Assessing the Impact Score of Scientific Publications through the Analysis of Abstracts and their Metadata

Artem Lebedev
UC Berkeley / Berkeley, CA
artem.lebedev@berkeley.edu

Farouk Ghandour
UC Berkeley / Berkeley, CA
fghandour18@berkeley.edu

Abstract

An impact score of the academic paper is a metric critically important for a scientist's career and attraction of funding for future research. The ability to predict the potential score of a paper could help scientists adjust their presentation strategy to achieve maximum impact. Natural language processing (NLP) is a logical tool to model the factors influencing the impact score. In this paper we show that the content of the abstract indeed influences the impact factor, but meta-factors, such as author's country and length of the paper also have surprisingly high predictive power. Augmenting initial data set with synthetic data we managed to create a model that predicts impact factor of a paper based on the abstract content and other information available to the authors at the time of submission. The model's predictions showed 73% correlation with the observed data and thus can be used to guide publishing decisions. [1]

1 Introduction

The ability to predict the paper's significance could help researchers adjust their presentation strategy. It would allow researchers to publish in the most impactful journal while avoiding delay caused by rejection from a journal with standards that the paper does not meet. In this paper we elucidate the factors that drive acceptance in high-impact journals and build a model that predicts it.

Our hypothesis is that variation in the impact of an academic paper can be mostly explained by the metadata of the paper: country of origin, number of authors, length of the paper etc. The metric of impact that is the most amenable for the quantitative analysis is Journal Impact Factor (JIF), which is, roughly, the average number of times papers published in this journal this year are cited in the works of others. The advantage of this metric is that it is less prone to stochastic variation, but as a result it reflects the opinion of the reviewer's more than the intrinsic value of the paper.

We selected Pearson's correlation coefficient between the predicted impact and observed impact as the metric to evaluate and compare models. We will assume that we failed to reject null hypothesis if observed-predicted correlation of the model including text data and metadata is not significantly larger than correlation achieved with the same model lacking text data (Eq. 1).

$$\begin{cases} H_0 : (r_{w\ text} - r_{only\ metadata}) \leq 0 \\ H_A : (r_{w\ text} - r_{only\ metadata}) > 0 \end{cases} \quad (1)$$

In the social science context correlation coefficient above 0.5 is generally considered moderately positive(Kahneman, 2021). Following this convention, we will consider our findings to be practically significant only if at least one coefficient exceeds 0.5. If both coefficients are below 0.5, it is likely that our research question can not be answered with this data set and our approach.

2 Background

A few attempts have been made to address this issue, demonstrating modest but encouraging success. (Macri et al., 2023; Alohali et al., 2022; Thelwall et al., 2023; van der Zwaard et al., 2020) The most sophisticated model utilized BERT context-aware encoding of the abstract to produce embedding that were then classified using SVM, logistic regression or XGBoost to predict the impact factor quintile. This approach demonstrated prediction accuracy of approx. 75

3 Data

We extracted data from the Web of Science, a platform that provides access to citation metrics for the majority of life science publications. The platform provides csv files with the article abstract

[1] The code and data available on GitHub and GDrive.

along with the ISSN identifier of the journal and other metadata. To enable use of the most developed NLP models and eliminate variability due to the type of the publication we selected only original peer-reviewed articles published in English between 2000 and 2022. To further reduce variability unrelated to the research question we focus on a narrow field of radioligand therapy. Trained on this narrow data set, our models are not transferable to even other areas of life science, but in return we avoid wide variability due to popularity of certain "fashionable" research topics. JIFs are available from Claritive Analytics “Journal Impact Factor Report” and contain the journal ISSN as well as the impact factor, by year. Merging these two datasets on ISSN we achieve our raw dataset, containing approximately 4500 records.

Since our aim is to support researcher decision on publication strategy, we focus on the features available to authors at the time of submission: ’Year Published’, ’Authors’, ’Document Title’, ’Author Keywords’, ’Abstract’, ’Author Address’, ’Funding Agency and Grant Number’, ’Cited Reference Count’, ’Page Count’. Using author identities as features would result in very sparse set of categorical variables, therefore we reduced the author information to just the number of authors and the author address to just the country. We also reduced the funding information to a binary variable funded/not-funded. ’Abstract’ is the most important and relevant feature of the paper. Together with the title and author’s keywords it represents condensed meaning of the paper.

EDA revealed that numerical variables are mostly uncorrelated with each other and label is roughly normally distributed around mean of 4.09 with standard deviation of 2.21. Less that 0.1% of the papers had JIF > 20, but some reached JIF as high as 86. To avoid undue influence of these high performers we filtered out all papers with JIF above 20. Unsurprisingly, USA and the UK dominate the countries of author’s origin, but in total more than 80 countries were represented. Majority of abstracts are less that 400 words long with noticeable peak around 250, a popular cut-off required by publishers. For the purpose of tokenization we set 400 tokens a the max length. Some abstracts contained publisher’s copyright statements that we removed to avoid a leak from label to the features. Full EDA is available in the notebook.

4 Methods

We use regression models based on neural networks to predict the impact of the publications. To evaluate the importance of text data, we compare predictive power of the model that includes text data vs the model that is based solely on metadata. Pearson’s correlation coefficient between observed and predicted impact values is a metric we use to test our hypothesis. One-tailed p-value for comparing two correlation coefficients is calculated after Fisher’s transformation. Whenever justified we assumed dependent sample to increase statistical power, but fell back on independent sample if data set were not identical between two models.

Features selected after EDA fall into three general categories: numerical features, categorical features and text. While numerical features can be included into the model as-is, categorical variables and text need to be transformed before they become inputs into neural network. One-hot-encoded representation is a common way to transform categorical variables, but it lacks the ability to capture patterns present in the data. For instance, papers originating from Canada are likely similar to ones written in the USA, but very different from papers submitted from China. To preserve these patterns, we have chosen to use learned embedding representation of the categorical variables (Guo and Berkhahn, 2016).

Transformation of text into contextually-aware embedding is commonly achieved using BERT (Devlin et al., 2018). sciBERT, a similar model trained on the large corpus of medical and scientific literature, was recently presented (Beltagy et al., 2019) and we compare these two models. Both variants of BERT yield several vectors that can represent the text: pooled vector, *cls* vector or a set of vectors for each input token. For the baseline model we concatenated *cls* vector with the embeddings representing categorical features and the vector of the numeric features, and used this vector as an input into a linear regression. LSTM and CNN approaches are investigated to generate a representation of text from all BERT embeddings, rather than relying on the summarization tokens produced by BERT model.

5 Results and discussion

5.1 Testing of the main hypothesis

Preliminary experiments Before testing our hypothesis we tested a few model architecture choices. First we investigated how the number of dense layers affects performance. We used *cls* token generated by standard uncased BERT and added no dense layers, one or two dense layers on top of the 768-dimensional vector. All dense layers had 768 neurons. Correlation between predicted and observed JIFs improved upon addition of the first layer (r = 0.54 vs 0.57 p = 8e-05, dependent samples), but did not improve with addition of the second dense layer (r = 0.56). Replacing standard BERT with sciBERT, a model trained on scientific text, we further improved the result (r = 0.57 vs 0.64 p = 2e-08) (notebook). In a separate experiment we teste mean pooling instead of *cls* token. Using mean pooling we aggregated the embeddings into a single vector, that was passed passed through an input layer and then multiple hidden layers with a relu activation and the final output layer is a dense layer with a linear activation. This resulted in a Pearson Correlation of 55%, lower than the model based on *cls* token.

We also attempted to test this model as a classification model. We first divided the target variable (JIF) into 10 equal quantiles. We then ran a CNN on top of this after initially running the encodings through SCI Bert. Our initial results we not very promising. This test returned a correlation coefficient of 35%. Another test was run with 5 quantiles and we received a correlation coefficient of 62%. Finally we ran a test with 20 quantiles and we received a correlation coefficient of 25%. This data makes clear sense. In the case of the 5 quantiles, since there were less options with the classification model, there was much higher likelihood that the model could detect a correlation. Whereas if the quantiles were increased to 10 or 20, it becomes much less likely that a correlation will be detected. These experiments are detailed in this notebook

Although the correlation score of this model yielded a strong result, running a classification to predict the JIF score doesn't seem to be a good approach. The data will be approximated too much and although the correlation could be high, we are sure to overfit if there is not enough quantiles within the dataset. Thus we continued with our original approach of utilizing a regression.

Main hypothesis testing For the baseline model we built three regression models. The one based on metadata only included numerical features concatenated with learned embeddings for each of the categorical features. The resulting vector was fed directly into the dense output layer with one element and linear activation. Supporting intuition behind our null-hypothesis, r value for this model was 0.55, far better than a random guess. This indicates that metadata alone, with no features related to the essence of the paper, can predict the impact of the paper. The second model included only sciBERT *cls* token and no meta data. This model exhibited r-value of 0.63, a statistically significant (p=0.0008, independent samples) improvement. The final model included both metadata and sciBERT *cls* token concatenated into one vector before the final dense and regression layer. This model demonstrated r = 0.67, an improvement over both baseline models. This notebook details these experiments. It exceeded performance of the pure metadata model with p-value 2E-07 (notebook). Based on that we can reject the null hypothesis and conclude that addition of a feature related to the paper's meaning does increase model's predictive power by approx. 20%.

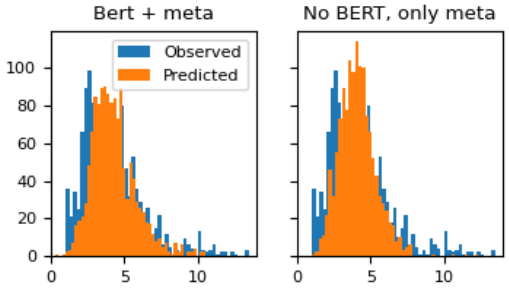

Figure 1: Observed-Predicted for models w metadata.

Figure 1 shows the distribution of the observed and predicted JIFs for the first and the third model. Interestingly, both models are reluctant to predict low JIFs. Additionally, learning curves showed clear signs of over-fitting with validation loss far exceeding training loss.

Rank prediction We investigated an alternative label representation, replacing JIF of the paper for its rank. To that end we calculated the proportion of papers that have the same or lower JIF for each paper (rank). We used max and min JIF obtained from training data and assumed it is representative of the whole population. Scaling label to (0,1) allowed us to use sigmoid activation in the regression layer, which we hoped would make

model more sensitive to the changes in the middle of the distribution. However, this approach resulted in lower model performance (r=0.61). Using linear activation instead of sigmoid resulted in even worse performance. Within the same experiment, we added drop-out layers after learned embeddings, that reduced over-fitting: validation loss exceeded training loss later in the training, pointing to the usefulness of drop-out layers (notebook).

5.2 Performance improvement

In attempt to increase predictive power of the model we hypothesized that incorporating the full set of BERT tokens would allow for better modeling of the abstract meaning. Two general approaches were investigated: using LSTM as an encoder to summarize the abstract and using a CNN set of layers to do the same.

LSTM for abstract representation The intuition behind using LSTM comes from observation that reviewer reads the abstract one word after another and after having read the full text, produces their opinion on weather to recommend the paper for publication or not. Mimicking this process we built a two-layer LSTM model with 768 nodes (to match BERT embedding) and trained it on the full output of sciBERT (400 tokens). Surprisingly, the correlation between predicted and observed JIFs was only r = 0.57 (notebook), substantially lower than the model that used *cls*.

CNN for abstract representation There is also a possibility that reviewer perceives the entire abstract as a whole, and it is the interactions between the words that produce the final judgment. From that perspective, using convolutional network to produce the final representation of the abstract appears to be a viable approach. We mimicked a CNN architecture typical in image processing using all tokens produced by BERT, and reshaping them into 399x768x1 image combined with three convolutional layers, each followed by batch normalization and maxpooling. The final convolution was flattened and concatenated with the embeddings for categorical features and the vector of numeric features, before passing it to a hidden layer and a regression layer. This architecture was not successful, yielding unsatisfying r = 0.47 (notebook).

Addition of the title In attempt to enrich the data set and reduce over fitting we added paper title to the abstract. Two approaches were investigated: generation of a septate *cls* token and concatenation of title and abstract. Our initial intuition was that title of a paper is meant to be information dense and should be represented in the final regression approximately equal to the full abstract. Therefore we generated two separate *cls* tokens and concatenated them together with the embeddings for the categorical features and a vector of numerical features (notebook). Contrary to our expectations, this approach resulted in reduced performance (r = 0.62).

A more straightforward and less computationally expensive approach is to concatenate title to the abstract and generate one *cls* (notebook). This method resulted in small, but significant performance improvement (r = 0.69) and also reduced over-fitting.

Data augmentation The fact that concatenation of title with abstract improved performance indicated that diverse set of training data, rather than sophisticated model, is the key to further improvements. Obtaining more training data would require paid subscription and we opted for data augmentation instead. First, all title+abstract paragraphs were ranked by JIF, and then for each record we generated a set of new abstracts by replacing n-th sentence of the given abstract first with the n-th sentence of the previous abstract and then with the n-th sentence of the next abstract in the dataset. Replacement was done 2k times, where k is the number of sentences in the shortest of the three abstracts. Each of the k replacements were done on the original abstract. For these new abstracts all other features and labels were kept equal to the abstract being augmented. The resulting training set was randomly shuffled before model training. Validation data set was not augmented. (notebook). This procedure increased the size of the training dataset approximately 10 fold, but introduced only minimal perturbance to the original abstract as only one sentence was replaced, and the implant was taken from a paper with a very similar JIF. This augmentation procedure yielded the best performing model, with correlation between predicted and observed JIF r=0.73. Figure 3 shows a 2D-histogram of the predictions done on the validation dataset.

5.3 Error analysis

Comparing distribution of the predicted JIFs and the observed values (Figure 2) we can see that distribution of the predicted values is a lot closer to normal than the distribution of the observed. Two

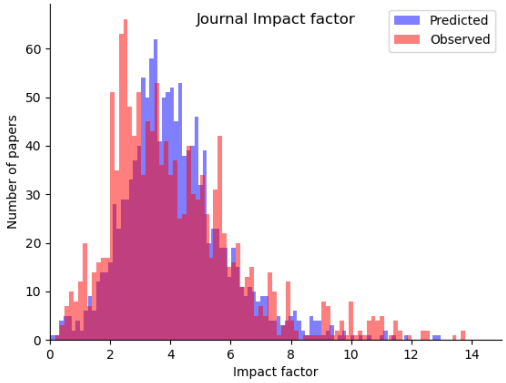

Figure 2: Overlaid distributions of observed and predicted values for the best model.

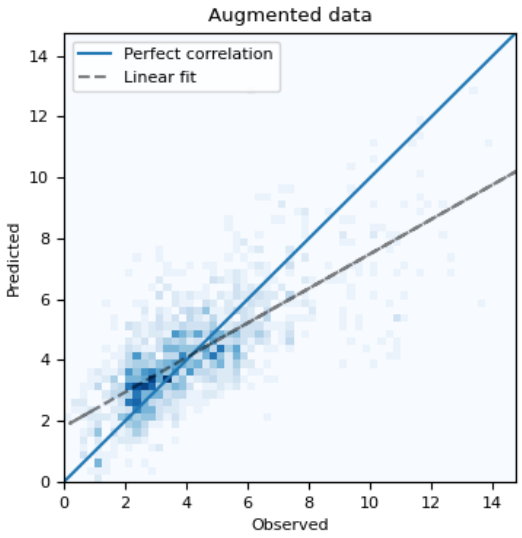

Figure 3: Observed-Predicted correlation for the best model.

departures from the normality for the observed values are immediately obvious: spikes in the frequency at JIFs 1 and 2.5. Our model failed to capture these two features. Instead, those papers that should have been in those spikes were assigned higher JIF resulting in over-prediction for JIFs around 4. Consequently, the correlation between predicted and observed values is far from perfect (Figure 3).

Cursory examination of the mispredicted records yield a plausible explanation. Some journals, in particular those with lower rank, allow for much longer abstracts, essentially turning the abstract into a mini-paper, including sections on experimental details, methods and discussion. This is not a standard practice, most other journals require abstract to be under 250 words and convey only general idea of the paper. The model likely assigned higher value to these unusually high quality abstracts. This hypothesis is partially confirmed by the modelling of the prediction error (*vide infra*).

To elucidate the role of various factors in making erroneous predictions, we built a series of simple linear models predicting the difference between observed and predicted JIF. The only factor that had reasonable predictive power (r = 0.71 for observed and predicted error, details are in the notebook) was the journal name. Interestingly, very different kinds of journals had the same coefficient explaining the error: prestigious "Proceedings of the National Academy of Science" was under-predicted by the same factor as an obscure "Chinese Chemical Letters". Some of this can be explained by the phenomenon mentioned above, but some of the journals are prone to error because their impact factor is influenced by factors not included in the model. For instance, "Chinese Chemical Letters" publish a lot of reviews that naturally have much higher impact factor, thus pushing overall journal impact factor up. We explicitly excluded reviews from our analysis and could not capture this trend.

6 Conclusions and Next Steps

The experiments reported in this paper strongly support rejection of our somewhat cynical null hypothesis. Moreover, unexpectedly good performance of our best model makes it potentially useful for scientists. With more than 70% correlation between predicted and observed, this model can give researchers a good estimate of what journals might accept their paper. Additionally, with small changes to abstract and title, authors can maximize the chances of acceptance in a more prestigious journal.

It is unlikely that the predictive power can be improved much further: acceptance in a journal is always a reviewer's judgement call, and predicting judgement calls is inherently difficult. However, access to a more diverse dataset can potentially reduce model bias. As a topic for future research, we would be interested in applying this methodology to predicting the actual citation index of individual papers. Citation index is a measure of individual project value, rather than the average journal performance, and might provide useful guidance to the scientists striving to generate the most impactful

research.

References

Yousef A Alohali, Mahmoud S Fayed, Tamer Mesallam, Yassin Abdelsamad, Fida Almuhawas, and Abdulrahman Hagr. 2022. A machine learning model to predict citation counts of scientific papers in otology field. *Biomed Res. Int.*, 2022:2239152.

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *CoRR*, abs/1903.10676.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Cheng Guo and Felix Berkhahn. 2016. Entity embeddings of categorical variables. *CoRR*, abs/1604.06737.

Lukas Haas and Michal Skreta. 2022. From roBERTa to aLEXa: Automated legal expert arbitrator for neural legal judgment prediction. In *Stanford CS224N Custom Project Final Report*.

Daniel Kahneman. 2021. *Noise : a flaw in human judgment / Daniel Kahneman, Olivier Sibony, Cass R. Sunstein.*, first edition. edition. Little, Brown Spark.

Carmelo Macri, Stephen Bacchi, Sheng Chieh Teoh, Wan Yin Lim, Lydia Lam, Sandy Patel, Mark Slee, Robert Casson, and Wengonn Chan. 2023. Evaluating the ability of open-source artificial intelligence to predict accepting-journal impact factor and eigenfactor score using academic article abstracts: Cross-sectional machine learning analysis. *J. Med. Internet Res.*, 25:e42789.

Mike Thelwall, Kayvan Kousha, Paul Wilson, Meiko Makita, Mahshid Abdoli, Emma Stuart, Jonathan Levitt, Petr Knoth, and Matteo Cancellieri. 2023. Predicting article quality scores with machine learning: The U.K. Research Excellence Framework. *Quantitative Science Studies*, 4(2):547–573.

Stephan van der Zwaard, Arie-Willem de Leeuw, L. (Rens) A. Meerhoff, Sue C. Bodine, and Arno Knobbe. 2020. Articles with impact: insights into 10 years of research with machine learning. *Journal of Applied Physiology*, 129(4):967–979. PMID: 32790596.