G. Drosatos, S. E. Kavvadias and E. Kaldoudi.

Topics and Trends Analysis in eHealth Literature.

In

H. Eskola et al. (eds.)
IFMBE Proceedings EMBEC & NBC 2017, vol 65, pp. 563-566, 2017
DOI: 10.1007/978-981-10-5122-7_141)

corrected after publication for corrigendum

Topics and Trends Analysis in eHealth Literature

G. Drosatos, S.E. Kavvadias and E. Kaldoudi

School of Medicine, Democritus University of Thrace, Alexandroupoli, Greece

Abstract— **eHealth is an interdisciplinary research area that fosters application of informatics and communication technologies for the improvement of healthcare delivery. In this paper, we present an overall analysis of eHealth topics and trends in published literature indexed in PubMed (all records till 31 Dec 2016, search on 25 Jan 2017), based on unsupervised topics modeling and trends analysis. Overall the analysis indicates a slightly declining (non significant) publication trend when compared to the overall PubMed corpus growth. Within the area of eHealth, a high negative trend is found for topics related to applications that support medical expert collaboration and consultation (e.g. teleradiology, image transmission, telesurgery, consultation between centres). On the contrary, a high positive trend is found for topics related to personalized eHealth applications, including mobile devices and patient empowerment.**

Keywords— **eHealth, trends analysis, topic modeling, LDA.**

I. Introduction

eHealth is an interdisciplinary research area that fosters application of informatics and communication technologies for the improvement of healthcare delivery. In general, eHealth systems and services are applied to assist patients in managing their disease at home and to help healthcare givers to provide remotely health services (and not only) [1]. The MeSH controlled vocabulary (by National Library of Medicine, USA) classifies eHealth in the group of terms including *telemedicine*, *telehealth* and *mobile health* (*mHealth*). A number of recent reviews cover different areas of the field, for example service models [2,3] or specific diseases [4,5,6]. In this paper, we present an overall analysis of eHealth topics and trends in published literature indexed in PubMed, based on unsupervised topics modeling [7,8].

II. Methodology

The analysis of eHealth literature trends followed a three-step approach. First, a generic PubMed query was used to build the corpus of published literature on eHealth. Then, main topics in the area of eHealth were identified via unsupervised topics modeling. Finally, trends were deduced based on the popularity of each topic per year.

A. Data collection

The corpus of publications that represents the eHealth domain was identified via a PubMed query that was built to include the most generic terms pertaining to the field, including the top MeSH term of the category and its synonyms as defined in MeSH. The search terms were restricted to title (TI) or abstract (AB) [Note: MeSH term “telemedicine” explodes the search to include synonyms: telehealth, ehealth, mobile health, mhealth]:

*("telemedicine"[MeSH Terms] OR "telemedicine"[TIAB] OR "ehealth"[TIAB]OR "e-health"[TIAB] OR "e*health"[TIAB] OR "tele-health"[TIAB] OR "telecare"[TIAB] OR "home monitoring"[TIAB] OR "telemonitoring"[TIAB]) AND ("0001/01/01"[PDAT] : "2016/12/31"[PDAT])*

The results of the query were exported as a XML file. The title, keywords and abstract of each publication was used to formulate the corpus for the topics modeling.

B. Topic modeling

Topic modeling algorithms are statistical methods that automatically extract topics from a large and unstructured collection of documents. In this work, we employ the algorithm of Latent Dirichlet Allocation (LDA) [7,8], as it has been shown to achieve highest precision in comparison to other topic modeling algorithms in corpora of Wikipedia and New York Times documents [9]. Furthermore, LDA has been successfully been applied in many other research areas, for example to analyze and classify genomic sequences [10], classify images based on visual words topic modeling [11], detect discussion themes in social networks [12] and analyze source code [13].

The LDA model assumes that each document is a mixture of topics. A topic is characterized by a collection of words, each word contributing with each own weight. A word can belong to multiple topics and documents can contain multiple topics. The algorithm starts by randomly assigns each word of a document in one of K topics. Then, it calculates conditional probabilities for each topic in each document ($P(t|d)$ where t denotes the topic and d denotes the document) and for each word in every topic ($P(w|t)$

where w denotes word). Through an iterative process, it reassigns words and topics until they reach a steady state. The algorithm requires setting the initial number K of assumed topics and the parameters that define the Dirichlet prior for the per document topic distribution (parameter α) and for the per topic word distribution (parameter β).

The implementation of LDA was based on the Java library jLDADMM [14] with a few required input/output and performance enhancements. The default values for the parameters of LDA were used, i.e. $\alpha = 0.1$ [15], and $\beta = 0.01$ [16], with 2000 iterations. Setting a small number for K, the identified topics may include several concepts, while setting a large value for K the identified topics may become over-compartmentalized. To identify a reasonable number K, we performed a series of investigative experiments using different number of topics (from K=40 to K=320) and concluded for an optimum value of K=160.

To avoid additive noise to the topic modeling algorithm from the free text of articles, we performed the following preprocessing cleaning process (implemented in Java): (1) remove all the punctuation and escape codes, (2) exclude all stop-words using the stop-words list from the Text Categorization Project [17], (3) convert all words to their lemmas by applying the stemming procedure of Krovetz stemmer [18] and (4) exclude articles with no words in their abstracts or less than 3 letters in their titles.

C. Trend analysis

Trends analysis follows the approach proposed by Priva and Austerweil [19]. First, the weight of each topic for each document is calculated as the percentage of the document words that belong to the particular topic. Then, the popularity $P(t,y)$ of the topic (t) for each year (y) is calculated as the mean of the weight of this topic for all documents published this year (D_y):

$$P(t,y) = \frac{1}{|D_y|}\sum_{d\in D_y}\frac{|\{w \in d{:}topic(w)=t\}|}{|d|} \qquad (1)$$

where t represents a topic and w is a word in document d of the documents' collection D_y for year y. The calculations of the above equation were implemented in Java.

Finally, we applied moving averaging (over 3 years interval) to smooth out short term fluctuations. Also, we used linear regression to identify the positive or negative trend for each topic.

III. Results

The PubMed query performed on 2017-01-25 with time limitation till 31 Dec 2016 and returned 25,824 publications (total XML file size of 207MB). After preprocessing, excluded 5,999 publications with no abstracts. The final corpus included title, abstract and keywords of 19,825 publications, corresponding to a total of 2,188,639 words, out of which 45,243 unique terms.

Fig. 1 shows the number of PubMed indexed publications per year retrieved via the query (blue line) and as a percentage of the total PubMed publications for the same year (as retrieved from the PubMed using as query *"0001/01/01"[PDAT]:"2016/12/31"[PDAT]*). Overall, there is an increase of the absolute number of publications on eHealth. When seen as the percentage of the overall PubMed corpus growth, linear regression shows a statistically significant increasing trend (regression coefficient =0.0000389, p-value < 0.05, R-squared =67.8%).

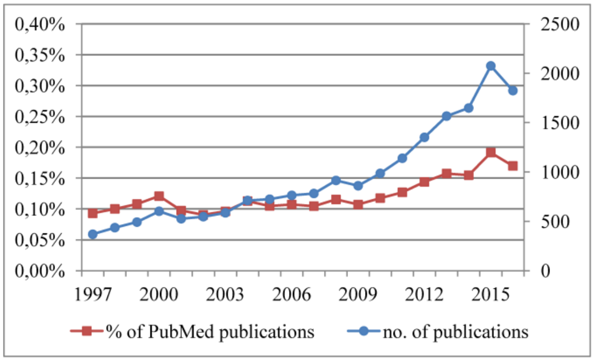

Fig. 1 eHealth publications per year: absolute number (circles) and percentage of total PubMed publications (squares).

Three eHealth experts reviewed independently the 160 topics and identified the major concept of each topic to serve as a topic title. The reviewing led to the identification of 96 clear topics (60% of all topics) which were organized by the experts in 7 categories as shown in Fig. 2.

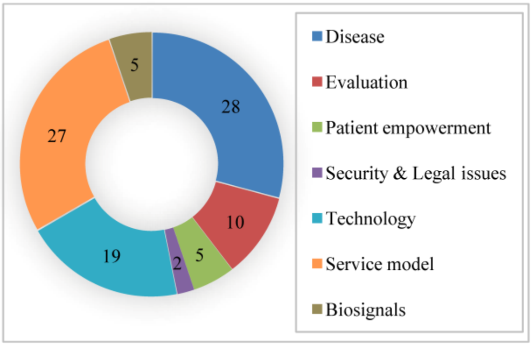

Fig. 2 Number of topics per category.

The top five most popular topics over the entire time span are: (1) wearables; (2) randomized control trials (RCT); (3) legal issues & ethics; (4) eye disease; and (5) teleconsultation.

Linear regression analysis showed a positive trend for 59 topics. The top 5 topics with the higher positive trend are: (1) randomized control trials; (2) depression; (3) remote patient monitoring; (4) mobile devices; (5) physical activity. A negative trend was found for 37 topics. The top 5 topics with the higher negative trend are: (1) teleradiology; (2) image transmission; (3) teleconsultation; (4) telesurgery; and (5) legal issues.

The topics of the most populated category of eHealth applications pertaining to a specific disease are shown in Table 1. Higher positive trends characterize eHealth applications specific to depression, stroke, heart failure, chronic disease, and chronic obstructive pulmonary disease (COPD). Higher negative trends characterize eHealth applications specific to dermatology, sudden infant death, and dental disease.

The analysis of topics related to eHealth service model show a positive trend for 9 topics and negative for 18. Topics with top positive trend are (1) remote patient monitoring; (2) rehabilitation; (3) adherence to medication; (4) health risk management; and (5) public health. Topics with top negative trend are (1) teleradiology; (2) teleconsultation; (3) telesurgery; (4) communication between tertiary establishment and remote clinic; and (5) space medicine.

Table 1 Diseases with positive and negative trends (ranked by regression coefficient).

Order	Disease category	Reg. Coeff.	R-squared
Positive trends			
1	Depression*	0.00072	94.88%
2	Stroke*	0.00050	87.95%
3	Heart failure*	0.00045	83.04%
4	Chronic*	0.00034	95.22%
5	COPD*	0.00033	92.81%
6-23	Diabetes*, Stress*, Smoke/alcohol cessation*, Elders, Pain*, Parkinson, ICU, HIV, Cancer*, Infectious*, Pediatrics, Cognitive function, Eye disease, INR lifetime anticoagulation therapy‡, Eating disorder, Wounds‡, Acute coronary disease‡, Psychiatry‡		
Negative trends			
1	Dermatology	-0.00037	88.47%
2	Infant – SID	-0.00024	73.88%
3	Dental	-0.00016	86.27%
4	Asthma‡	-0.00007	18.55%
5	Pregnancy‡	-0.00002	1.97%

* *Diseases with a good linear regression fit, R-squared > 80%*

‡ *Diseases with NON-significant reg. coefficient, p-value > 0.05*

An interesting category is the one with topics related to biosignals monitoring, as showed in Fig. 3. Positive trend is found in physical activity, blood pressure, while electrocardiogram (ECG) shows almost a neutral trend, and sleep monitoring and blood glucose show a negative trend.

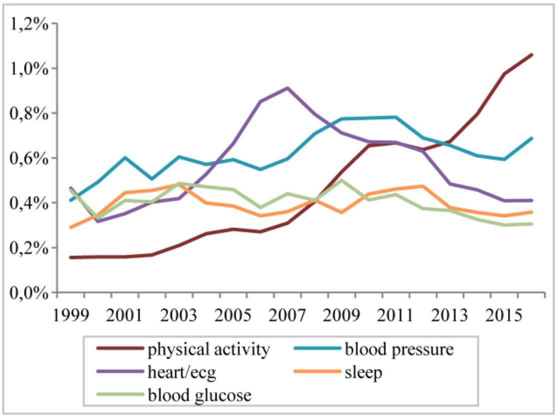

Fig. 3 Trends of topics in biosignal monitoring.

The topics related to different technologies in eHealth are shown in Table 2. Higher positive trends characterize mobile devices, wearables, data mining and classification, mobile phones and short messages, and web-based technologies. Higher negative trends characterize image transmission and real time video transmission.

Table 2 Technologies with positive and negative trends (ranked by regression coefficient).

Order	Technology category	Reg. Coeff.	R-squared
Positive trends			
1	Mobile devices	0.00064	64.62%
2	Wearables‡	0.00038	21.12%
3	Data mining and classification*	0.00037	92.51%
4-10	Mobile phone/SMS*, Web-based, Wireless sensors, Electronic health records, Portable/mobile devices, Speech/voice analysis‡, Decision support system		
Negative trends			
1	Image transmission*	-0.00114	93.46%
2	Real time video transmission*	-0.00047	90.38%
3	Internet based*	-0.00046	96.03%
4-9	Virtual reality/worlds*, Data compression, Telephone, Videoconference, Robotic systems‡, Standards‡		

* *Technologies a good linear regression fit, R-squared > 80%*

‡ *Technologies with NON-significant reg. coefficient, p-value > 0.05*

The trends of the five topics related to patient empowerment are shown in Fig. 4. They all exhibit positive trends, and especially the topic related to social media shows an exponential growth.

Finally, as previously mentioned, the topic on security and privacy issues shows an exponential growth, especially after year 2011. On the contrary, the topic on legal issues shows a strong negative trend.

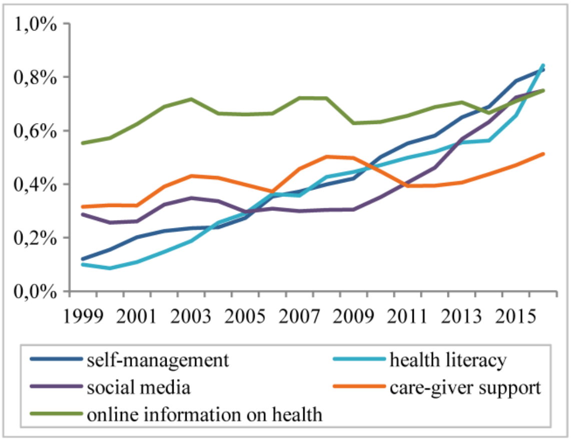

Fig. 4 Trends of topics in patient empowerment category.

IV. Discussion

Probabilistic topics modeling in a corpus of eHealth publications was used to identify in an unbiased way major topics in eHealth publications and calculate trends for the last 20 years. The LDA algorithm used in this study has been shown to exhibit the highest performance in topics modeling of textual corpora, albeit of a different nature [9]. As a confirmation, we compared the trends curve of the topic ‘remote patient monitoring’ as found in our study with trends analysis performed in a systematic review [3] of the same topic which identified 55 publications in the time span 2005-2014. This comparison showed similar trends. Major limitations of the study include the restriction to the corpus available in the PubMed indexing database and the subjective naming and grouping of the topics.

Overall the analysis indicates a statistically significant increasing trend of eHealth publications compared to the overall PubMed corpus growth. Within the area of eHealth a high negative trend is found for topics related to applications that support medical expert collaboration and consultation (e.g. teleradiology, image transmission, telesurgery, consultation between centres). On the contrary, a high positive trend is found for topics related to personalized eHealth applications, including mobile devices and patient empowerment.

Acknowledgment

This work was supported by the FP7-ICT project CARRE (Grant No. 611140), funded in part by the European Commission and Greek National Matching funds (DUTH Grant No. 81442).

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Boogerd EA, Arts T, Engelen LJ et al. (2015) “What is eHealth”: time for an update? JMIR research protocols 4
2. Silva BM, Rodrigues JJ, de la Torre Díez I et al. (2015) Mobile-health: A review of current state in 2015. JBI 56:265-272.
3. Vegesna A, Tran M, Angelaccio M et al (2016) Remote patient monitoring via non-invasive digital technologies: a systematic review. Telemed J E Health 23:3-17
4. Himes BE, Weitzman ER (2016) Innovations in health information technologies for chronic pulmonary diseases. Respir. Res 17:38
5. Rigla M (2011) Smart telemedicine support for continuous glucose monitoring: the embryo of a future global agent for diabetes care. J Diabetes Sci Technol 5:63-67
6. Hughes HA, Granger BB (2014) Racial disparities and the use of technology for self-management in blacks with heart failure: a literature review. Curr Heart Fail Rep 11:281-289
7. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet Allocation. J Machine Learning Res 3:993-1022
8. Blei DM (2012) Probabilistic topic models. Comm ACM 55:77-84
9. Chang J, Boyd-Graber JL, Gerrish S et al. (2009) Reading tea leaves: How humans interpret topic models. NIPS 31, pp 1-9
10. La Rosa M, Fiannaca A, Rizzo R et al (2015) Probabilistic topic modeling for the analysis and classification of genomic sequences. BMC bioinformatics 16:S2
11. Rasiwasia N, Vasconcelos N (2013) Latent Dirichlet allocation models for image classification. IEEE Trans PAMI 35:2665-2679
12. Lau JH, Collier N, Baldwin T (2012) On-line Trend Analysis with Topic Models. COLING, pp 1519-1534
13. Binkley D, Heinz D, Lawrie D et al (2014) Understanding LDA in source code analysis. 22nd Int. Conf. on Program Comprehension, ACM, pp 26-36
14. Nguyen DQ (2015) jLDADMM: A Java package for the LDA and DMM topic models. http://jldadmm.sourceforge.net
15. Yin J, Wang J (2014) A Dirichlet Multinomial Mixture Model-based Approach for Short Text Clustering. 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp 233-242
16. Griffiths TL, Steyvers M (2004) Finding scientific topics. National Academy of Sciences of the United States of America, 101:5228-5235
17. Text Categorization Project at http://code.google.com/p/text-categorization/
18. Krovetz R (1993) Viewing morphology as an inference process. 16th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, NY, USA, pp 191-202
19. Priva UC, Austerweil JL (2015) Analyzing the history of Cognition using topic models. Cognition 135:4-9

Corresponding author:

Eleni Kaldoudi
School of Medicine, Democritus University of Thrace
University Campus, Dragana
Alexandroupoli, Greece
kaldoudi@med.duth.gr