Exploring Features for Predicting Policy Citations

Christian Bailey
Northern Illinois University
cbailey11@niu.edu

Bharat Kale
Northern Illinois University
bkale@niu.edu

Jamieson Walker
Northern Illinois University
jwalker21@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 and
Northern Illinois University
papka@niu.edu

ABSTRACT

In this study we performed an initial investigation and evaluation of altmetrics and their relationship with public policy citation of research papers. We examined methods for using altmetrics and other data to predict whether a research paper is cited in public policy and applied receiver operating characteristic curve on various feature groups in order to evaluate their potential usefulness. From the methods we tested, classifying based on tweet count provided the best results, achieving an area under the ROC curve of 0.91.

KEYWORDS

Altmetrics, Social Media, Public Policy

ACM Reference format:
Christian Bailey, Bharat Kale, Jamieson Walker, Harish Varma Siravuri, Hamed Alhoori, and Michael E. Papka. 2017. Exploring Features for Predicting Policy Citations. In *Proceedings of ACM/IEEE Joint Conference on Digital Libraries, Toronto CA, June 2017 (JCDL2017),* 2 pages.
https://doi.org/10.475/123_4

1 BACKGROUND

1.1 Introduction

The growth of social media in the academic community has enabled scholars to develop new methods to evaluate the impact of research. Historically, evaluations of the impact of research have been limited to the reception by the scholarly community. However, with the advent of altmetrics we are able to track the social impact of research [8][2] . For example, Ding et al. [3] explored the use of social media tagging as it relates to scholarly works.

Policy documents have a vital role in generating demand for scientific innovation [4]. Haunschild and Bornmann [5] study the relation between Web of Science fields and the researches' use in public policy and found that less than 2% of every category is cited in public policy. Orduna-Malea, Thelwall, and Kousha [6] 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. Winterfeldt [9] presented a framework to bridge the gap between science and decision making in the policy sphere.

To better understand the possibilities of altmetrics, we conducted an exploratory study to determine the potential of social media data for creating valuable models to describe the use of research articles in public policy.

JCDL2017, June 2017, Toronto CA

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

1.2 Collection

The primary source of data for our analysis was a database dump from altmetric.com [1]. The dataset, which is from June 4th 2016, consists of 5.2 million articles. We separated articles into 2 classes: papers cited in public policy documents and papers not cited in such documents. Policy documents, as used in this paper and by altmetric.com, currently includes mostly policy published by medical organizations. The dataset comprised 89,350 research articles referenced in policy documents and 5,097,207 articles not referenced in that context.

We drew on the altmetric.com dataset for meta-info about each research paper, specifically journal, publisher, and Scopus subject information, as well as social media activity. To augment our dataset, we collected the citation counts for articles of interest. Initially, we collected citation counts for all the policy documents and 120,000 of the non-policy documents from the Thomson Reuters Web of Science. In addition, we collected journal impact factors for 9,000 journals.

1.3 Feature Selection and Filtering

We selected four groups of features to evaluate: meta-info consisting of journal, publisher, and Scopus subject; social media information consisting of unique user mentions on Facebook, Google Plus, Twitter, Reddit, and Stackoverflow; traditional citation counts; and mention counts from several online platforms. We filtered the social media mentions to consider only those that had occurred before the policy citation of any given article. We treated the count features as simple numeric variables, whereas for meta-info and unique users we used a collection of boolean variables for each value.

2 METHODS

We used two methods to explore the efficacy of our selected feature sets for predicting policy citations. The first approach we used was to perform linear regression on our various feature sets, and the

second involved training classifiers and evaluating various classification metrics. In both instances, we used the Scikit-Learn [7] software package to develop our models.

2.1 Linear Regression and Correlation

To analyze the relations between our feature sets and policy citations, we used least squares regression and calculated the coefficients of determination (r^2) for several relationships. We related each of citation count, unique users, meta-info, and mention counts on a number of platforms including Twitter, Google Plus, Facebook, Wikipedia, Mendeley, blogs, and Stackoverflow to two targets: policy citation presence, i.e., a boolean target, and policy citation count, which is the number of policy documents citing an article.

We list the coefficient of determination for each relation in the Table 1. While overall, the correlations are relatively weak, we observe a wide range in values, so it is clear that some features have a stronger bearing on whether a document is cited in policy documents.

Table 1: r^2 for features regressed on Presence and Count

	Policy Presence	Policy Count
Citation Count	0.01118	0.01557
Unique Users	0.32163	0.15005
Meta-info	0.41748	0.27011
Blog Posts	0.04242	0.01449
Mendeley Readers	0.03977	0.03830
Wikipedia	0.01152	0.00617
Tweet Count	0.01460	0.00612
Facebook Posts	0.00771	0.00361
Google Plus	0.00265	0.00012
Stackoverflow	0.00150	0.00036

2.2 Classification

Table 2: Area under ROC curve with a 90% CI

	ROC AUC using Bernoulli NB
Citation Count	0.57 ± 0.0
Unique Users	0.71 ± 0.05
Meta-info	0.81 ± 0.04
Tweet Count	0.91 ± 0.01
Facebook Posts	0.62 ± 0.04
Blog Posts	0.54 ± 0.06
Mendeley Readers	0.54 ± 0.01
Wikipedia	0.52 ± 0.01
Google Plus	0.52 ± 0.01
Stackoverflow	0.50 ± 0.01

To determine which, if any, feature sets provide the best predictions for citations in policy, we evaluated Bernoulli Naive Bayes using unique user mentions, meta-info, citation count, and mention counts on platforms including several social media sites, Wikipedia, Mendeley, and blogs. Instead of the standard method of counting policy citations, we used binary labels for the purpose of classification: either cited in policy or not cited in policy.

We performed a ten-fold cross validation evaluation on each of the stated feature sets. From this evaluation, we determined the area under the receiver operating characteristic curve when classifying the presence of a policy citation with each feature set, as shown in Table 2. We found that citation and mention counts performed poorly. However, classifiers using unique users and journal meta-info performed better.

3 CONCLUSIONS AND FUTURE WORK

In this initial study, we took a handful of alternative metrics and classic metrics for research papers to examine how they relate to the use of scholarly research in policy documents. We found that citations to be very poor at predicting research use in public policy. However, altmetrics grouped in specific ways, such as unique users and meta-info, show better potential.

With this study, we have discovered some promising directions for additional research and discounted a handful of other avenues of study. Moving forward, we plan to develop a more in-depth study of correlations between altmetrics and the use of research in public policy. We plan to focus on applying clustering algorithms and methods to altmetrics to determine which clusters produce the most accurate predictions. We will continue to compare our results from working with altmetrics to similar tests based on more classic ways to evaluate research articles and determine the value of given scholarly research papers. We plan to develop a model that will provide accurate and timely predictions pertaining to whether any given scholarly research will be credited in public policy documents.

4 ACKNOWLEDGEMENTS

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.
[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 (JCDL '14)*. IEEE Press, Piscataway, NJ, USA, 375–378. http://dl.acm.org/citation.cfm?id=2740769.2740833
[3] Ying Ding, Elin K. Jacob, Zhixiong Zhang, Schubert Foo, Erjia Yan, Nicolas L. George, and Lijiang Guo. 2009. Perspectives on social tagging. *Journal of the American Society for Information Science and Technology* 60, 12 (2009), 2388–2401. https://doi.org/10.1002/asi.21190
[4] Jakob Edler and Luke Georghiou. 2007. Public procurement and innovationâĂŤResurrecting the demand side. *Research Policy* 36, 7 (sep 2007), 949–963. https://doi.org/10.1016/j.respol.2007.03.003 arXiv:arXiv:1011.1669v3
[5] Robin Haunschild and Lutz Bornmann. 2016. How many scientific papers are mentioned in policy-related documents? An empirical investigation using Web of Science and Altmetric data. *Scientometrics* (2016), 1–8.
[6] Enrique Orduna-Malea, Mike Thelwall, and Kayvan Kousha. 2017. Web citations in patents: Evidence of technological impact? *Journal of the Association for Information Science and Technology* (2017). https://doi.org/10.1002/asi.23821
[7] F. Pedregosa, G. Varoquaux, A. Gramfort, and others. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12 (2011), 2825–2830.
[8] Jason Priem, Heather A. Piwowar, and Bradley M. Hemminger. 2012. Altmetrics in the wild: Using social media to explore scholarly impact. *CoRR* abs/1203.4745 (2012). http://arxiv.org/abs/1203.4745
[9] Detlof von Winterfeldt. 2013. Bridging the gap between science and decision making. *Proceedings of the National Academy of Sciences* 110, Supplement_3 (aug 2013), 14055–14061. https://doi.org/10.1073/pnas.1213532110