(Bio)medical Publications in the Age of Big Data: Yes, They Are Different

Allard Jan-Jaap van Altena
Amsterdam Medical Center - University of Amsterdam
a.j.vanaltena@amc.uva.nl

Sílvia Delgado Olabarriaga
Amsterdam Medical Center - University of Amsterdam
s.d.olabarriaga@amc.uva.nl

CCS CONCEPTS

• **Applied computing** → **Life and medical sciences**; • **Information systems** → *Data mining*; *Clustering*; • **Computing methodologies** → Classification and regression trees;

KEYWORDS

Big Data, Topic Modelling, Biomedical Literature

ACM Reference format:
Allard Jan-Jaap van Altena and Sílvia Delgado Olabarriaga. 2017. (Bio)medical Publications in the Age of Big Data: Yes, They Are Different. In *Proceedings of DH '17, London, United Kingdom, July 02-05, 2017,* 2 pages.
https://doi.org/10.1145/3079452.3079474

1 INTRODUCTION

In 2011 the term 'Big Data' was introduced by Gartner [5], and since then its use in literature has ever increased, also in the (bio)medical research field [1]. Although the term Big Data is widely used, studies show that its meaning is much debated and many different definitions exist [10]. This variety of definitions may lead to different understandings and therefore difficulties in communication. For example, a researcher that is looking for 'Big Data' solutions might miss an interesting method that is not tagged as such.

In previous work we studied major topics that appear in Big Data literature using a Topic Modelling approach [8]. However, from that study it was not possible to know whether those topics are exclusive to publications self-identified as Big Data (BD), or not. Therefore, here we investigate the research question: What are the differences between topics in BD and non-Big Data (NBD) corpora?

2 METHODS

Figure 1 shows steps taken in this study. Firstly, the BD and NBD corpora were constructed by searching PubMed and PubMed Central. For the BD corpus search queries were constructed to search for the literal use of "Big Data" after 2011 (i.e., after introduction of the term 'Big Data' by gartner). Searches for the NBD corpus were performed for each journal and restricted to the same time period as the publications in the BD corpus. The BD+NBD corpus was created by matching on journal with a 1:2 ratio. BD publications without matching NBD publications were excluded.

DH '17, July 02-05, 2017, London, United Kingdom

ACM ISBN 978-1-4503-5249-9/17/07.
https://doi.org/10.1145/3079452.3079474

Text analysis was implemented using various packages in R. For extracting features from the corpora, the text mining method Topic Modelling (TM) was applied [2]. TM was chosen because it provides richer features (i.e., topics) than the frequency of words in the corpus alone. For more details on TM, pre-processing, and post-processing please refer to our previous paper [8].

To determine the difference between the BD and NBD publications, the TM topics were used as features to the machine learning method 'Random Forest' (RF) [3]. RF was applied because of its suitability for classifying binary outcomes (i.e., BD or NBD), it was implemented in R using the `caret` package [6, 7]. The performance of RF results was assessed using the F_1-measure [9]. The RF outcome was interpreted using the `varImp` function, which ranks the input features based on their importance in the RF model. For each RF model the two most important features (i.e., topics) were selected for further analysis. From these two topics, the top-30 relevant words (according to the TM) were selected and analysed through a wordcloud.

For assessing the stability of the obtained results, we considered variations in the number of topics in the TM, cross validations for the RF, and variations in the corpus. Multiple corpora were created to get a larger representation and to test the stability of the resulting TM and RF models. Four TMs were built to assess the influence of the number of topics, and lastly, the RF model was cross-validated using 10 folds.

3 RESULTS

The publication search yielded 2,176 results from PubMed and 745 from PubMed Central. From these, 542 empty abstracts and 629 duplicates were removed. Furthermore, in total 32 journals (277 papers) were excluded from the corpus because they were not focused on (bio)medical research. After cleaning 1,473 results remained in the BD corpus with 747 distinct journals. By matching on journal and publication year of the BD corpus, 14,146 results were found on PubMed, of which 12,263 remained after removing empty abstracts and duplicates. Four BD+NBD corpora were constructed, including a total of 10,179 publications, thereby covering 74% of all BD and matched NBD publications that were found.

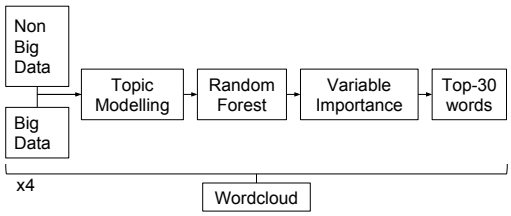

Figure 1: Overview of data processing pipeline.

A total of sixteen TMs were fitted, which were used as input to sixteen runs of the RF method respectively. For each RF run, the average F_1-measure over the 10 cross-folds was measured, resulting in a mean of 0.72, standard deviation: 0.015. Each of the RF models showed a similar behavior of the variable importance curve.

Lastly, the final wordcloud is shown in Figure 2. From the 960 words, 365 were unique, with frequencies between 14 to 1 occurrences. The minimal frequency of words in the wordcloud is 3 and a maximum of 100 words are shown.

4 DISCUSSION & CONCLUSION

In this study we attempted to identify differences between a BD corpus and NBD corpus through text mining and machine learning methods. Publication inclusion in the corpus was designed to be as broad as possible. All (bio)medical relevant publications labelled with 'Big Data' in PubMed and PMC were included. However, some publications had to be excluded because the number of matching NBD publications was insufficient.

We succeeded to collect a large BD corpus ($n = 1,473$), therefore we expect that missing BD publications have a small impact in the major outcomes and conclusions from our study. Furthermore, although the NBD corpus ($n = 12,263$) can be considered small compared to the complete set of papers in PubMed, these were most of the available matching papers for the BD corpus.

The measured outcomes (i.e., F_1-measure and variable importance curves, data not shown) indicate that the TMs and RFs are robust when varying the input parameters (e.g., different BD+NBD corpora, varying number of topics, and cross-folds).

Visual inspection of the wordcloud shows that a couple of discriminating topics include words related to themes such as big data, the research field, and buzzwords. Words such as technologies, large, and computational were also found in our previous study and can be associated with existing Big Data definitions [8]. Furthermore, various research fields are represented, such as: bioinformatics, neuroscience, and epidemiology. This suggests that Big Data is more commonly identified in these fields of application. Also note that words such as opportunities, challenges, and era can be interpreted as buzzwords, confirming that the Big Data term is also associated with a hype.

Figure 2: Wordcloud combining top-30 words in most discriminating topics from 16 RF models. Size and color denote word frequency.

To our knowledge, no previous study has investigated the differences between Big Data publications with other comparable research publications. Nevertheless, prior studies have researched the definition of Big Data [4, 10]. Unlike those studies, however, here no semi-formal definitions are used to gather the BD corpus. Instead, PubMed queries were used to find publications self-identified as Big Data. Furthermore, as compared to previous qualitative studies [4, 10], more publications could be assessed here due to the adoption of quantitative text mining and machine learning methods.

Our results show that a difference between BD and NBD publications can be detected using text mining and machine learning methods with a fair amount of precision and recall (F_1-measure of 0.72 on average). This performance is stable throughout a large variety of set-ups. Furthermore, by analysing the most discriminating features in the RF model (Figure 2), some insight is gained into what kind of differences exist.

In future work we plan to further investigate the meaning and uses of 'Big Data'. While this paper stopped at finding the differences between BD and NBD we plan to qualitatively analyse those differences. Furthermore, we are investigating the use of Big Data methods in medical research — specifically systematic reviews — and the barriers and facilitators that occur.

Acknowledgements: This work was carried out on the High Performance Computing Cloud resources of the Dutch national e-infrastructure with the support of SURF Foundation.

The authors have declared that no competing interests exist.

REFERENCES

[1] Javier Andreu-Perez, Carmen CY Poon, Robert D Merrifield, Stephen TC Wong, and Guang-Zhong Yang. 2015. Big Data for Health. *IEEE Journal of Biomedical and Health Informatics* 19, 4 (7 2015), 1193–1208.
[2] David M Blei. 2012. Probabilistic Topic Models. *Commun. ACM* 55, 4 (April 2012), 77–84.
[3] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32.
[4] Andrea De Mauro, Marco Greco, and Michele Grimaldi. 2016. A formal definition of Big Data based on its essential features. *Library Review* 65, 3 (2016), 122–135.
[5] Jackie Fenn and Hung LeHong. 2011. *Hype cycle for emerging technologies, 2011.* Technical Report. Gartner.
[6] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, and Can Candan. 2016. *caret: Classification and Regression Training.* R package version 6.0-71.
[7] R Core Team. 2015. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.
[8] Allard J. van Altena, Perry D. Moerland, Aeilko H. Zwinderman, and Sílvia D. Olabarriaga. 2016. Understanding big data themes from scientific biomedical literature through topic modeling. *Journal of Big Data* 3, 1 (2016), 23.
[9] Cornelis Joost van Rijsenbergen. 1979. *Information Retrieval.* Buttersworth.
[10] Jonathan Stuart Ward and Adam Barker. 2013. Undefined By Data: A Survey of Big Data Definitions. *arXiv preprint arXiv:1309.5821* (2013).