, Tang J, Dong X, He B, Qiu J, Wild DJ. Finding complex biological relationships in recent pubmed articles using bio-lda. PLOS ONE. 2011;6(3):1–14. https://doi.org/10.1371/journal.pone.0017243.
12. Newman D, Karimi S, Cavedon L. Using topic models to interpret medline's medical subject headings. In: Nicholson A, Li X, editors. AI 2009: Advances in Artificial Intelligence. Berlin, Heidelberg: Springer; 2009. p. 270–9.
13. Wu Y, Liu M, Zheng WJ, Zhao Z, Xu H. Ranking gene-drug relationships in biomedical literature using latent dirichlet allocation. In: Biocomputing 2012. Singapore: World Scientific; 2012. p. 422–33.
14. Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci. 2004;101(suppl 1):5228–35. https://doi.org/10.1073/pnas.0307752101.
15. Barbosa-santill LI. Analysis of medical publications with latent semantic analysis method. In: IMMM 2013 The Third International Conference on Advances in Information Mining and Management. Lisbon: International Academy Research and Industry Association; 2013. p. 81–86.
16. Magerman T, Looy BV, Baesens B, Debackere K. Assessment of latent semantic analysis (LSA) text mining algorithms for large scale mapping of patent and scientific publication documents. SSRN. 2011;78:1–78. https://doi.org/10.2139/ssrn.2096159.
17. Goic AG. La Revista Métdica de Chile y la educación en medicina. Rev Med Chile. 2002;130:719–22.
18. Feinerer I, Hornik K, Meyer D. Text mining infrastructure in r. J Stat Softw. 2008;25(5):1–54.
19. Feinerer I, Hornik K. Tm: Text Mining Package. R package version 0.7-7. 2019. https://CRAN.R-project.org/package=tm. Accessed 1 Sept 2020.
20. Grün B, Hornik K. Topicmodels: An r package for fitting topic models. J Stat Softw Artic. 2011;40(13):1–30. https://doi.org/10.18637/jss.v040.i13.
21. Cao J, Xia T, Li J, Zhang Y, Tang S. A density-based method for adaptive LDA model selection. Neurocomputing. 2009;72(7):1775–81.
22. Arun R, Suresh V, Veni Madhavan CE, Narasimha Murthy MN. On finding the natural number of topics with latent dirichlet allocation: Some observations. In: Zaki MJ, Yu JX, Ravindran B, Pudi V, editors. Advances in Knowledge Discovery and Data Mining. Berlin, Heidelberg: Springer; 2010. p. 391–402.
23. Sievert C, Shirley K. LDAvis: A method for visualizing and interpreting topics. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. Baltimore, Maryland, USA: Association for Computational Linguistics; 2014. p. 63–70. https://doi.org/10.3115/v1/W14-3110 https://www.aclweb.org/anthology/W14-3110.
24. Blei DM. Probabilistic topic models. Commun ACM. 2012;55(4):77–84. https://doi.org/10.1145/2133806.2133826.
25. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, Zhao L. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. Multimed Tools Appl. 2019;78(11):15169–211. https://doi.org/10.1007/s11042-018-6894-4.
26. Gao Y, Li Y, Lau RYK, Xu Y, Bashar MA. Finding semantically valid and relevant topics by association-based topic selection model. ACM Trans Intell Syst Technol. 2017;9(1):1–22. https://doi.org/10.1145/3094786.
27. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.
28. Chuang J, Manning CD, Heer J. Termite: visualization techniques for assessing textual topic models. In: Advanced Visual Interfaces. New York: ACM Press; 2012. http://vis.stanford.edu/papers/termite.
29. Graham S, Weingart S, Milligan I. Getting started with topic modeling and mallet. Programm Historian. 2012;1:. https://programminghistorian.org/en/lessons/topic-modeling-and-mallet.
30. Havre S, Hetzler B, Nowell L. Themeriver: Visualizing theme changes over time. In: IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings. Pscataway: IEEE; 2000. p. 115–123. https://doi.org/10.1109/INFVIS.2000.885098.
31. Cao N, Sun J, Lin Y, Gotz D, Liu S, Qu H. Facetatlas: Multifaceted visualization for rich text corpora. IEEE Trans Vis Comput Graphic. 2010;16(6):1172–81. https://doi.org/10.1109/TVCG.2010.154.
32. Lynch S. Introduction to Applied Bayesian Statistics and Estimation for Social Scientists. New York: Springer; 2007. https://doi.org/10.1007/978-0-387-71265-9.
33. Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika. 1964;29(1):1–27. https://doi.org/10.1007/BF02289565.
34. Shapiro HD, Petroski H. Reviewed work: The pencil: A history of design and circumstance by henry petroski. JSTOR. 1991;82(2):355–56.

Publisher's Note