Journal of Big Data

SURVEY PAPER Open Access

# Knowledge discovery from a more than a decade studies on healthcare Big Data systems: a scientometrics study

Fatemeh Soleimani-Roozbahani[1], Ali Rajabzadeh Ghatari[2*] and Reza Radfar[1]

*Correspondence:
alirajabzadeh@modares.ac.ir
[2] Department of Management, Tarbiat Modares University, Tehran, Iran
Full list of author information is available at the end of the article

**Abstract**

Annually, lots of research papers are published in scientific journals around the world. The knowledge of the status of research is a prerequisite for research planning and policy making. This type of knowledge could be gained through a scientometrics study on the published literature that analyzes research products in a scientific field. Always healthcare was a permanent concern of researchers and also rapidly expanding field of Big Data analytics has started to play a pivotal role in the evolution of healthcare practices and research. It leads attracting attention from academia, industry and even governments around the world to "Big data in Healthcare". Therefore, this paper has done a meta-analysis on published researches methodology in this field in the period of 2008–2018. Statistical finding shows the "Meta-analysis and evidence" is the most used methodology in published papers. We applied data mining techniques for predicting using methodologies in the various databases to achieving knowledge discovery in the field. Naïve Bayes classifier in RapidMiner has been applied and results show eight main categories for words used in papers while "Developing methods to evaluate of care" averagely is the most intended using methodology for publishing papers and "Agent-based modeling" in nature is most using methodology and could be better predicted.

**Keywords:** Big Data, Healthcare, Knowledge discovery, Scientometrics study, Naïve Bayes, Published papers

## Introduction

The increasing development of information and communication technologies (ICT) has brought many achievements for human society and greatly influenced people's lives [1], and it has been adding significant benefits to various aspects of it [2]. Captured and stored huge quantities of information about people, their daily interactions and even their biotic signs via a variety of digital devices, potentially processed and analyzed by academic researchers, corporations, and governments [3]. Fortunately, the cost of information processing is cheap today [4], while organizations are using information systems for optimizing processes in order to increase coordination and interoperability across the organizations [5], and helps them to increase the integration and standardization of processes [6].

In the same vein, cutting-edge technologies such as Big Data have the potential to leverage the adoption of circular economy concepts by organizations and society, becoming

Springer Open

© The Author(s) 2019. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

more present in our daily lives [7]. Nowadays, scientific, research and commercial literature about Big Data corroborate penetrating its capabilities in all areas [8–11].

In another side, healthcare plays an important role in our societies. Improving the healthcare efficiency, accuracy, and quality of people is the main goal set forth by both the government and researchers [12]. The healthcare industry historically has generated large amounts of data, driven by record keeping, compliance and regulatory requirements, and patient-care [13].

The importance of healthcare to individuals and governments and its growing costs to the economy have contributed to the emergence of healthcare as an important area of research for scholars in business and other disciplines [14]. By now, the electronic collection, organization, annotation, storage, and distribution of heterogeneous data are essential activities in the contemporary biomedical, clinical, and translational discovery processes [15]. Therefore, Big Data in healthcare has become an emerging and remarkable research field. So that in the middle of 2018, Google Scholar displays 17,000 results for searching "Big Data in Healthcare" for the only year of 2018. And Big Data in healthcare has drawn substantial attention in recent years [12]. Big healthcare data has considerable potential to improve patient outcomes, predict outbreaks of epidemics, gain valuable insights, avoid preventable diseases, reduce the cost of healthcare delivery and improve the quality of life in general [16]. Based on this importance, there are numerous current areas of research within the field of Health Informatics, including Bioinformatics, Image Informatics (e.g. Neuroinformatics), Clinical Informatics, Public Health Informatics, and also Translational BioInformatics (TBI) [17]. Scientific publications in (bio) medicine show a massive increase in the number of papers published yearly that mention Big Data [18]. However, the identification of major hot topics and related research methodologies in big data in healthcare still lacks a comprehensive quantitative analysis. Provide tools by scientometrics approaches well suit to address questions of interdisciplinary integration in research fields [19]. They can help us identify cross-sectional patterns within scientific communities and can explicate how those patterns evolve over the life course of fields [19, 20]. The scientometrics studies help in get information about research areas that researchers are attentive in, how they like presenting the results of the research, the journals and publications they are interested in and the importance of a research topic in a specific time period that based on the information research policies can be made with a less probability of mistakes.

Scientometrics aims at the advancement of knowledge on the development of science and technology [21]. According to Van Raan's claim about the relationship between knowledge discovery and scientometrics, it can be resulted in one of the functions of it is the 'knowledge discovery' [21].

Some scholars and practitioners use the notion of 'V' to define 'Big Data' [22–25], 3V, 5V and even 7V. volume, velocity, variety in 3V [22–25], and then value and veracity have been added for 5V [26] and recently 7V added variability and visualization [27, 28]. In the field of "Big Data in Healthcare" few researchers have focused on this issue and they considered the 5V. Jatrniko et al. mentioned 5V as the defining factors of Big Data [29]. In other study 5V introduced as patient data attributes [30]. Van and Alagar believe 5V characterize Big Data and motivate their relevance to healthcare data [31]. In a study named "Big Data stream computing in healthcare real-time analytics", the challenges in

big data analysis on health care can be understood by 5V s characteristics [32]. "Big Data, Big Knowledge: Big Data for Personalized Healthcare" fully described 5V in healthcare: *Volume* the community wishes to exploit the vast entirety of clinical data records, but the datasets that support these analyses are often very expensive to acquire, and currently the penetration is limited [33]. *Variety* it explains the diversified data sets with respect to the structured, semi-structured and unstructured data sets [30], in healthcare field variety could be defined as clinical data, data from medical imaging, data from wearable sensors, lab exams, and simulation results [33]. *Velocity* is expressed in terms of data arrival rate from the patients. *Veracity* while data collected as part of clinical studies are in general of good quality, clinical practice tends to generate low quality data. This is due in part to the extreme pressure medical professionals face, but also to a lack of "data value" culture; most medical professionals see the logging of data a bureaucratic need and a waste of time that distracts them from the care of their patients. *Value* refers to the "economic value" that results in saving and analyzing Big Data [31]. For example in general, healthcare expenditure in most developed countries is astronomical: the 2013/2014 budget for NHS England was £95.6 billion, with an increase over the previous year of 2.6%, at a time when all public services in the UK are facing hard cuts. In OECD countries, we spend on average USD$3395 per year per inhabitant in healthcare [34].

This paper aims to address an analysis of the more considerable research output (papers published in the seven important databases) "Big Data in Healthcare" for achieving a deep and comprehensive trend study and based on it, makes a knowledge discovery from the publications. Using Naïve Bayes, results identified a classification of methodologies used in duplicated papers in journals.

## Methods and materials

The statistical population of research is 82,313 papers, shown as the search result of "Big Data in Healthcare" in intended databases. The source of data is articles (conference papers, articles, reviews, articles in press and survey) published in selected databases. Since the studies in the field of the Big Data are toddle and the subject of using Big Data in the field of healthcare is not more than 10 years old, we chose the period 2008–2018. The reason for choosing these databases are based on the carried-out evaluation on 20 well-known databases (IEEE, Elsevier, Wiley Online Library, Springer, Nature, Taylor and Francis Online,[1] ACM Digital Library,[2] ASP Publication,[3] JStore,[4] AIP,[5] Emerald Insight[6] and ASME,[7] Sage journals,[8] Oxford Journals, World Scientific,[9] AMS,[10] Annual

[1] http://taylorandfrancis.com/journals/.

[2] https://dl.acm.org/.

[3] https://pubs.acs.org/.

[4] https://www.jstor.org/.

[5] https://www.aip.org/.

[6] https://www.emeraldinsight.com/.

[7] https://www.asme.org/.

[8] http://online.sagepub.com/.

[9] http://www.worldscientific.com/.

[10] http://www.ams.org/journals.

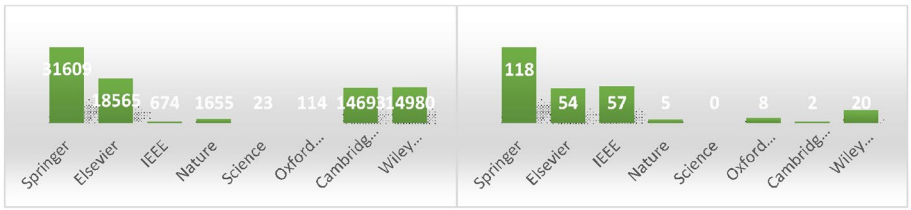



**Fig. 1** Frequency chart of databases' publication (left) vs number of selected paper to study from every database (right)

Reviews,[11] Cambridge University Press and Royal Society);[12] the largest number of papers in the field of interest has been published in the selected databases, but after the first review, it turned out that the number of articles published on some of these sites was very small, so these databases were deleted and Elsevier, IEEE, Springer, Nature, Science, Oxford Journals, Cambridge University Press and Wiley Online Library remained. Figure 1 shows the frequency chart of databases' publication vs the number of selected paper to study from every database after refining.

Two different databases have been ready for this research. The first one contains 265 records (refined papers). The other one contains eight datasets of papers, seven for training data mining model and the other one for testing the model. Researches applied VOSviewer 1.6.9 to draw up the maps and RapidMiner Studio 8.2 for data mining.

## Quality control

Quality control was carried out in several stages: firstly, in the weekly meetings, doubtful papers, reviews, and feedbacks were given. Second, ten percent of all papers were reviewed by the research coordinator and provided feedback to colleagues. The evaluation of the re-examination process indicated a significant reduction in errors and disagreements. Third, collected data forms were checked and be ready for data entry after completing, in the end of data entry stage, the accuracy of 20% of the entered data is reviewed.

Although all the shown results were not aligned with the researchers' purpose, all of them had been investigated and the appropriated ones had been selected to analyze. Therefore, the total number of investigated papers is 82,313 and the number of selected papers is 265. It needs to be notifying Science database had no related paper to use in current research.

## Results and discussion

In this study, in first phase the overall status of researches on Big Data in healthcare, and related science was studied. Totally 265 papers were evaluated in a 12-year period in seven databases. Most of the first-ever authors had a Ph.D. or higher degree whose affiliations were universities. Furthermore, each article had an average of 4.1 writers.

[11] https://www.annualreviews.org/.

[12] https://royalsociety.org/journals/.

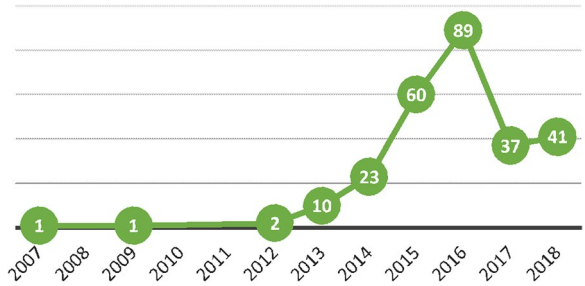



**Fig. 2** Number of published papers in various years

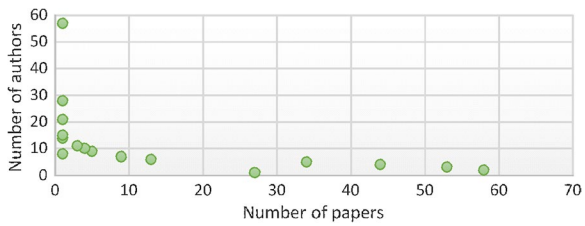



**Fig. 3** Number of authors of papers

The process of publishing papers is shown in Fig. 2. The number of articles published until 2013 was limited and insignificant and did not fluctuate significantly, but since 2013 there has been a noticeable upward trend, we are published papers.

Figure 3 shows the number of papers' authors. The most number of authors belongs to one paper titled "MAKING SENSE OF BIG DATA IN HEALTH RESEARCH: TOWARDS AN EU ACTION PLAN" written by 57 persons in 2016 and has been published by GENOME MEDICINE journal in Springer database. Its authors' information is available at the end of paper. They are scientists working in universities and research centers in different European countries. After this paper, the most number of authors respectively are 28 (one paper), 21 (one paper), 15 (one paper), 14 (one paper), 11 (three papers), 10 (four papers).

The "subject area" of the study that was categorized into 17 areas adopted with minor adaptation of the criteria used in Hermon and Williams [35] study. "Healthcare Data Analysis" is with the 74 papers is the most noteworthy subject of the article. "E-health" with 50 papers placed after it. The other places and the number of their publications are shown in Fig. 4.

Usually, data mining techniques are used in most of the studied papers to analyze their data (179 paper). Figure 5 shows the frequency of used data mining techniques (in percentage). The decision tree is the most used technique while the Appriori algorithm is the least applied technique.

The journals and the number of papers published shows in Fig. 6. "Journal of Medical Systems" has published the largest number (seven paper) of papers on the field. The other journals have published at least two papers.

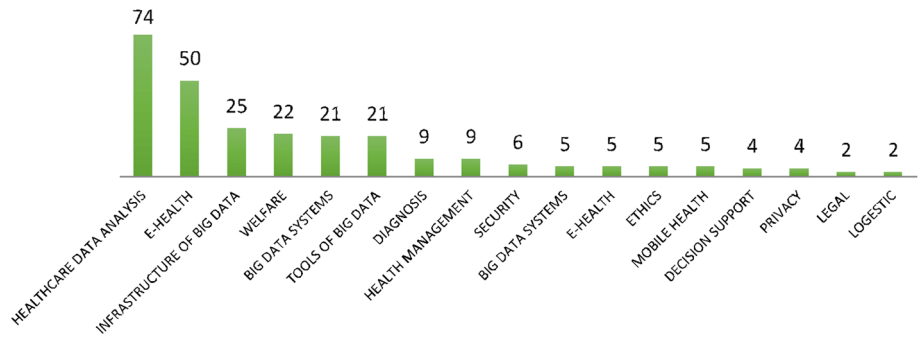



**Fig. 4** Papers subject categories

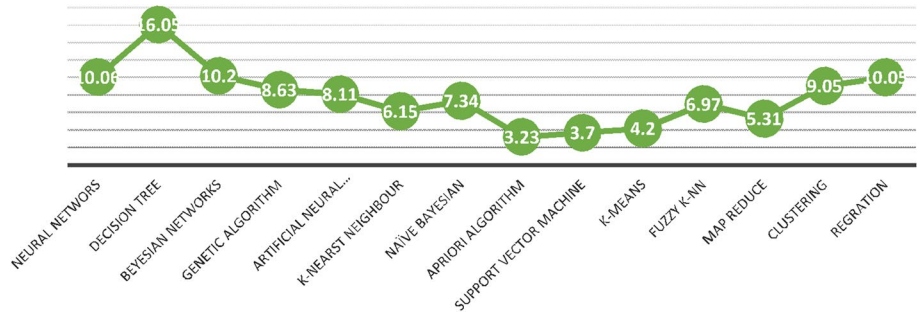



**Fig. 5** Data mining techniques used in papers

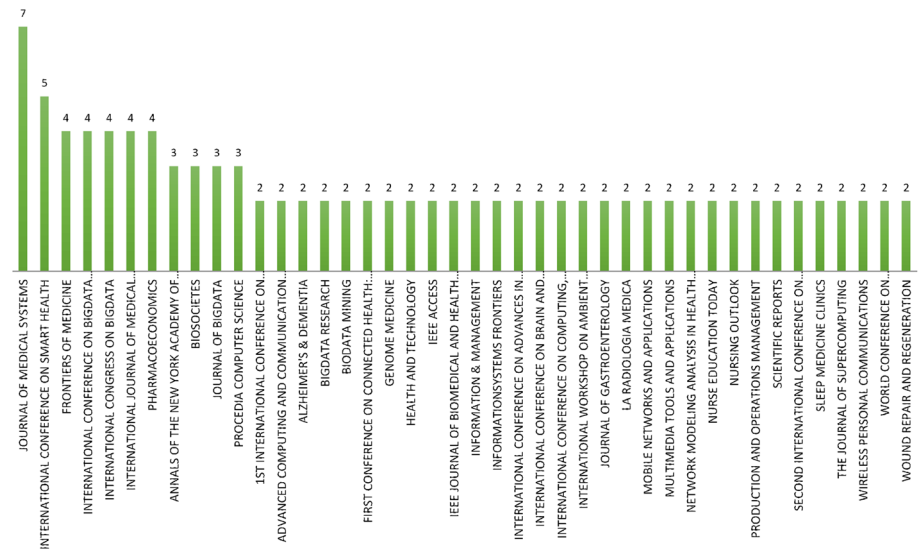



**Fig. 6** The number of published paper in each journal

Based on some categories used by researches [35] and some seen constant terms in papers 17 categories (as it previously mentioned) were considered for papers. The frequency distribution of the Big Data in healthcare published papers in the period 2007–2018 in terms of papers categories shows in Fig. 7.

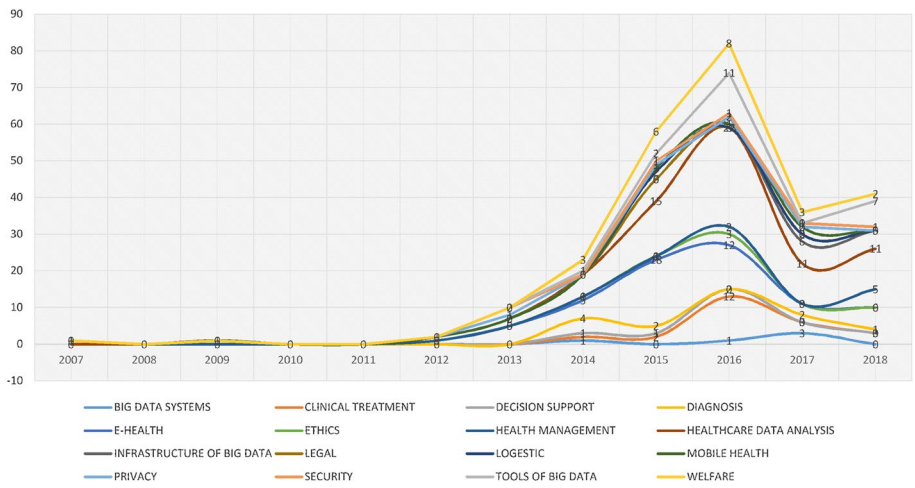



**Fig. 7** The frequency distribution of the Big Data in healthcare published paper in terms of papers categories

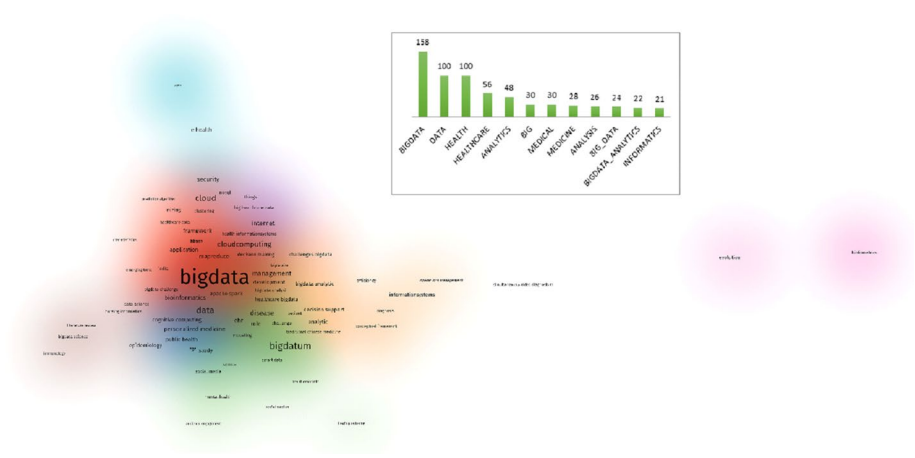



**Fig. 8** Conceptual map of the keywords vs frequency chart of them

Figure 8 shows the conceptual map of the words used as keywords of papers (made by VOSviewer) vs the bar chart of frequency of the most repeated keywords (made by RapidMiner).

Conceptual map of the words used in the title, abstract and conclusion of the papers is shown in Fig. 9. As it can be seen, these words are located in eight clusters. These clusters are based on the most frequent words.

Big Data methodologies used in papers based on the content and methodology applied were classified into nine categories, including data quality grading and assurance, identifying "unusual" data segments, machine learning and transactional data, developing methods to evaluate of care, meta-analysis and evidence, agent-based modeling, early warning systems, text data mining, tracking interactions among users. Figure 6 shows frequency of used Big Data methodology in papers (in percentage). The most used methodology is meta-analysis and evidence used in 99 papers, and the least one is "Tracking interactions among users" used in 37 papers (Fig. 10).

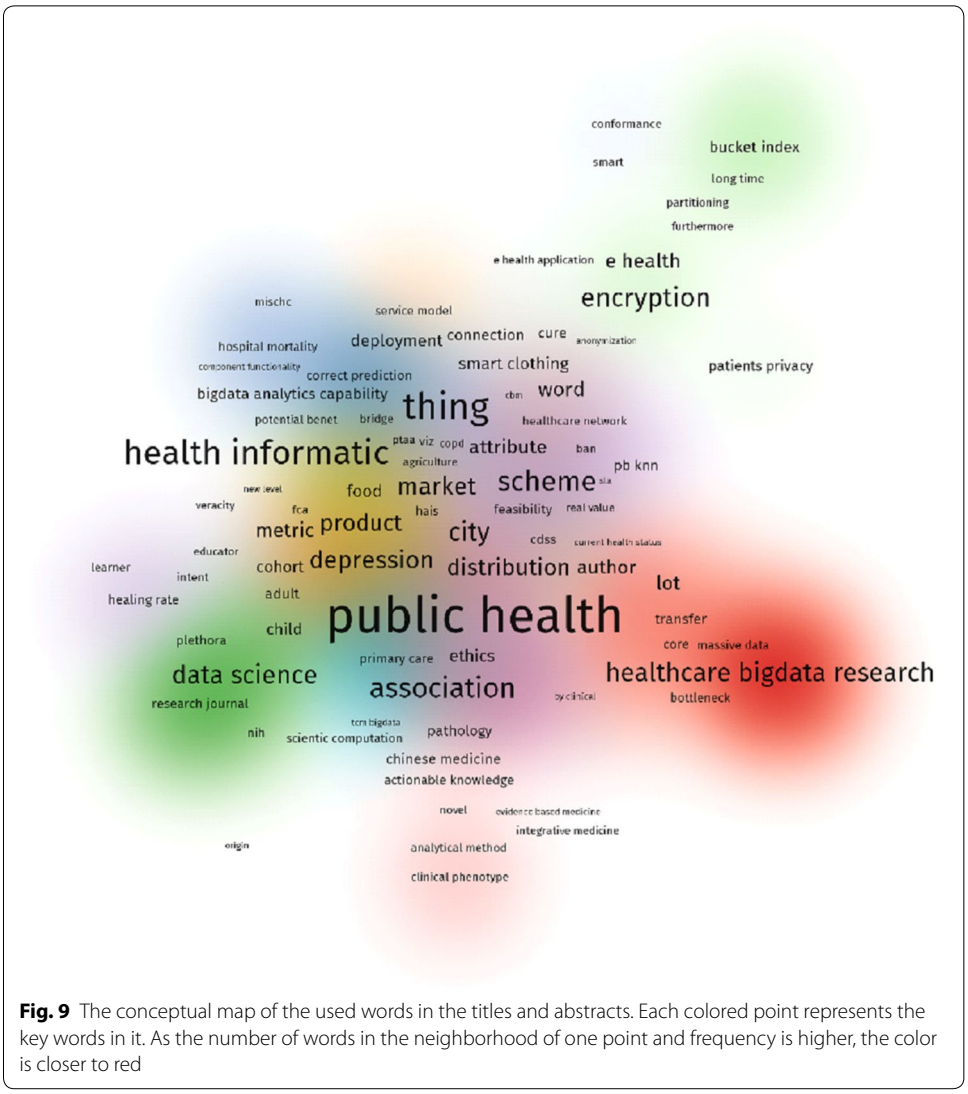



**Fig. 9** The conceptual map of the used words in the titles and abstracts. Each colored point represents the key words in it. As the number of words in the neighborhood of one point and frequency is higher, the color is closer to red

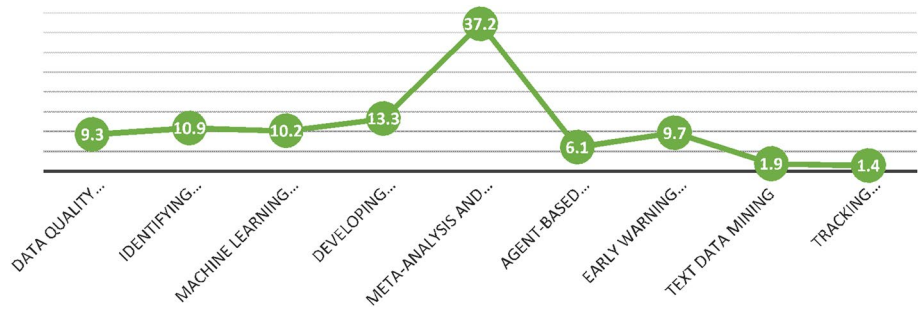



**Fig. 10** Frequency of used Big Data methodology

**Table 1 Technical overview of best six data mining open source tools [37]**

| S. N | Tool name | Release date | Operating system | Language | Website |
|---|---|---|---|---|---|
| 1 | RapidMiner | 2006 | Cross platform | Language independent | http://www.rapidminer.com |
| 2 | ORANGE | 2009 | Cross platform | Python C++, C | http://www.orange.biolab.si |
| 3 | KNIME | 2004 | Linux, OS X, Windows | Java | http://www.knime.org |
| 4 | WEKA | 1993 | Cross platform | Java | http://www.cs.waikato.ac.nz/~ml/weka |
| 5 | KEEL | 2004 | Cross platform | Java | http://www.sci2s.ugr.es/keel |
| 6 | R | 1997 | Cross platform | C, Fortran and R | http://www.r-project.org |

**Table 2 Tool with best accuracy in tested datasets [36]**

| BM[a] | W[b] | N[c] | CE[d] | BC[e] | SB[f] | Technique |
|---|---|---|---|---|---|---|
| WEKA (88.04) | RapidMiner (100) | WEKA (90.67) | WEKA (87.58) | RapidMiner (97.06) | R (98.01) | NB[g] |
| WEKA (88.00) | RapidMiner (98.30) | WEKA (90.32) | KNIME (86.57) | WEKA (97.42) | KNIME (89.91) | |
| R (90.30) | WEKA (98.36) | R (97.30) | WEKA (95.40) | RapidMiner (97.60) | WEKA (93.15) | DT[h] |
| R (90.20) | WEKA (98.87) | R (98.10) | WEKA (95.08) | R (95.60) | WEKA (93.24) | |
| R (89.10) | WEKA (95.08) | WEKA (97.52) | R (91.05) | WEKA (95.79) | WEKA (89.00) | KNN[i] |
| R (89.10) | WEKA 94.94 | WEKA (98.37) | WEKA (93.51) | WEKA (94.84) | WEKA (90.76) | |

[a] Bank Marketing (https://archive.ics.uci.edu/ml/datasets/Bank+Marketing)

[b] Wine (https://archive.ics.uci.edu/ml/datasets/Wine)

[c] Nursery (https://archive.ics.uci.edu/ml/datasets/Nursery)

[d] Car evaluation (https://archive.ics.uci.edu/ml/datasets/Car+Evaluation)

[e] Breast Cancer Wisconsin (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original))

[f] Spambase (https://archive.ics.uci.edu/ml/datasets/Spambase)

[g] Naïve Bayes

[h] Decision tree

[i] K nearest neighbor

Based on descriptive statistics, "Meta-analysis and evidence" is a methodology used in most papers, but this research is based on the knowledge discovery, applied data mining techniques to predict the different methodologies used in the published papers in various databases.

The second part of the analysis has been done by RapidMiner Studio 8.2. In first step, text mining was done on data. RapidMiner is the best to handle continues data type [36]. The reason of using Rapid Miner, compared with other data mining tools like WEKA, orange, and R is that it provides the fully automatic parameter optimization of machine learning operator and presents good validation and cross validation. Introduced by some studies as the first of best open source data mining tools [36, 37]. Tables 1 and 2 show the results of studies.

Since naive Bayes is a high-bias, low-variance classifier, and it can build a good model even with a small data set [38], it has been used to achieve the aim. In classification, the goal of a learning algorithm is to construct a classifier given a set of training examples with class labels [39].

Naïve Bayes does the calculation for all possible label values and selects the label value that has maximum calculated probability.

The naive Bayes classifier is the simplest of these models, in that it assumes that all attributes of the examples are independent of each other given the context of the class. This is the so-called "naive Bayes assumption". While this assumption is clearly false in most real-world tasks, naive Bayes often performs classification very well. This paradox is explained by the fact that classification estimation is only a function of the sign (in binary cases) of the function estimation; the function approximation can still be poor while classification accuracy remains high [40, 41].

Document classification is just such a domain with a large number of attributes. The attributes of the examples to be classified are words, and the number of different words can be quite large indeed. While some simple document classification tasks can be accurately performed with vocabulary sizes less than one hundred, many complex tasks on real-world data from the Web, UseNet and newswire articles do best with vocabulary sizes in the thousands. Naive Bayes has been successfully applied to document classification in many research efforts [42–45].

In this research process was a little different because the datasets were text files of papers, therefore, we used documentation operators of RapidMiner such as "Process Documents from Files" and "Split Validation".

Datasets used for knowledge discovery, as it has mentioned, are published papers in Big Data in healthcare, generally in natural language. Therefore, firstly, they have been processed by RapidMiner text mining operators then the classifiers have been used for training as well as testing. It needed to classify papers into nine types of class labels (previously mentioned methodologies, Fig. 10). These labels are used to train the classifier operator and then based on it; the classifier will predict the label of the test dataset. This supervised process has been repeated for every seven databases. Figure 11 shows Naïve Bayes prediction for IEEE and Fig. 11 shows the accuracy of Naïve Bayes classifier for IEEE for the same database.

Accuracy is the most important criterion for determining the efficiency of a model calculating the exact criterion for the entire category. Based on it, Naïve Bayes for Wiley Online Library has the best result. In another side, recall measures the completeness, or sensitivity, of a classifier. Higher recall means few false negatives, while lower recall means more false negatives [46].

As it can be seen in Fig. 12, it is predicted 71.43% publishing papers in IEEE will use "Machine learning and transactional data" methodology by accuracy rate of 41.49%.

This test has been done for all of databases. Based on results: for Cambridge university press by accuracy rate of 46.67%, it is predicted 100% of publishing papers will use "Agent-based modeling" as their methodology (Fig. 13).

The result of Naïve Bayes test on Nature papers by 36.25% accuracy shows the probability of using "Developing methods to evaluate of care" methodology in this database is 67.86% (Fig. 14).

This result for Elsevier presents "Developing methods to evaluate of care" methodology is predicted for 75% of papers by accuracy rate of 47.95% (Fig. 15).

| confidence (data | confidence (iden | confidence (mac | confidence (de | confidence (met | confidence (agent bas | confidence (war | confidence (tex | confidence (use | prediction (label) |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | agent based modellin |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | warning systems |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | text mining |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | meta analysis |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | developing methods |
| 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | data quality grading a |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | meta analysis |
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | user interaction |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | dentifying unusual da |
| 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | meta analysis |
| 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | developing methods |

**Fig. 11** Naïve Bayes prediction for IEEE database

**accuracy: 41.49%**

| | true data quality grading and assurance | true identifying unusual data | true machine learning and transactional data | true developing methods | true meta analysis | true agent based modelling | true warning systems | true text mining | true label | Class precision |
|---|---|---|---|---|---|---|---|---|---|---|
| pred. data quality grading and assurance | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. identifying unusual data | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. machine learning and transactional data | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 100.00% |
| pred. developing methods | 0 | 0 | 2 | 19 | 7 | 6 | 1 | 7 | 3 | 42.22% |
| pred. meta analysis | 1 | 0 | 0 | 4 | 9 | 5 | 0 | 2 | 1 | 40.91% |
| pred. agent based modelling | 0 | 0 | 0 | 2 | 3 | 1 | 0 | 4 | 0 | 10.00% |
| pred. warning systems | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. text mining | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 3 | 0 | 50.00% |
| pred. label | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 2 | 33.33% |
| class recall | 0.00% | 0.00% | 71.43% | 67.86% | 47.37% | 7.69% | 0.00% | 17.65% | 33.33% | |

**Fig. 12** Accuracy of Naïve Bayes classifier for IEEE

**accuracy: 46.67%**

| | true data quality grading and assurance | true identifying unusual data | true machine learning and transactional data | true developing methods | true meta analysis | true agent based modelling | true warning systems | true text mining | true label | Class precision |
|---|---|---|---|---|---|---|---|---|---|---|
| pred. data quality grading and assurance | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. identifying unusual data | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. machine learning and transactional data | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. developing methods | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. meta analysis | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. agent based modelling | 1 | 1 | 1 | 1 | 1 | 7 | 1 | 1 | 1 | 46.67% |
| pred. warning systems | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. text mining | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. label | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| class recall | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 100.00% | 0.00% | 0.00% | 0.00% | |

**Fig. 13** Accuracy of Naïve Bayes classifier for Cambridge University Press

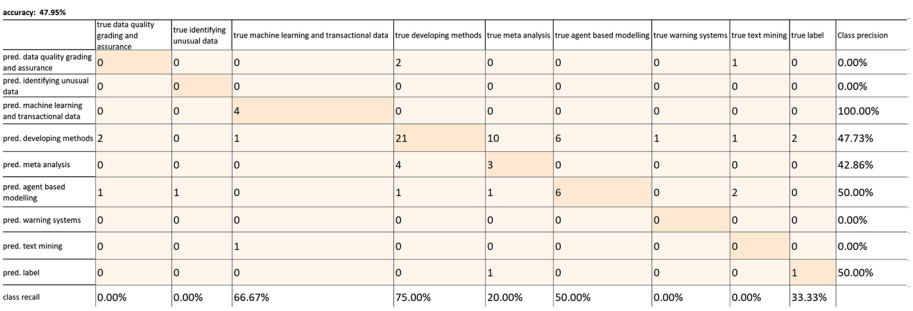
**accuracy: 36.25**

| | true data quality grading and assurance | true identifying unusual data | true machine learning and transactional data | true developing methods | true meta analysis | true agent based modelling | true warning systems | true text mining | true label | Class precision |
|---|---|---|---|---|---|---|---|---|---|---|
| pred. data quality grading and assurance | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 1 | 0.00% |
| pred. identifying unusual data | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0.00% |
| pred. machine learning and transactional data | 0 | 0 | 4 | 0 | 0 | 2 | 0 | 0 | 0 | 66.67% |
| pred. developing methods | 1 | 0 | 1 | 19 | 15 | 7 | 1 | 3 | 3 | 38.00% |
| pred. meta analysis | 2 | 0 | 0 | 5 | 2 | 0 | 0 | 0 | 0 | 22.22% |
| pred. agent based modelling | 0 | 1 | 0 | 2 | 1 | 2 | 0 | 1 | 0 | 28.57% |
| pred. warning systems | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. text mining | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. label | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 66.67% |
| class recall | 0.00% | 0.00% | 66.67% | 67.86% | 10.53% | 16.67% | 0.00% | 0.00% | 33.33% | |

**Fig. 14** Accuracy of Naïve Bayes classifier for nature

The "Meta-analysis and evidence" is predicted dominant methodology using in Springer by probability of 92.86% and accuracy of 62% (Fig. 16).

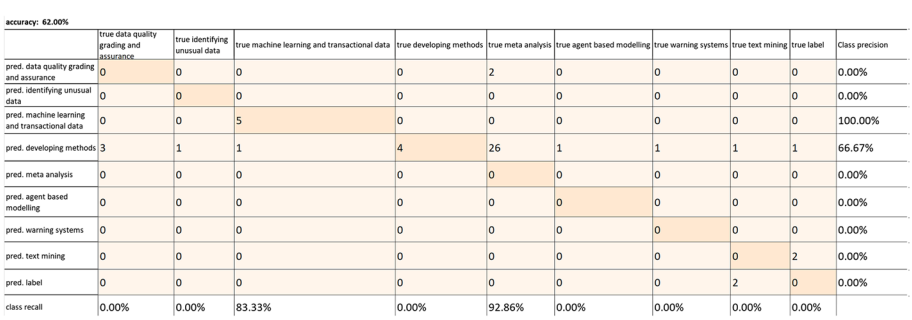
It is predicted in 93.02% of papers of Wiley Online Library will use "Developing methods to evaluate of care" by the accuracy rate of 64.06% (Fig. 17).



accuracy: 47.95%

| | true data quality grading and assurance | true identifying unusual data | true machine learning and transactional data | true developing methods | true meta analysis | true agent based modelling | true warning systems | true text mining | true label | Class precision |
|---|---|---|---|---|---|---|---|---|---|---|
| pred. data quality grading and assurance | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0.00% |
| pred. identifying unusual data | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. machine learning and transactional data | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 100.00% |
| pred. developing methods | 2 | 0 | 1 | 21 | 10 | 6 | 1 | 1 | 2 | 47.73% |
| pred. meta analysis | 0 | 0 | 0 | 4 | 3 | 0 | 0 | 0 | 0 | 42.86% |
| pred. agent based modelling | 1 | 1 | 0 | 1 | 1 | 6 | 0 | 2 | 0 | 50.00% |
| pred. warning systems | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. text mining | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. label | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 50.00% |
| class recall | 0.00% | 0.00% | 66.67% | 75.00% | 20.00% | 50.00% | 0.00% | 0.00% | 33.33% | |

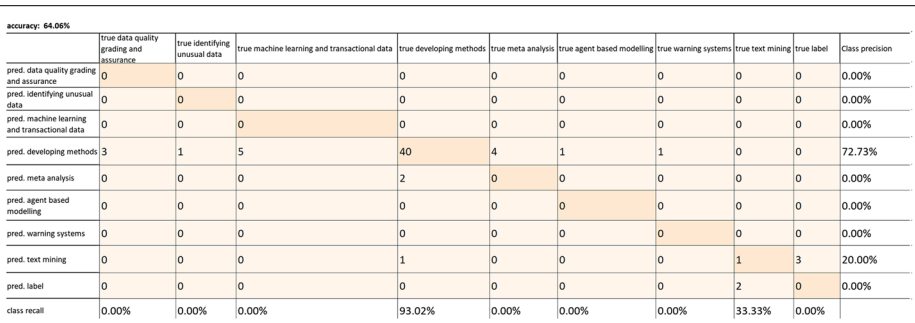
**Fig. 15** Accuracy of Naïve Bayes classifier for Elsevier



accuracy: 62.00%

| | true data quality grading and assurance | true identifying unusual data | true machine learning and transactional data | true developing methods | true meta analysis | true agent based modelling | true warning systems | true text mining | true label | Class precision |
|---|---|---|---|---|---|---|---|---|---|---|
| pred. data quality grading and assurance | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0.00% |
| pred. identifying unusual data | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. machine learning and transactional data | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 100.00% |
| pred. developing methods | 3 | 1 | 1 | 4 | 26 | 1 | 1 | 1 | 1 | 66.67% |
| pred. meta analysis | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. agent based modelling | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. warning systems | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. text mining | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0.00% |
| pred. label | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0.00% |
| class recall | 0.00% | 0.00% | 83.33% | 0.00% | 92.86% | 0.00% | 0.00% | 0.00% | 0.00% | |

**Fig. 16** Accuracy of Naïve Bayes classifier for Springer



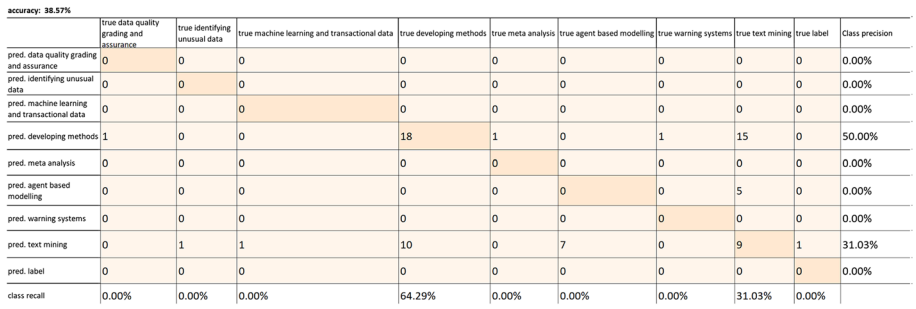
accuracy: 64.06%

| | true data quality grading and assurance | true identifying unusual data | true machine learning and transactional data | true developing methods | true meta analysis | true agent based modelling | true warning systems | true text mining | true label | Class precision |
|---|---|---|---|---|---|---|---|---|---|---|
| pred. data quality grading and assurance | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. identifying unusual data | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. machine learning and transactional data | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. developing methods | 3 | 1 | 5 | 40 | 4 | 1 | 1 | 0 | 0 | 72.73% |
| pred. meta analysis | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. agent based modelling | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. warning systems | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. text mining | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 3 | 20.00% |
| pred. label | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0.00% |
| class recall | 0.00% | 0.00% | 0.00% | 93.02% | 0.00% | 0.00% | 0.00% | 33.33% | 0.00% | |

**Fig. 17** Accuracy of Naïve Bayes classifier for Wiley Online Library

This test results show the probability of using "Developing methods to evaluate of care" methodology in papers publishing in Oxford Journals is 64.29% by accuracy rate of 38.57% (Fig. 18).

Therefore, averagely it is predicted "Developing methods to evaluate of care" methodology is the most intended methodology for publishing papers and in another side "Agent-based modeling" in Wiley Online Library has fewer false results.



accuracy: 38.57%

| | true data quality grading and assurance | true identifying unusual data | true machine learning and transactional data | true developing methods | true meta analysis | true agent based modelling | true warning systems | true text mining | true label | Class precision |
|---|---|---|---|---|---|---|---|---|---|---|
| pred. data quality grading and assurance | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. identifying unusual data | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. machine learning and transactional data | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. developing methods | 1 | 0 | 0 | 18 | 1 | 0 | 1 | 15 | 0 | 50.00% |
| pred. meta analysis | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. agent based modelling | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0.00% |
| pred. warning systems | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. text mining | 0 | 1 | 1 | 10 | 0 | 7 | 0 | 9 | 1 | 31.03% |
| pred. label | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00% |
| class recall | 0.00% | 0.00% | 0.00% | 64.29% | 0.00% | 0.00% | 0.00% | 31.03% | 0.00% | |

**Fig. 18** Accuracy of Naïve Bayes classifier for Oxford Journals

## Conclusion and further direction

In this paper we performed a scientometric study on published research papers during last 11-year period to "Big Data in Healthcare" researches characterization while has used the Naïve Bayes data mining technique to explore knowledge from them.

Results show the most duplicated papers belong to Springer database, and year of 2016 had the high frequency of publication. Big Data" are the high-frequency words and also key words and results verified the expectation. Results of applying VOSviewer for keywords, title, abstracts and conclusion of papers show eight clusters of words. The clusters are: public health, health informatics, healthcare big data research, data science, association, e-health encryption and things. Journal of medical systems published most papers in the field. decision Tree was most used techniques in data mining in papers applied data mining. The most number of author is 57. Health data analytics has the first rank among the subject. Males having Ph.D. degree with university affiliations had the dominant rate of authors. Meta-analysis and evidence was the most used Big Data methodology. In addition to descriptive statistics methods, in order to perform scientometrics study, a prediction technique (classification) has been done on Big Data methodology used in the papers of various databases and knowledge discovered from them. According to the results, the Nature database had the maximum accuracy in the results, and the "Agent-based modeling" had the maximum call in Wiley's database. It shows applied Big Data methodology in papers of Nature could be better predicted, and the other papers of this database are more consonant in Big Data methodology. Moreover in papers of Wiley database there was no papers with "Agent-based modeling" methodology which its methodologies predicted false.

Future researchers can be utilized more refine the strategy to give more precision and manage some other issue like regional health systems, or do the work on more databases and content (like as books). Additionally, build the span of the testing dataset and can look at the more brand of the cellular telephone as a huge number of versatile brand are accessible to the market while this study can be performed on more labels and apply classification. In addition, the future studies can exanimate the predictions of this study.

**Abbreviations**

ICT: information and communication technologies; IEEE: The Institute of Electrical and Electronics Engineers; ACM: Association for Computing Machinery; ASP: American Scientific Publishers; AIP: American Institute of Physics; ASME:

The American Society of Mechanical Engineers; AMS: American Mathematical Society; WEKA: Waikato Environment for Knowledge Analysis; NHS: National Health Service; OECD: The Organization for Economic Co-operation and Development; UK: United Kingdom.

**Authors' contributions**
FSR has contributed for acquisition of data, analysis and interpretation of data, drafting of the manuscript. ARG has served as the advisor in study conception, and for critical revision. RR has critically reviewed the study proposal and for design. All authors read and approved the final manuscript.

**Author details**
[1] Department of Information Technology Management, Science and Research Branch, Islamic Azad University Tehran, Tehran, Iran. [2] Department of Management, Tarbiat Modares University, Tehran, Iran.

**Acknowledgements**
None.

**Competing interests**
The authors declare that they have no competing interests.

**Availability of data and materials**
Results of searching in mentioned scientific databases.

**Funding**
None.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 30 October 2018 Accepted: 22 December 2018
Published online: 31 January 2019

**References**

1. Roozbahani FS, Hojjati SN, Azad R. The role of E-payment tools and E-banking in customer satisfaction case study: Pasargad bank E-payment company. Int J Adv Netw Appl. 2015;7(2):2640–9.
2. Poorebrahimi A, Roozbahani FS. Effects of security and privacy concerns on using of cloud services in energy industry, an oil and gas company: a case study. Int J Adv Netw Appl. 2015;7(3):2779–83.
3. Skoric MM. The implications of big data for developing and transitional economies: extending the Triple Helix? Scientometrics. 2013;99(1):175–86.
4. Roozbahani FS, Azad R. Security solutions against computer networks threats. Int J Adv Netw Appl. 2015;7(1):2576–81.
5. Roozbahani FS, Barjouei RS, Hojjati SN. Identifying an appropriate model for information systems integration in the oil and gas industry. Int J Adv Netw Appl. 2018;10(1):3687–91.
6. Poorebrahimi A, Razavi F, Roozbahani FS. Presenting VALIT frameworks and comparing between them and other enterprise architecture framework. Adv Netw Appl. 2016;7(4):2805–9.
7. Nobre GC, Tavares E. Scientific literature analysis on big data and internet of things applications on circular economy: a bibliometric study. Scientometrics. 2017;111(1):463–92.
8. How Big Data is transforming every business, in every industry. Bernard Marr & Co. https://www.bernardmarr.com/default.asp?contentID=767. Accessed 28 May 2018.
9. Esselink J. Today Big Data and analytics is everywhere for everyone. IBM, 03.08.2017. https://www.ibm.com/blogs/think/nl-en/2017/03/08/today-big-data-analytics-everywhere-everyone/. Accessed 28 May 2018.
10. Dull T. I see big data. All the time. It's everywhere. SAS. https://www.sas.com/en_us/insights/articles/data-management/i-see-big-data.html. Accessed 28 May 2018.
11. Adrianto B. The influence of Big Data implementation towards business models in different sectors (Master Thesis). Delft University of Technology. 2017.
12. Li J, Ding W, Cheng HK, Chen P, Di D, Huang W. A comprehensive literature review on Big Data in healthcare. 2016.
13. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst. 2014;2(1):1–10.
14. Fichman RG, Kohli R, Krishnan R. Editorial overview—the role of information systems in healthcare: current research and future trends. Inf Syst Res. 2011;22(3):419–28.
15. Toga AW, Dinov ID. Sharing big biomedical data. J Big Data. 2015;2:7.
16. Abouelmehdi K, Beni-Hessane A, Khaloufi H. Big healthcare data: preserving security and privacy. J Big Data. 2018;5(1):1–18.
17. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. J Big Data. 2014;1:2.
18. van Altena AJ, Moerland PD, Zwinderman AH, Olabarriaga SD. Understanding big data themes from scientific biomedical literature through topic modeling. J Big Data. 2016;3:23.
19. Adams J, Light R. Mapping interdisciplinary fields: efficiencies, gaps and redundancies in HIV/AIDS research. PLoS ONE. 2014;9(12):1–13.

20. Alan P, Alex C, David Roessner J, Marty P. Measuring researcher interdisciplinarity. Scientometrics. 2007;72(1):117–47.
21. Van Raan A. Scientometrics: state-of-the-art. Scientometrics. 1997;38(1):205–18.
22. Big Data. Gartner, 2012. https://www.gartner.com/it-glossary/big-data/. Accessed 2018.
23. Kwon O, Sim JM. Effects of data set features on the performances of classification algorithms. Expert Syst Appl. 2013;40(5):1847–57.
24. McAfee A, Brynjolfsson E, Davenport TH, Patil DJ, Barton D. Big data: the management revolution. Harv Bus Rev. 2012;90(10):60–8.
25. Russom P. The three vs of Big Data analytics. TDWI BLOG. 2011.
26. White M. Digital workplaces: vision and reality. Bus Inf Rev. 2012;29(4):205–14.
27. DeVan A. The 7V's of Big Data. Impact, 7 Apr 2016. https://impact.com/marketing-intelligence/7-vs-big-data/. Accessed 10 Dec 2018.
28. Pant P, Tanwar R. An overview of Big Data opportunity and challenges. In: Smart trends in information technology and computer communications. 2016. pp. 691–7.
29. Jatrniko W, Arsa DMS, Wisesa H, Jati G, Ma'sum MA. A review of big data analytics in the biomedical field. In: Big Data and information security (IWBIS), international workshop on IEEE. 2016. pp. 31–41.
30. Sahoo PK, Mohapatra SK, Wu SL. Analyzing healthcare Big Data with prediction. IEEE Access. 2016;4:9786–99.
31. Wan K, Alagar V. Characteristics and classification of Big Data in health care sector. In: Natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD), 2016 12th international conference on. IEEE. 2016.
32. Ta VD, Liu CM, Nkabinde GW. Big Data stream computing in healthcare real-time analytics. In: Cloud computing and Big Data analysis (ICCCBDA), international conference on. IEEE. 2016. pp. 37–42.
33. Viceconti M, Hunter PJ, Hose RD. Big data, big knowledge: big data for personalized healthcare. IEEE J Biomed Health Inform. 2015;19(4):1209–15.
34. OECD Indicators. Health at a Glance 2011. Paris: OECD Publishing; 2011.
35. Hermon R, Williams PAH. Big Data in healthcare: what is it used for?. In: Australian eHealth informatics and security conference. 2014.
36. Al-Khoder A, Harmouch H. Evaluating four of the most popular open source and free data mining tools. Int J Acad Sci Res. 2015;3(1):13–23.
37. Rangra K, Bansal KL. Comparative study of data mining tools. Int J Adv Res Comput Sci Softw Eng. 2014;6:4.
38. Naive Bayes. rapidminer Documentation. https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/bayesian/naive_bayes.html. Accessed 31 Jul 2018.
39. Zhang H. The optimality of naive Bayes. AA. 2004;1(2):3.
40. Friedman JH. On bias, variance, 0/1—loss, and the curse-of-dimensionality. Data Min Knowl Discov. 1997;1(1):55–77.
41. Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss, Domingos, Pedro, and Michael Pazzani. Mach Learn. 1997;29(2–3):103–30.
42. McCallum A, Nigam K. A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization, vol. 752, no. 1. 1998.
43. Robertson SE, Sparck Jones K. Relevance weighting of search terms. J Am Soc Inf Sci. 1976;27(3):129–46.
44. Lewis DD. An evaluation of phrasal and clustered representations on a text categorization task. In: Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval. ACM. 1992.
45. Kalt T, Croft WB. A new probabilistic model of text classification and retrieval. Technical Report IR-78, University of Massachusetts Center for Intelligent Information Retrieval. 1996.
46. Jacob. Text classification for sentiment analysis—precision and recall. Streamhacker, 17 May 2010. https://streamhacker.com/2010/05/17/text-classification-sentiment-analysis-precision-recall/. Accessed 31 Aug 2018.



**Submit your manuscript to a SpringerOpen journal and benefit from:**

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

**Submit your next manuscript at ► springeropen.com**