Predicting Research that will be Cited in Policy Documents

Rahul Pothireddy
Northern Illinois University
rpothireddy1@niu.edu

Aravind Muvva
Northern Illinois University
amuvva1@niu.edu

ABSTRACT

Scientific publications and other genres of research output are increasingly being cited in policy documents. Citations in documents of this nature could be considered a critical indicator of the significance and peak societal impact of the research output. In this study, we built regression 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 plat- forms. We evaluated the regression models based on their accuracy, precision, recall and accuracy values. We found that Random Forest and Polynomial regression of degree 3 performed better overall.[1]

KEYWORDS

Public Policy, Policy documents, Altmetrics, Social Media, Precision, Recall, F1 measure, R^2 Values, mean square error

1 INTRODUCTION AND PREVIOUS WORK

Policy documents make a huge and intensified social impact on major parts of the society. The ultimate significance and importance of policy documents throughout various companies, organizations, determines the usefulness and importance of predicting the rate at which the citations are being made.. The importance of author as well as the importance of the policy citations are equally focused. Likely, in this scenario, appropriate assigning of a policy document citation more weight than a regular citation included in a literature review of scholarly paper.

Haunschild and Bornmann 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 on a range of subjects had been mentioned at least once in policy-related documents. Charles Thomas analyzed patterns in the types of altmetric attention received by papers that make it into policy documents and found that papers are often being referenced quickly, i.e., within 2 years of publication, such that they are having a real-world impact at a highly increasing rate.

Orduna- Malea, Thelwall, and Kousha 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.

There is always an increasing amount of scholarly content that has been shared regularly on social media platforms like Quora, Discosession etc. Whereas citations measure research impact within scholarly boundaries,web-based metrics or altmetrics make it possible to measure different infiuences, including readers who read an article or share, and/or discuss it with others, but do not formally cite it in tradition- ally published articles.

One of the great scientist Karl Landsteiner utilized the peak value of Altmetrics for evaluating funding criteria and found that some metrics can be helpful in this sphere. Sarewitz and Roger Pielke proposed a method to strengthen the connection between science policy decisions, scientific research, and social outcomes using the example of climate change research. To date, most studies focus on understanding and using altmetrics in reference to only a few measures. The present study is the first to explore modeling altmetrics in order to predict citations in policy documents using regression models. Our study inturn helps in predicting the the rate of policy documents citations being cited on social media and accordingly rank them as well as identify its importance.

2 DATASET COLLECTION

The dataset in this study is a database dump that we obtained from altmetric.com, which consists of 5.2 million articles[3]. Initially analysis showed that of these articles, 89,350 had been cited in at least one policy document whereas 5,097,207 had not been included in a document of this kind. To create a balanced dataset we have taken into consideration, along with the 89,350 articles that had been cited in a policy documents, another 89,350 articles that had not been cited in a policy document.The result was a balanced dataset with approximately 180,000 records, half of which had been cited in policy documents. Our dataset has features that mainly seeks online attention and for the further processing, the dataset has been divided into 80% test set and 20% training-set.

3 FEATURE SELECTION AND DATA PREPROCESSING

The dataset has a very rich set of features for each article. However, in our analysis, we considered only features related to online attention. [4].The dataset consists of mention counts on various online sources including reference managers, mainstream news outlets, blogs, peer-review platforms , social media , public policy and Wikipedia.

We used mention counts on Twitter, Facebook, Reddit, Mendeley, Google+, Wikipedia, Weibo, mainstream news outlets, blogs, videos, and peer review sites as features to build the regressors. We ignored a few sources out of our account, including Connotea, which was discontinued in 2013, and Pinterest and Stackoverfiow, which together contributed to less than 1% of the articles in the sample.[6] After selecting the features we have then performed data preprocessing i.e, data cleaning where we have handle the missing values. The missing values are filled with 0's . This is done to avoid the Not a Number (NaN) error.

4. REGRESSION MODELS AND RESULTS

4.1 Regressors

To predict the citations in a research article being cited in a policy document and to test the usability of features a multiple regression models were employed.[7] we have implemented five regressors: the Multiple Linear regression, the Random Forest regression, logistic regression,[8] Polynomial regression of degree 2 and degree3.[9] We have calculated the average values of precision, recall and f1 measure for all the regression models and the similar among them are written below. We build prediction graphs for each feature vs policy.

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

	Log Reg	Pol deg 3	Rand. For
Precision	0.75	0.79	0.84
Accuracy	81.8%	83.2%	88%
F1 meas.	0.78	0.81	0.86
Mean Error	0.47	0.49	0.34
R^2 value	0.35	0.50	0.53

3.1 Features Ranking and Regressor Values

The regressor models we have built, along with the values of mean square error, precision, recall, f1 measure and r^2 values,. Random forest is said to be the best regressor which has the least mean square error value, quite high r^2 value along with quite compatible values of precision, recall and f1 measure. Based on the correlation coefficient, we ranked the features in regard to their importance with peer_review being the most popular one and new being the least influential feature. [8]

4 CONCLUSIONS AND FUTURE WORK

In this study, we used a specific set of features that track online attention received by scholarly articles to build classifiers and regressions to predict the likelihood of an article being cited in public policy. The Random Forest classifier and Random Forest

Table 2: Feature ranking for different models using correlation coefficients.

Features	Correlation coefficient
Peer_reviews	0.5641
Googleplus	0.0260
Reddit	0.1696
Video	0.3362
Twitter	0.0227
Weibo	0.0876
Mendeley	0.0268
Wikipedia	0.0931
Blogs	0.0686
Facebook	0.0682
News	0.0151

Regression showed better results in making predictions with high accuracy. We found peer_review as the most influential or popular feature. The accurate results obtained in this work Proves that dependency or a connection exists between the online attention that a scholarly work receives and the policy citations it generates. We intend to extend our work in this area by taking into consideration more and more features and build regression,[11] classification models to predict the number of policy citations a given work is likely to receive and then compare the results.

REFERENCES :

1. M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions.* New York: Dover, 1972.
2. L. Breiman. Random forests. *Machine Learning*, 1(45):5–32, 2001.
3. C. Chang and C. Lin. Libsvm: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm, detailed documentation at http://www.csie.ntu.edu.tw/ cjlin/papers/libsvm.ps.gz, 2001.
4. P. Christoffersen and F. Diebold. Further results on forecasting and model selection under asymmetric loss. *Journal of Applied Econometrics*, 11:561–571, 1996.
5. J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of 23rd International Conference on Machine Learning*, 2006.
6. P. Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining (KDD-99)*, pages 155–164. ACM Press, 1999.
7. C. Elkan. The foundations of cost-sensitive learning. In *Proc. of 7th International Joint Conference of Artificial Intelligence (IJCAI'01)*, pages 973–978, 2001.
8. P. Flach. The geometry of roc space: understanding machine learning metrics through roc isometrics. In *Proceedings of the 20th International Conference on Machine Learning*, 2003.
9. J. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–141, 1991.
10. David Meyer. *Support Vector Machines, the interface to libsvm in package e1071.* Technische Universitat Wien, Austria, 2002.
11. C. Van Rijsbergen. *Information Retrieval.* Dept. of Computer Science, University of Glasgow, 2nd edition, 1979.