Predicting Research Impact: A Novel Linguistic Approach for Forecasting Citation Counts

Presented to
Professor Taha Havakhor

By
DeSilva, Dhevin - 261177497
Delisle, Audrey - 261142504
Barabasz, Michelle - 261152119
Oluwole-Rotimi, Angel - 261171068

INSY 669 - Section 076

McGill University - Desautels Faculty of Management
Tuesday February 13th 2023

Table of Contents

1. Introduction & Background....2

1.1 Introduction.... 2

1.2 Problem at hand.... 2

1.3 Objective....2

2. Methodology.... 2

2.1 Data Collection.... 2

3. Results.... 3

3.1 Key Findings....3

4. Expected Impact....3

4.1 Novelty of project.... 3

4.2 Business, economic and societal impact....3

4.3 Importance of topic....3

5. References.... 4

6. Appendix A : Code Outputs....5

Model 1 Classification Report.... 5

Model 2 Classification Report.... 5

Model 3 Classification Report.... 5

Pipeline Examples using Model 3:.... 6

1. Introduction & Background

1.1 Introduction

In the realm of scientific research, accurately predicting the impact of research papers is crucial for researchers, publishers, and academic institutions. Citation counts serve as a primary metric for gauging a paper's impact, influencing career advancements, funding opportunities, and future research directions.

1.2 Problem at hand

Traditional methods of predicting citation counts have primarily focused on bibliometric features such as the number of authors and publication sources, but the inherent lead time in the academic publication cycle means these can take years to manifest. Additionally, these methods do not account for the value of the content itself, particularly the novelty of the research presented. This project aims to bridge the gap by investigating the effect of incorporating a novel approach to citation prediction using a novelty score derived from text analysis of their titles.

1.3 Objective

The primary objective of this project is to develop and compare three classification models for predicting the citation counts of research papers, thereby evaluating the added value of incorporating a novelty score derived from the text analysis of paper titles. This project also intends to contribute to the understanding of how the novelty of research, as reflected in its title, correlated with its citation count.

2. Methodology

2.1 Data Collection

The project sourced data from the Publish or Perish database, targeting the Natural Language Processing (NLP) field and collecting 500 titles annually from 2015 to 2021. This dataset, featuring citation counts, publication years, and titles (over abstracts) for their distillation of paper themes and reliability of data collection. Data cleaning involved removing blanks or N/A fields to ensure dataset consistency.

2.2 Text and Citation Analysis

2.2.1 Model 1

The initial model, Model 1, which served as the baseline for comparison, was created using a straightforward approach that relied solely on the variable "Years Since Publication" to predict whether a paper would be classified as having a high or low number of citations. The process began by calculating the median citation count for the dataset, which was then used to create a binary “High Impact” target column, assigning a 1 to papers whose citation count was above the median, which was calculated being 7 citations. Additionally, the model computed the approximate years since each paper's publication.

2.2.2 Model 2

Advancing from Model 1, Model 2 incorporated both the years since publication and a set of novelty scores calculated using Euclidean distance. This method was chosen over cosine because of its effectiveness in capturing the intensity of thematic novelty within titles. This model also experimented with annual vs. overall corpus-based novelty score calculations, ultimately favoring the latter for its simplicity. Additionally, Model 2 used dissimilarity scores—derived from the inverse of cosine similarities among document vectors—to further refine the novelty assessment. The model utilized count vectorization for text analysis, due to its straightforward representation of thematic elements over TF-IDF, and included linear and non-linear transformations—namely logarithmic and squared modifications—of the novelty score to capture the complex relationship between novelty and citation impact.

2.2.3 Model 3

The most advanced iteration, Model 3, introduced an analysis of unigrams, bigrams, and trigrams extracted from paper titles to quantify novelty more granularly in addition to the metrics from Model 2. Unigrams represent single words, bigrams, pairs of consecutive words, and trigrams sequences of three words, each offering different levels of contextual insight into the thematic content of titles. This model compared these linguistic features against a baseline corpus from 2015 to identify unique elements signaling thematic novelty. By counting the occurrences of new unigrams, bigrams, and trigrams not in the baseline, Model 3 provided an even more advanced measure of novelty.

3. Results

3.1 Key Findings

In developing predictive models for research paper citation counts, using a general novelty score (Model 1 receiving 0.62 accuracy, Model 2 achieving a 0.64 accuracy) did not significantly enhance prediction accuracy. However, Model 3 showed improvement (achieving 0.65 accuracy) by adding relevant linguistic details, highlighting their importance in prediction (please refer to Appendix A). Furthermore, a predictive pipeline was developed as part of Model 3, designed to determine whether a new title is likely to receive a low or high citation count. This pipeline was tested on three different example title names (view Appendix A), giving each one a novelty score and classification. A title like NLP got a score of 0 as expected, but a title about a new healthcare innovation got a score of 1 because there are no healthcare words in the corpus. This reaffirms our pipeline works successfully.

4. Expected Impact

4.1 Novelty of project

The novelty of this project is in using text analysis to predict citation counts, whereas before, only bibliometric indicators were used. This methodology not only highlights the value of innovative research but also challenges existing methods, offering a new perspective on assessing academic contributions.

4.2 Business, economic and societal impact

The project's predictive pipeline has the potential to transform how publishers and academic platforms make informed decisions about which research papers to publish by evaluating the potential impact based on title analysis. It serves as a safeguard against the rise of "clickbait" titles, ensuring that attention-grabbing headlines match the substantive quality of the research. Economically, it can reallocate resources towards high-impact innovations, speeding up advancements in science and technology, leading to a more efficient allocation of funds, ensuring that investment is directed towards research with the highest potential for societal benefit. Socially, it encourages the exploration of diverse research topics, some of which may directly inform policy making and lead to breakthroughs in healthcare and other fields critical to societal advancement. This approach ensures that pioneering work, especially in areas that can shape the future in health, technology, and environmental sustainability, is identified and promoted for its ability to drive meaningful change.

4.3 Importance of topic

In a time where scientific research is both a driver of global progress and subject to intense scrutiny, developing more accurate methods of evaluating research impact is crucial. This project not only addresses a gap in the current evaluation methods but also proposes a solution that could reevaluate the understanding of what makes research influential. The goal of this project is to highlight the need for metrics that reflect the evolving nature of knowledge creation and dissemination.

5. References

Beirlant, J., Bornmann, L., Cirillo, P., Didegah, F., Levitt, J., Mingers, J., Wallace, M. L., Wang, M., Abramo, G., Acuna, D., Adams, J., Albarrán, P., Brody, T., Burrell, Q., Clauset, A., Costas, R., Price, D. J. D. S., & Dekkers, A. L. (2015, August 25). *Predicting the long-term citation impact of recent publications*. Journal of Informetrics. https://www.sciencedirect.com/science/article/pii/S1751157715300080?casa_token=e1KC-fZAiiwAAAAA%3AFAXFZluplW51zTaFRZ6Gcwd2-AElfZn7LUjIWgyge_KZk03MEI_3_TEkelqmBaA8oQymRuc_mw

DOWNLOAD47 million PDFsfor Free. Academia.edu - Share research. (n.d.). ttps://www.academia.edu/

Publish or perish. Harzing.com. (n.d.). https://harzing.com/resources/publish-or-perish

6. Appendix A : Code Outputs

Model 1 Classification Report

Model 1 Classification Report:

	precision	recall	f1-score	support
0	0.65	0.47	0.55	225
1	0.60	0.76	0.67	238
accuracy			0.62	463
macro avg	0.63	0.62	0.61	463
weighted avg	0.63	0.62	0.61	463

Model 2 Classification Report

Model 2 Classification Report:

	precision	recall	f1-score	support
0	0.65	0.57	0.61	225
1	0.64	0.71	0.67	238
accuracy			0.64	463
macro avg	0.64	0.64	0.64	463
weighted avg	0.64	0.64	0.64	463

Model 3 Classification Report

Model 3 Classification Report:

	precision	recall	f1-score	support
0	0.65	0.58	0.61	225
1	0.64	0.71	0.67	238
accuracy			0.65	463
macro avg	0.65	0.64	0.64	463
weighted avg	0.65	0.65	0.64	463

Pipeline Examples using Model 3:

```
Paper Title: 'Natural Language Processing review'
Predicted Class: 0
Novelty Score: 0.60
New Unigrams: 0, New Bigrams: 1, New Trigrams: 1
------------------------------------------------------------
Paper Title: 'jack jack jack jack jack jack jack jack jack jack jack'
Predicted Class: 0
Novelty Score: 1.73
New Unigrams: 1, New Bigrams: 1, New Trigrams: 1
------------------------------------------------------------
Paper Title: 'Transplantation of cultured islets from two-layer preserved pancreases'
Predicted Class: 1
Novelty Score: 2.24
New Unigrams: 6, New Bigrams: 5, New Trigrams: 4
------------------------------------------------------------
```