Modelling the Names of Jeffrey Beall's List of Possible Predatory Journals

Yves Tillé[*]

Abstract

On his web site, Jeffrey Beall proposed a list of possible predatory journals. The names of these journals are compared to the source publication list for the Web of Science® of Thomson Reuter™. By means of the lasso statistical method and logistic regression, we identify a list of words of titles that enables to discriminate between both lists. The titles of the possible predatory journals contains more often vague terms as 'advanced', research', 'international' or 'sciences'. Some journal names are so characteristic that it is possible to almost predict with certainty that they are possible predatory ones.

Keywords: Lasso; Logistic regression; Publications; Web of Science

MSC 97K80 JEL C10

1 Beall's list of journals

Jeffrey Beall (2012, 2013) proposed a list standalone, 'potential, possible, or probable predatory scholarly open-access journals'. These journals are suspected to charge publication fees to the authors without respecting the usual academic standards. This list is based on criteria that are clearly defined in Beall (2015) (see also Mehrpour and Khajavi, 2014). The phenomenon has taken on considerable proportions (Xia et al., 2015) particularly in developing countries (Shen and Björk, 2015). Jeaffrey Beall's web site was suddenly removed in January 2017. Beall (2017b) wrote an opinion paper, in which he describes the different pressures of the listed journals and of the authors who publish in these journals. Some journals also contacted the university officials with numerous emails and letters. One of the last version of Beall's list is however still available in the Internet Archive (Beall, 2017a).

At a first glance, one might notice that the names of these journals are a little bit bombastic. For instance, the names of a large number of these journals begin with 'International journal of'. A challenging question consists of trying to predict that a journal belongs to Jeffrey Beall's list based solely on the words that are used in its name. In December 2016, Jeffrey Beall's list contained 1293 journal names. This list is compared to the 2016 source publication list for the Web of Science® of Thompson Reuter™ (2015) that contained 8761 journal names.

Our aim is to try to predict if a journal belongs either in the Jeffrey Beall's list or in the Web of Science® list by using only the words that are used in the journal names. Are some words typical of a possible predatory journal? In this paper, we try to answer to these questions and we propose a model that enables to estimate a probability that a journal belongs to Beall's list instead of the Web of Science® list.

2 Methods and Results

The analysis is restricted to the 45 words that appear at least 100 times in the journal names. A table $\mathbf{X} = (x_{ij})$ of occurrences has been constructed, where x_{ij} is the number of times that word number j appears

[*]Yves Tillé, University of Neuchâtel, Rue de Bellevaux, 51, 2000 Neuchâtel, Switzerland, Tel.: +41-32-7181475, yves.tille@unine.ch

in the name of journal i. For the statistical treatment, all the titles are put to uppercase. Punctuation and abbreviations in parenthesis have been removed. However, the character '&' has been left and is thus different from the word 'and'.

Each of the 45 words that occurs 100 times or more is an explanatory variable used in order to predict either the journal is in Beall's list or not. Since the number of explanatory variable is large, we used the lasso method proposed by Tibshirani (1996) with a logistic regression model. The lasso method enables at the same time to select the variables and to estimate the parameters. The 'glmnet' R package that contains an implementation of the lasso method has been used for the statistical treatment (Friedman et al., 2010; Simon et al., 2011). Penalization parameter λ has been estimated by cross-validation.

Table 1 contains the estimated regression coefficients of the words selected by the lasso method. The first column is dedicated to the negative coefficients and the last three columns to the positive ones. Occasionally a word can appear several times in the same journal names. The most characteristic words of the journals of Beall's list are 'research', 'international', 'sciences', 'journal', 'advanced'. These words say nothing specific about the content of the journals. They are just bombastic. "Management', 'computer' 'science', and engineering' seem to be disciplines where the predatory journals are more active. On the opposite, the words 'materials', 'society', 'systems' and 'ieee' increase the probability of being in the list of 'Web of Sciences'. The word 'society' often appears when the journal belongs to an academic society and IEEE means 'Institute of Electrical and Electronics Engineers' that is a recognized professional organization.

Tab. 1: Table of regression coefficients corresponding to each word obtained by the lasso logistic regression

Words	coef.	Words	coef.	Words	coef.	Words	coef.
intercept	-4.35						
materials	-1.27	american	0.02	letters	0.49	advanced	1.49
society	-0.70	acta	0.07	engineering	0.56	journal	1.71
systems	-0.70	the	0.19	current	0.58	sciences	1.76
ieee	-0.25	applications	0.20	computer	0.65	international	1.84
clinical	-0.24	european	0.34	and	0.72	research	1.93
surgery	-0.24	&	0.37	applied	0.79		
of	-0.10	de	0.39	advances	0.93		
on	-0.06	technology	0.40	science	1.04		
physics	-0.01	medical	0.44	review	1.22		
health	-0.003	in	0.45	management	1.26		

The value of the logit function of a journal is obtained by adding to the intercept the sum of the values of the coefficients multiplied by the number of times the corresponding words appears in the name. The probability of being in Beall's list is $prob = 1/(1 + \exp(-\text{logit}))$. If the logit is positive (or equivalently if $prob > 0.5$), we predict that the journal should belong Beall's list.

Table 2 shows that the model allows proper classification. For Beall's list, 87.8% of the predicted values are correct. For the list of Web of Science, 93.4% of the predicted values are correct. Nevertheless, 616 journals of Beall's list were not identified by the model. A little bit less than one half of the journals cannot be identified thanks to their names.

Tab. 2: Table with the number of journals in both lists with respect to the predicted values.

	Web of Science	Beall's list	Total
Predicted Web of Science	8667 (93.4%)	616 (6.6%)	9283 (100%)
Predicted Beall's list	94 (12.2%)	677 (87.8%)	771 (100%)

More than half of the journal names of the Beall list are so characteristic that they can directly be identified as belonging to this list based on the words of the name. A little less than half of the Beall

list journals remain unnoticed. The model could however be improved by adding other variables as the localization of the web site based on the IP address of the web site, the number of published paper, the age of the journal, the last volume number, the existence of an editorial board, and the affiliation of their members. A hard work is however necessary to collect this information.

3 Conclusions

The model does not enable to completely discriminate between the lists but it underscores the most important words that are typical of possible predatory journals. However the model provides a simple tool that enables to estimate a probability for a particular journal to be a possible predatory one and thus to know if a more complete investigation should be realized. If one wants to create a new journal, the model also enables to advise avoiding a name as *International Journal of Research and Review in Advanced Management Sciences and Advances in Applied Computer Engineering (IJRRAMSAACE)* unless to be directly suspected to be a predator.

References

Beall, J. (2012). Predatory publishers are corrupting open access. *Nature*, 489(7415):179–179.

Beall, J. (2013). Predatory publishing is just one of the consequences of gold open access. *Learned Publishing*, 26(2):79–84.

Beall, J. (2015). Criteria for determining predatory open-access publishers. Scholarly open access `https://web.archive.org/web/20150320190303/https://scholarlyoa.files.wordpress.com/2015/01/criteria-2015.pdf`,(accessed 2015-02-14).

Beall, J. (2017a). List of standalone journals: Potential, possible, or probable predatory scholarly open-access journals. Scholarly open access `https://web-beta.archive.org/web/20170103170852/https://scholarlyoa.com/individual-journals/`,(accessed 2017-06-14).

Beall, J. (2017b). What i learned from predatory publishers. *Biochemia Medica*, 27(2):273–278.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Mehrpour, S. and Khajavi, Y. (2014). How to spot fake open access journals. *Learned Publishing*, 27(4):269–274.

Shen, C. and Björk, B.-C. (2015). 'predatory'open access: a longitudinal study of article volumes and market characteristics. *BMC medicine*, 13(1):1.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13.

Thompson Reuter™ (2015). Source publication list for Web of Science®, science citation index expanded. Scholarly open access `http://ip-science.thomsonreuters.com/mjl/publist_sciex.pdf`,(accessed 2016-01-01).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 73(3):267–288.

Xia, J., Harmon, J. L., Connolly, K. G., Donnelly, R. M., Anderson, M. R., and Howard, H. A. (2015). Who publishes in predatory journals? *Journal of the Association for Information Science and Technology*, 66(7):1406–1417.