amon_2022_is_it_all_bafflegab_linguistic_and_meta_characteristics_of_research_articles_in_prestigious_economics_journals

Year

2022

Author(s)

Julian Amon and Kurt Hornik

Title

Is it all bafflegab? – Linguistic and meta characteristics of research articles in prestigious economics journals

Venue

Journal of Informetrics

Topic labeling

Manual

Focus

Secondary

Type of contribution

Established approach

Underlying technique

Manual labeling

Topic labeling parameters

\

Label generation

"…we investigate the posterior word distributions to find the most frequent terms in each topic and manually label these topics accordingly"

Table A.2
Labels of the topics from LDA output.

Topic ID	Topic label	Topic ID	Topic label
1	Modeling & testing	16	Energy economics
2	Econometrics	17	Macroeconomics
3	Business administration	18	Game theory
4	Leadership	19	Supply chain management
5	Marketing	20	Economics and society
6	Socioeconomics	21	Resource economics
7	Stock market	22	Governance
8	Project management	23	Trade economics
9	Education	24	Optimization
10	Model description	25	Forecasting
11	Financial system	26	Urban economics
12	Monetary economics	27	Microeconomics
13	International economics	28	Innovation & technology
14	Mathematical economics	29	Mathematics
15	Corporate finance	30	Transport economics

Table A.2 matches the topic IDs used in Table 3 to their labels. From the LDA algorithm, we obtain posterior word distributions for each topic and use this information to assign suitable labels to the topics. Word clouds of all topics that illustrate these posterior word distributions are contained in Figures S.1 and S.2 in the supplementary materials.

Motivation

\

Topic modeling

LDA

Topic modeling parameters

Nr of topics (K): 30

Nr. of topics

30

Label

Single or multi-word manually provided label

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Domain (paper): Article prestige discovery
Domain (corpus): Economics

Problem statement

In this paper, we propose an approach that understands scientific prestige in terms of the rankings of the journals that the articles appeared in, as such rankings are routinely used as surrogate research quality indicators.
For the purpose of determining the most important drivers of suchlike prestige, we use state-of-the-art text mining tools on journal articles in economics.
We then estimate beta regression models to investigate the relationship between these predictors and a cross-sectionally standardized version of SCImago Journal Rank (SJR) in multiple topically homogeneous clusters.
In so doing, we also reinvestigate the bafflegab theory, according to which more prestigious research papers tend to be less readable.

Corpus

Origin: Elsevier, Springer Nature
Nr. of documents: 255,644 (221,008 after pre-processing)
Details:

- All available articles in all the accessible economics journals from the Elsevier Article Metadata API until the end of January 2020
- Articles after pre-processing come from 290 different journals

Document

Full text economics article (title, abstract, full text, footnotes and appendices) with additional meta information contained in the XML format
Each article is mapped to a vector of 344 (linguistic and meta) characteristics and a table of term frequencies after pre-processing

Pre-processing

- removing double punctuation, expanding abbreviations and trimming white space
- tokenization, lemmatization, part-of-speech (POS) tagging, named entity recognition, constituency parsing and coreference resolution

The annotation objects generated by this procedure then allow the straightforward calculation of 127 linguistic features categorised in 7 groups:

- linguistic base, POS, entity, parse tree, word overlap, semantic and coreference.

From the XML, 30 meta features for each article are extracted. Those are grouped into the subcategories:

- base, author, bibliometric and mathematical

Recording the frequencies of all (lemmatized) unigrams, bigrams and trigrams that appear at least twice in the full text of a given article

We stipulated that each article in the data set have a publication date, at least one author, a title and abstract, at least one full-text section and at least one reference in its bibliography.

- A total of 29,950 documents did not meet these requirements and were thus removed.
- Moreover, we had to remove 4520 papers for which we could not accurately compute certain features, mostly because their XML files deviated slightly from the otherwise

standardized structure of the publisher, which occurs, for instance, with errata and editorial notes.

- Finally, a further 166 articles were discarded as they were published more than once in exactly the same form and thus also appeared multiple times in our data set.

```
@article{amon_2022_is_it_all_bafflegab_linguistic_and_meta_characteristics_of_research_articles_in_prestigious_economics_journals,
  abstract = {In competitive research environments, scholars have a natural interest to maximize the prestige associated with their scientific work. In order to identify factors that might help them address this goal more effectively, the scientometric literature has tried to link linguistic and meta characteristics of academic papers to the associated degree of scientific prestige, conceptualized as cumulative citation counts. In this paper, we take an alternative approach that instead understands scientific prestige in terms of the rankings of the journals that the articles appeared in, as such rankings are routinely used as surrogate research quality indicators. For the purpose of determining the most important drivers of suchlike prestige, we use state-of-the-art text mining tools to extract 344 interpretable features from a large corpus of over 200,000 journal articles in economics. We then estimate beta regression models to investigate the relationship between these predictors and a cross-sectionally standardized version of SCImago Journal Rank (SJR) in multiple topically homogeneous clusters. In so doing, we also reinvestigate the bafflegab theory, according to which more prestigious research papers tend to be less readable, in a methodologically novel way. Our results show the consistently most informative predictors to be associated with the length of the paper, the span of coreference chains in its full text, the deployment of a personal and moderately informal writing style, the ``density'' of the article in terms of sentences per page, international and institutional collaboration in research teams and the references cited in the paper. Moreover, we identify various linguistic intricacies that matter in the association between readability and scientific prestige, which suggest this relationship to be more complicated than previously assumed.},
  author = {Julian Amon and Kurt Hornik},
  date-added = {2023-03-10 14:02:53 +0100},
  date-modified = {2023-03-10 14:02:53 +0100},
  doi = {https://doi.org/10.1016/j.joi.2022.101284},
```

```
  issn = {1751-1577},
  journal = {Journal of Informetrics},
  keywords = {Research impact, SJR indicator, NLP, Readability, Gradient
boosting, GLMLSS},
  number = {2},
  pages = {101284},
  title = {Is it all bafflegab? -- Linguistic and meta characteristics of
research articles in prestigious economics journals},
  url = {https://www.sciencedirect.com/science/article/pii/S1751157722000360},
  volume = {16},
  year = {2022}}
```

#Thesis/Papers/Initial