Predicting Research that will be Cited in Policy Documents

Bharat Kale
Northern Illinois University
bkale@niu.edu

Harish Varma Siravuri
Northern Illinois University
hsiravuri@niu.edu

Hamed Alhoori
Northern Illinois University
alhoori@niu.edu

Michael E. Papka
Argonne National Laboratory
Northern Illinois University
papka@niu.edu

ABSTRACT

Scientific publications and other genres of research output are increasingly cited in policy documents. Citations in documents of this nature could be considered a critical indicator of the significance and societal impact of the research output. In this work, we have built classification models that predict whether a particular research work is likely to be cited in a public policy document based on the attention it received online, primarily on social media platforms. We evaluated the classifiers based on their accuracy, precision and recall values. We found that the Random Forest classifier performed best.

KEYWORDS

Public Policy, Policy documents, Altmetrics, Social Media

ACM Reference format:
Bharat Kale, Harish Varma Siravuri, Hamed Alhoori, and Michael E. Papka. 2017. Predicting Research that will be Cited in Policy Documents. In *Proceedings of International ACM Web Science Conference, Troy, NY (USA), June 2017 (WebSci'17),* 2 pages.
https://doi.org/10.475/123_4

1 INTRODUCTION

Policy documents by their nature influence large sections of society [4]. By virtue of the nature and impact of these influential documents, the citations they include both support the stated policy and strengthen the authority of the authors cited [6]. Because of the unique importance of policy documents across diverse organizations, citations included in this type of material bear more weight than crediting an author by supporting the credibility of the policy document itself. Likewise, in this context, it may be appropriate to assign a policy document citation more weight than one included in a literature review in a scholarly paper, for example.

Haunschild and Bornmann [7] studied the percentage of papers in Web of Science that are mentioned in policy related documents and found that less than 0.5% of the papers in different subjects have been mentioned at least once in policy related documents. Lauren [5] analyzed patterns in the types of altmetric attention received by papers that make it into policy documents and found that the inclusion into policies is happening more quickly, within 2 years of publication, thereby having a real world impact sooner.

Winterfeldt [12] presented a framework to bridge the gap between science and decision making in the policy sphere. Orduna-Malea, Thelwall, and Kousha [7] explored the relationship between citations in patents and technological impact and found that the number of patents citing a resource indicates the technological capacity or relevance of that resource. Black [3] concluded that although evidence-based policy-making is being encouraged in all areas of public service, research is currently under-used in policy-making and that there is a need for a better understanding between research and policy communities.

Citation analysis is self-limiting because it ignores many other signals through which research receives attention. An increasing amount of scholarly content is being shared and discussed daily on social media platforms [1]. Whereas citations measure research impact within scholarly boundaries, non-traditional web-based metrics or altmetrics [8][2] provide the ability to measure different influences, including readers who share, read, or discuss an article with others, but do not formally cite it within traditionally published articles.

Thelwall et al. [11] studied the potential value of altmetrics for evaluating funding criteria and found that some metrics could be helpful in this sphere. Sarewitz and Pielke [10] suggested a method to improve the connection between science policy decisions, science, and social outcomes using the example of climate change research. Pawson [9] discussed various ways to incorporate results from research into the policy-making process. To date, the focus of most studies is on understanding and using altmetrics in reference to only a few measures, but modeling altmetrics for predicting citations in policy documents has not been explored.

2 DATA COLLECTION

The dataset in this study is a database dump that we obtained from altmetric.com, which consists of 5.2 million articles. Our initial analysis showed that 89,350 articles have at least one policy citation and 5,097,207 articles have no citation in any policy document. To create a balanced dataset for further analysis, along with the 89,350 articles that have been cited in policy documents, we randomly chose another 89,350 articles that did not receive any citation in policy documents. The result was a balanced dataset with approximately 180,000 records, half of them being cited in policy documents.

WebSci'17, June 2017, Troy, NY (USA)

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

3 FEATURE SELECTION

The dataset has a very rich set of features for each article but for our analysis, we have considered only the features related to online attention. The dataset consists of mention counts on various online sources including reference managers, mainstream news outlets, blogs, peer-review platforms (e.g., PubPeer and Publons), social media, public policy documents, and Wikipedia.

We used mention counts in Twitter, Facebook, Reddit, Mendeley, Google+, Wikipedia, Weibo, mainstream news outlets, blogs, videos, and peer-review paltforms as features to build the classifiers. A few sources were not considered. "Connotea" that has been discontinued since 2013 and two other sources, "Pinterest" and "Stackoverflow" contributed to less than 1% of the articles in the sample. We replaced the policy citation count with a binary class label denoting whether or not the article had been cited in policy documents.

4 METHODS AND RESULTS

4.1 Classification

To predict the likelihood of a research article being cited in a policy document, we implemented three classifiers which include Multinomial Naive Bayes, Random Forest with number of trees set at 100, and a C-Support Vector Machine with the Radial Basis Function (RBF) kernel. We then divided the entire dataset into training and test sets comprising of 80% and 20% of the entire dataset respectively. The models were trained using 10 fold cross validation technique and evaluated based on the accuracy, precision, recall, and F1-measure metrics as shown in Table 1.

Table 1: Accuracy, Precision, Recall and F1-Measure for different models

	Multinomial Naive Bayes	Random Forest	SVM
Accuracy	0.842	0.870	0.868
Precision	0.802	0.826	0.820
Recall	0.905	0.870	0.868
F1-Measure	0.850	0.844	0.824

4.2 Feature Ranking

With the classification models built, we calculated weights for each feature to determine their significance in making the final prediction. Since feature weights in the case of Support Vector Machines can only be determined for linear kernels, we ranked features based on their relevance only to the Random Forest and Multinomial Naive Bayes classifiers. We ranked the features in decreasing order of their importance to the Random Forest classifier as shown in Table 2.

5 CONCLUSIONS AND FUTURE WORK

In this work we used a specific set of features that track online attention received by scholarly articles to build classifiers that predicted the likelihood of an article being cited in public policy. The Random Forest classifier showed better results in making predictions. We found that mention counts in peer-review platforms to be the most influential feature while news was the least influential feature.

Table 2: Feature ranking for different models

Platform	Random Forest	Multinomial Naive Bayes
peer-review	0.273595	4.4267
Google+	0.197488	3.4210
Reddit	0.151016	4.4087
video	0.098035	4.9458
Twitter	0.068745	2.2421
Weibo	0.088242	3.7988
Mendeley	0.030116	0.3210
Wikipedia	0.026027	4.9668
blogs	0.018631	4.4571
Facebook	0.016189	3.2314
news	0.008926	3.7307

The promising results obtained in this work show that a relationship exists between the online attention that scholarly work receives and the policy citations they generate, which we were able to exploit. We intend to build upon this work and build regression models to predict the number of policy citations a particular work is likely to receive. We also plan to build more classifiers with different feature sets and compare our results.

ACKNOWLEDGMENTS

MEP was supported in part by the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract DE-AC02-06CH11357.

REFERENCES

[1] Euan Adie and William Roe. 2013. Altmetric: enriching scholarly content with article-level discussion and metrics. *Learned Publishing* 26, 1 (2013), 11–17. https://doi.org/10.1087/20130103

[2] Hamed Alhoori and Richard Furuta. 2014. Do altmetrics follow the crowd or does the crowd follow altmetrics?. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*. 375–378. https://doi.org/10.1109/JCDL.2014.6970193

[3] Nick Black and Anna Donald. 2001. Evidence based policy: proceed with careCommentary: research must be taken seriously. *BMJ* 323, 7307 (2001), 275–279. https://doi.org/10.1136/bmj.323.7307.275

[4] Lutz Bornmann, Robin Haunschild, and Werner Marx. 2016. Policy documents as sources for measuring societal impact: how often is climate change research mentioned in policy-related documents? *Scientometrics* 109, 3 (2016), 1477–1495. https://doi.org/10.1007/s11192-016-2115-y

[5] Lauren Cadwallader. 2016. Papers, policy documents and patterns of attention. In *3:AM The Altmetrics Conference*. https://doi.org/10.17863/CAM.4844

[6] Richard Freeman and Jo Maybin. 2011. Documents, practices and policy. *Evidence & Policy: A Journal of Research, Debate and Practice* 7, 2 (2011), 155–170.

[7] Robin Haunschild and Lutz Bornmann. 2017. How many scientific papers are mentioned in policy-related documents? An empirical investigation using Web of Science and Altmetric data. *Scientometrics* 110, 3 (2017), 1209–1216. https://doi.org/10.1007/s11192-016-2237-2

[8] Kim Johan Holmberg. 2015. *Altmetrics for information professionals: Past, present and future*. Elsevier. https://books.google.com/books?id=GhdiBQAAQBAJ

[9] Ray Pawson. 2002. Evidence-based Policy: In Search of a Method. *Evaluation* 8, 2 (2002), 157–181. https://doi.org/10.1177/1358902002008002512

[10] Daniel Sarewitz and Roger A Pielke. 2007. The neglected heart of science policy: reconciling supply of and demand for science. *Environmental Science & Policy* 10, 1 (2007), 5–16. https://doi.org/10.1016/j.envsci.2006.10.001

[11] Mike Thelwall, Kayvan Kousha, Adam Dinsmore, and Kevin Dolby. 2016. Alternative metric indicators for funding scheme evaluations. *Aslib Journal of Information Management* 68, 1 (2016), 2–18. https://doi.org/10.1108/AJIM-09-2015-0146

[12] Detlof von Winterfeldt. 2013. Bridging the gap between science and decision making. *Proceedings of the National Academy of Sciences* 110, Supplement 3 (2013), 14055–14061. https://doi.org/10.1073/pnas.1213532110