Don't judge a journal by its cover?: Appearance of a journal's website as predictor of blacklisted open-access status

Lizhen Liang and Daniel Acuna

School of Information Studies

Syracuse University

lliang06@syr.edu, deacuna@syr.edu

Abstract

The nature of scientific research has motived an open-access model of publication supported by article processing fees. Under this rapidly evolving environment and financial incentives, some dubious venues would publish almost anything—for a fee. Many entities keep track of the standards of these new journals, "blacklisting" those deemed problematic. Anecdotal evidence suggests that blacklisted journals tend to have websites with subpar appearance (e.g., old web technologies, unprofessional design). In this work, we systematically explore whether this anecdotal evidence is true. In particular, we evaluate the websites of journals whitelisted and un-whitelisted by the Directory of Open Access Journals (DOAJ). We use a convolutional neural network to predict whether a journal is whitelisted based on a screenshot of its website and analyze the factors that predict one output vs. the other. Our results show that appearance is indeed a predictive factor, achieving a medium performance (AUC of 0.736). Further, our interpretation suggests that the network considers whitelisting those websites with a table of content, social media links and packed content. Conversely, our model mistakenly whitelists blacklisted journals hosted by Elsevier and blacklists whitelisted websites with sans fonts and non-Latin characters.

1 Introduction

Scientific knowledge is efficiently disseminated when there is no barrier to access and when it benefits everybody (Stephan, 2012). Critics of for-profit publishers such as Elsevier have proposed to have alternatives based on open-access scientific publication (Smith et al., 2016). This has led to the creation of many venues that publish scientific work and that, instead of asking payment from readers or institutions, they ask authors to pay (Solomon and Björk, 2012). While open access is in principle achieved by pre-print servers (Till, 2001), articles do not through peer review in these venues. Some evidence suggests that open access articles are more cited than non-open access ones (Eysenbach, 2006), further incentivizing scientists to prefer open access peer-reviewed journals. Unfortunately, predatory journals started to be created to take advantage of article processing fees. These predatory journals, as previous research has shown (Grudniewicz et al., 2019), have little or no guideline for publishing. Lacking transparency and deviating from the best editorial and publishing practices, articles published under a predatory journal may often include misleading information. Predatory journals disseminate false information that mislead both normal readers and researchers. Research has shown that 7% of academics in Australia have published in unethical journals (Downes, 2020). While in Italy, 5% of academics have published in predatory journals (Bagues et al., 2019).

It is hard to spot predatory journals at scale. A journal can simply be created using widely available web tools (Dadkhah et al., 2016) and sometimes they seem legitimate by listing (without permission) reputable experts as reviewers (Severin et al., 2020). A systematic way to discriminate between normal journals and predatory journals is to compare a journal with best practice of publishing provided by institutions like the Directory of Open Access Journals (DOAJ). Yet it is not explicit and also hard to automatically and efficiently detect predatory journals. There have been research on criteria and checklists about spotting predatory journals, but those methods also require manual selection. Thus detecting predatory journals using these checklists is time consuming.

In pursue of an automatic approach to detect predatory journal, in this work, a very obvious aspect of a journal is taken into account: the website's appearance of the journal. By training image and text classifiers on websites using whitelisted and blacklisted journals from DOAJ, we introduce a method to find predatory journals by a screenshot of their website, or in other words, "judge a journal by its website". We interpret the features learned from our models and discuss future extensions.

2 Curated list of whitelisted and blacklisted journals

Some institutions have curated blacklists or whitelists indicating whether journals are predatory. The Directory of Open Access Journals (DOAJ) provides a list of journals being whitelisted along with journals being removed from the list(DOAJ, 2020). Journals being listed in the DOAJ whitelist aligns its publication policies with the best practices of publishing defined by the directory. On the contrary, an "un-whitelisted" journal is a journal being removed from the whitelist for failing to comply with the guideline, including not having a dedicated page for the journal, bad homepage quality and the inclusion of commercial and ads. In this work, we consider those "un-whitelisted" journals inferior and regarded as predatory journals or "blacklisted".

These two lists contain the URL of their journal. We scrapped those websites with Python program and get a list of screenshots and source code for the corresponding URLs from the year 2019 using Archive.org's Wayback Machine. We were able to get screenshots and html source code of webpages for 262 journals, 85 of them un-whitelisted. We then build classifiers that predict whitelisted or un-whitelisted (predatory) status based on a screenshot or the source code of a website.

3 Analyzing the appearance of journals

3.1 Models for predicting based on web pages

We built two classifiers to predict whether a journal is whitelisted or unwhitelisted, one based on the HTML source code and another based on an screenshot image.

Source code classifier based on logistic regression. The source code classifier is based on HTML tags (e.g., link) and content of the tag (e.g., text with a link). We combine both sets of texts and compute the TF-IDF transformation of the text, which is a relatively straightforward way of transforming text into vector representation (Manning et al., 2008). We then build an elastic-net regularized logistic regression to predict a journal status.

Screenshot classifier based on convolutional neural network. We built an image classifier discriminating webpages by fine-tuning a pre-trained residual neural network model. Introduced in 2015, residual neural network has been widely used in image recognition because it enables effective deep neural network training (He et al., 2015). We took screenshots of each website and reprocessed them by reshaping each into the size accepted by the residual neural network. For fine-tuning, we froze the coefficients in the pre-trained model and replace the last layer of the feed-forward layer with a Sigmoid layer to predict the status of the journal.

3.2 Measuring performance

In this work, we use area under the curve (AUC) to measure the performance of both models. It is a reliable and consistent way of measuring the performance of classifiers by measuring how close true predictions are to the ground truth. A classifier with good performance generally have an AUC close to 1 while a model with bad performance normally has an AUC of nearly 0.

Since we have a relatively small sample size, it is important that we cross-validate properly. For both classifiers, we split 20% of the data set as testing set and 80% as training and validation set. For the source code classifier, we used 8-fold cross-validation using the training and validation set to find the best set of

Weight	Token	Weight	Token
0.33	`at`	-0.13	`fitplayback`
0.16	`div`	-0.10	`gs`
0.08	`span`	-0.08	`link`
0.06	`recommended`	-0.06	`v`
0.06	`sites`	-0.06	`br`
0.06	`class`	-0.06	`wb`
0.06	`a`	-0.05	`color`
0.05	`share`	-0.05	`script`
0.05	`themes`	-0.05	`main`
0.05	`addthis`	-0.05	`td`

Table 1: Weights and tokens when predicting whether a journal is whitelisted

hyper-parameter using grid search. Then we get the area under the curve using the test set to measure the performance. For the webpage screenshot classifier, we split the training and validation set several times to get several area under the curve score and get the average of the score to measure the performance.

For both models, classifiers provides probabilities of how likely a data point is a positive case (i.e., whitelisted journal). By getting the differences between the likelihood and the label, we are able to measure how good or how bad a prediction was made. By sorting the measured differences, we get the top "best positive prediction", "best negative prediction", "worst positive prediction", and "worst negative prediction". In this work, we will interpret those extreme cases to see if it makes sense to judge a journal by its websites.

4 Results

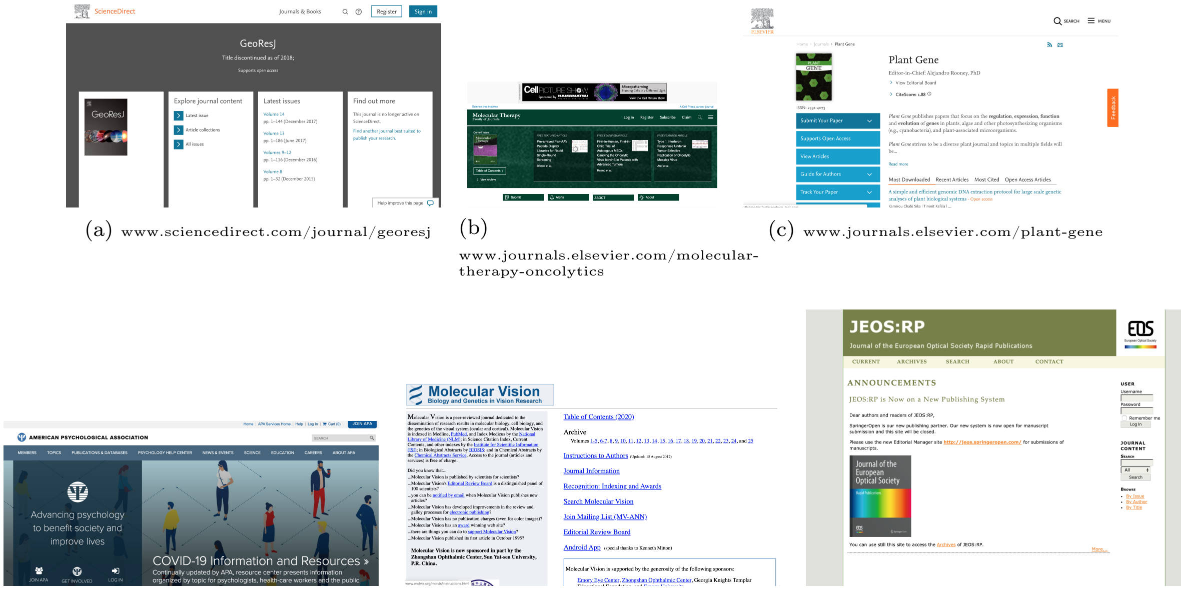
In this work, we are trying to understand whether the appearance of a journal is related to whether it is predatory or not. Relying on a curated dataset of open access publications provided by DOAJ, we built classifiers to detect whether a journal is whitelisted or predatory based on a journal's website screenshot and source code. Further, we examine the characteristics of the features that make a website be classified as predatory or not.

4.1 Results from source code model

One of the most basic characteristics to analyze of a website is its source code. We used the HTML tags of a journal's website source code to understand whether it is predictive of its status (see Methods). Our elastic-net regularized logistic regression displays a medium level of performance with an area-under-curve of 0.756. This result suggests that even by looking at the code of a website, we can determine relatively well whether it is predatory.

With the simple analysis presented above, we can try to understand the features of a website that makes it predatory or not. In particular, we can look at the *weights* in the regression. Positive weights indicate that the feature is important for determining whether it is whitelisted whereas negative weights indicates the opposite. Using these guidelines (Table 1), we observe that the token "fitplayback" has a very negative coefficient, which means that the token is a strong predictor that a web page is hosted by a un-whitelisted journal. We investigated further and found that the word "fitplayback" is the name of a user-defined custom Javascript function. The repetition of the same naming convention indicates that there are several un-whitelisted journals using a shared Javascript functionality. We may infer that those websites might be from the same group of developers or they are products of outsourcing companies (e.g., see Byrne and Christopher, 2020). Among tokens with negative coefficients, we also have "color", "script" and "autocomplete". Among tokens with positive coefficients, we have "at", "div", "span", which could potentially be the indicator that a website is more organized. Thus, these weights give us a sense what is important in the prediction of predatory journals or not.

We then wanted to understand what kinds of errors our method is making. This error analysis allows us to understand the extend to which the problem might need more data to improve or whether the problem is



ScienceDirect
GeoResJ
Title discontinued as of 2018;
Explore journal content
Latest issues
Find out more
Molecular Therapy
Plant Gene
American Psychological Association
Advancing psychology to benefit society and improve lives
COVID-19 Information and Resources »
Molecular Vision
Biology and Genetics in Vision Research
JEOS:RP
Journal of the European Optical Society Rapid Publications
ANNOUNCEMENTS
JEOS:RP is Now on a New Publishing System

(a) www.sciencedirect.com/journal/georesj

(b) www.journals.elsevier.com/molecular-therapy-oncolytics

(c) www.journals.elsevier.com/plant-gene

(d) apa.org/journals

(e) molvis.org

(f) jeos.org

Figure 1: Failed predictions by source code model. **Top row**: Incorrectly predicted as whitelisted. **Bottom row**: Incorrectly predicted as un-whitelisted

fundamentally hard. By getting the error of predictions from the classifier on the validation dataset (Figure 1), we were able to spot predictions that are especially bad. We take top three worse false positive.

For those un-whitelisted websites falsely predicted as whitelisted websites, they look visually presentable and organized but follow problematic publishing principles. The classifier tends to believe journals hosting such visually pleasant features are more likely to be a whitelisted. The classifier thus has falsely used the appearance of the website to judge its quality. As for whitelisted journals falsely predicted as un-whitelisted journals, they tend to be websites that adopt aesthetics from the early 2000. For those cases, the classifier makes false predictions based again on the appearance of websites of whitelisted journals. Taken together, these results suggest that the classifier seem to have captured the aesthetic feature of websites as predictor of the quality of journals. This means that indeed the appearance of websites is a valid signal for predicting quality.

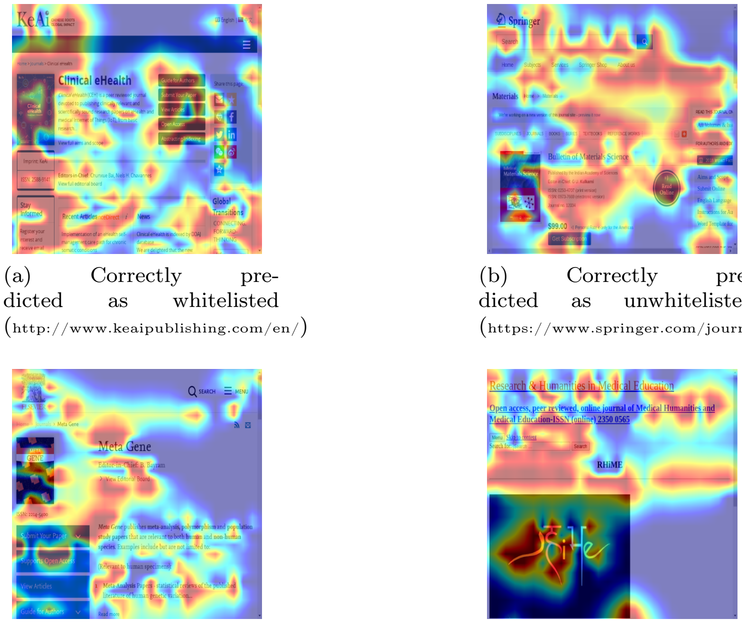
4.2 Results from website's screenshot-based model

Remarkably, we get a model with an area under curve of 0.736. This suggests that only looking at an actual picture of the website might in principle give us some information about its quality.

We then picked the top false positive, false negative, true positive and true negative data point in the validation split and interpret them based on what the classifier considers the most important image patch. Although interpreting deep learning models is challenging, we use a technique proposed by Fong and Vedaldi (2017) that computes the pixels "important" for a making a prediction.

As we may observe from the interpretation for top true predictions (Figure 2, Figure 2b), the classifier paid close attention to how the web page is structured and its logos. For example, the classifier would pay attention to whether a web page has an email logo on a web page. The classifier also pays attention to what kind of logo the web page has, the font of the web page. In general, whitelisted journals seem to have a good sense of design and be good quality.

On the other hand, for false predictions, we have a similar pattern to that found in the source code classifier. False positive samples (Figure 2c) are websites that look clean and visually organized while false negative websites are below-average looking but hosted by a known publisher. For example, the classifier pays attention to the side bar because it might indicate a good information organization. The classifier has falsely predicted an un-whitelisted journal based on the apparent good organization of it. On the other hand,


KeAi
Clinical eHealth
Springer
Bulletin of Materials Science
Meta Gene
Research & Humanities in Medical Education
RHiME

(a) Correctly predicted as whitelisted (http://www.keaipublishing.com/en/)

(b) Correctly predicted as unwhitelisted (https://www.springer.com/journal/12034/)

(c) Incorrectly predicted as whitelisted (https://www.journals.elsevier.com/meta-gene)

(d) Incorrectly predicted as unwhitelisted (https://www.rhime.in/)

Figure 2: Correct and incorrect predictions by screenshot-based model

for the false positive case (Figure 2d), the classifier has paid attention to the font, large images and whether the website has a menu bar. It is somewhat obvious that the website is relatively worse organized and the classifier has paid close attention to that, which the prediction is heavily counted on.

Based on the interpretation, we know that even though the classifier is predicting whether a journal is whitelisted, it takes the appearance of the website into account and the performance of the classifier shows that the logic of the classifier makes sense.

5 Discussions and Conclusions

In this work, websites for journals are used to generate features for predicting whether a journal is whitelisted or un-whitelisted, in other words, whether a journal has aligned with good publishing guideline by the Directory of Open Access Journals (DOAJ). Even though the classifiers we built are trained on merely 262 journals, the models have achieved reasonable performance. Our interpretation reveals that the classifiers would put a large importance to the appearance of the website when deciding whether a journal is predatory or not.

One limitation of our study is that we are using the DOAJ as our guiding institution but there is no universal principle for judging the quality of journal. For example, the term predatory journal remains fuzzy and lacks formal definition (Grudniewicz et al., 2019). However, the DOAJ is one of the first organization to openly share their list and has one of the largest sets of curated journals. In the future, we will explore how good our models are at predicting other lists such as Beall's and Cabells's lists.

The performance of our models leave a great deal to be desired. A performance of AUC 0.75 would necessary require human intervention for making a final judgement. One of the issues is that the dataset size and features are not proportional to the complexity of the problem. For example, the image classifier is only using 262 images whereas typical deep learning classifiers need thousands of data points. In the future, we will explore other features that we could engineer from other data and better ways of operationalizing the characteristics of good journals favored by DOAJ.

In spite of our limitations, we successfully explored whether the appearance of a website is important and the features that make this happen. We think our pipeline could be significantly expanded in the future and be a complementary tool when analyzing new websites. Our method significantly accelerate the process, saving valuable time for researchers who are sometimes lost in a sea of new publishing venues.

References

Bagues, M., Sylos-Labini, M., and Zinovyeva, N. (2019). A walk on the wild side:‘predatory’journals and information asymmetries in scientific evaluations. *Research Policy*, 48(2):462–477.

Byrne, J. A. and Christopher, J. (2020). Digital magic, or the dark arts of the 21st century—how can journals and peer reviewers detect manuscripts and publications from paper mills? *FEBS letters*, 594(4):583–589.

Dadkhah, M., Maliszewski, T., and Jazi, M. D. (2016). Characteristics of hijacked journals and predatory publishers: our observations in the academic world. *Trends in pharmacological sciences*, 37(6):415–418.

DOAJ (2020). Directory of open access journals.

Downes, M. (2020). Thousands of australian academics on the editorial boards of journals run by predatory publishers. *Learned Publishing*.

Eysenbach, G. (2006). The open access advantage. *Journal of Medical Internet Research*, 8(2):e8.

Fong, R. C. and Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437.

Grudniewicz, A., Moher, D., Cobey, K. D., Bryson, G. L., Cukier, S., Allen, K., Ardern, C., Balcom, L., Barros, T., Berger, M., et al. (2019). Predatory journals: no definition, no defence.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. corr abs/1512.03385 (2015).

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*. Cambridge university press.

Severin, A., Strinzel, M., Egger, M., Domingo, M., and Barros, T. F. (2020). Who reviews for predatory journals? a study on reviewer characteristics. *BioRxiv*.

Smith, E., Gunashekar, S., and Parks, S. (2016). A framework to monitor open science trends in the eu. *New Media & Society*, 14(5):729–747.

Solomon, D. J. and Björk, B.-C. (2012). A study of open access journals using article processing charges. *Journal of the American Society for Information Science and Technology*, 63(8):1485–1495.

Stephan, P. E. (2012). *How economics shapes science*, volume 1. Harvard University Press Cambridge, MA.

Till, J. E. (2001). Predecessors of preprint servers. *Learned publishing*, 14(1):7–13.