Journals' agendas versus actual publications: A first look at article dynamics in innovation journals

Ana Teresa Santos[1] and Sandro Mendonça[2]

[1] *atmss@iscte-iul.pt*
Instituto Universitário de Lisboa (ISCTE-IUL) – Business Research Unit (BRU-IUL), Lisbon, Portugal.

[2] *sfm@iscte-iul.pt*
Instituto Universitário de Lisboa (ISCTE-IUL) – Business Research Unit (BRU-IUL), Lisbon, Portugal.
UECE/REM – ISEG/ University of Lisbon, Lisbon, Portugal.
SPRU, University of Sussex, UK.

Abstract
In this paper, we address the problem faced by researchers attempting to decide the appropriate journal to submit their works for. Based on content analysis, we studied how semantically similar are journals blurb's sections with the articles published by the outlets. By considering such a methodology, we propose a new strategy for journal's selection for manuscript submission decision based on endogenous outcomes instead of traditional ones like journal's scores centered on dissemination achievements. Throughout, we illustrate our analysis with data from twenty current innovation-oriented journals. We use the articles published from 2010 to 2019 to develop a framework for understanding how historical contents shape publication opportunities for researchers. We emphasize the usefulness of contents already published to understand journals' selection practices. Current statistical approaches to content analysis can grasp the usefulness of already published abstract articles or journal blurbs section as a path to drive further submission decisions and offer reliable measures of influence that may have potential policy implications.

Introduction
Turning a scientific manuscript into a published article and reaching multiple stakeholders is the greatest desire of any scientist. However, it is a tough decision to select an appropriate outlet to submit a research work. In the modern world where both knowledge and technology progress promptly, delays in the publication or wrong audience envisioned might negatively impact a yearly academic review and prevent a pioneering idea from entering the desired field.
By their side, journals publish articles selected by editors who ultimately depend on reviewers' opinions. Expert reviewers evaluate the rigor and value of new discoveries to gauge how they advance the field. Such peer-review constitutes an important approach to evaluating scientific output and it will continue to play a critical role in many forms of evaluation. For having an inbuilt quality filter, journal articles seem to be the most appropriate unit to count.
However, peer review is limited by its subjective nature and weakly correlates with the manuscripts true value (Starbuck, 2005). As a result, highly prestigious journals are publishing a considerable number of low-value articles while lower prestige ones are distributing some admirable papers. This random editorial selection process is also making outstanding manuscripts receive sequential rejections from different journals before being accepted.
The first attempts to describe the motivations of authors reassemble to the 1950s and 1960s when De Solla Price (1963) treating science as a measurable entity, developed some quantitative techniques and introduced the scientometrics concept. Later and to realise the main interests of authors when selecting a journal for submission purposes, Kochen and Tagliacozzo (1974) identified five basic factors which intervene in the choice of a journal: relevance, acceptance rate, circulation, prestige, and publication lag. Within the years, many other studies contributed for the corpus of knowledge in multiple perspectives. For instances, it was perceived publication timelines are field-dependent. Björk and Solomon (2013) determined submission-to-publication times were approximately twice as long in business and economy as in chemistry and the same happened also for earth sciences and chemistry (Garg, 2016).

The number of factors driving this judgement also changed as well as their importance. Rowlands and Nicholas (2005) found the top factors for senior authors were the journal's reputation, readership, and impact factor (IF). Solomon and Björk (2012) also surveyed authors to evaluate the importance of different factors when considering a journal. The most important factor was whether the research fit the scope of the journal, followed by the journal's quality/impact, speed of review and time-to-publication, type of readership, open access option, and likelihood of acceptance. Salinas and Munch (2015) reported journal prestige, likelihood of acceptance, turnaround time, target audience and IF.

Along with those studies, multiple web services were developed to support authors selecting a publishing venue. Besides of being free to use, they support scholars offering multiple measurements of performance to be used as filters for journals selection. Cofactor Journal Selector, created by the London-based firm Cofactor, leads users through a detailed list of filters to match author's publishing requirements (*Journal Selector | Edanz Group*, n.d.). Others like Elsevier Journal Finder or EndNote Match developed by Elsevier and Thomson Reuters respectively, requires user to input key pieces of information about the article to publish (e.g. title, abstract, keywords/phrases) and uses it to find the best matching journals (Kang et al., 2015). However, services provided by publishers are limited to their own pool of publications, assuming authors begin their decision process by first picking a publisher (Forrester et al., 2017).

To the best of our knowledge, scholarly journals are currently compared through four different means: 1) based on directly available information like IF calculated yearly by the Institute of Scientific Information; 2) data from publishers including acceptance rates, number of subscribers and Web load statistics; 3) data calculated from openly available information such as publishing fees and mean time from submission to publication; and 4) data obtained via surveys with authors who have experienced in publishing in a specific journal (Björk & Öörni, 2009).

Since all these methods are not content sensitive, we propose to match descriptive data ('external') with the 'real fine content' ('internal') of a set of journals based on semantic document classification strategies. Taking advantage of the short promotional statement self-prepared by the journals, the blurbs, as well as the journals' portfolio made available through the acknowledged sources of scientific information like the Web of Science platform, we aim we propose a novel recommendation system for those seeking to publish their scientific work.

Bringing some content-sensitivity to journals' comparison

The exponential growth of the number of scientific publications accompanied with the huge progression on the number of scientific outlets makes it difficult for researchers to decide about the journal to choose for publishing their works (Bornmann & Mutz, 2015; Evans, 2013; Gu & Blackmore, 2016; Shiffrin et al., 2018). Several metrics were developed to measure journals impact, importance and ultimately to guide science makers about the channels to submit their novels (Bornmann & Marx, 2015).

Journal IF, proposed by Garfield (1972), was one important metrics guiding authors (Eugene Garfield, 2006). In ecological field, 85,6% of authors revealed that a journal with a high IF is a 'very important' to 'important' criterion when selecting an outlet to submit their works (Aarssen et al., 2008). And, in a large-scale survey, covering 923 scientific journals between 2006 and 2008, it was found a resubmission pattern suggesting a flow from higher to lower IF journals (Calcagno et al., 2012). Seeking to maximize citation counts, researchers choose to submit their works to the journal with the highest IF and then work down the IF list as the manuscript is rejected. However, this kind of strategy ignores the value of following actions and the opportunity costs involved (i.e., every time a paper is rejected). It is seen as a poor predictor of the ultimate success (Wang et al., 2013). Among the IF's limitations, it should be considered as a measure of scientific use by other researchers rather than scientific quality (Callaway, 2016).

However, one of the limitations noticed was the content insensitivity. Disregarding the matter of which journals are made of, it biases the authors' decisions about submissions. Computer-aided screening and analysis might help to both overview and classify contents more efficiently. Already applied in multiple environments ranging from tweets and other personal opinions to sports news and movie reviews, machine learning algorithms has enabled efficient text classification encompassing subjective sentence contents, source identification and sentiment expression classification.

Machine learning in general allows for classifying data into specifically predefined classes based on manually annotated training data. The algorithm then can identify by itself key features specific of the defined classes. For text analysis, the words used, and their frequencies commonly represent the entry data. This method is called "bag of words". The current work was planned to assess the current potential of machine learning to extract content from articles published and compare it with journals blurb section. Machine learning encompasses several models that are implemented in code in different ways. For this purpose, we selected the Rain Forest model.

Innovation-oriented journals as the interest topic

More than half a century old, innovation studies as a research field emerged from different disciplines as Economics (Nelson, 1959), Management (Burns, 1961) and Sociology (Rogers, 1983) bringing such heterogeneity along (Fagerberg et al., 2012). This interdisciplinary was its nature as an eclectic borrowing of cognitive resources was used in its progress (Martin, 2012). Journals from multiple scientific fields are publishing articles related to innovation studies topic.

With the purpose of identify the outlets which accept and publish the most works on such an area of interest, Fagerberg et al. (2012) studied both the most productive as well as those using further innovation findings. Supporting the establishment of innovation studies as a field, we restricted our analyses to the twenty most influential journals Fagerberg et al. (2012) identified. Table 1 catalogues the principal set of outlets with their subject area, year of launch and quartile they occupy currently according to the IF achieved.

Although almost all journals are in the first quartile (the exception is *Technology Analysis & Strategic Management*) evidencing a prestigious position among their pairs, the subject categories where they are coming from vary greatly. The Business, Management and Accounting category attracts sixteen of our outlets and Economics, Econometrics and Finance are the common interest for three. Both *Administrative Science Quarterly* and *Human Relations* fit on Arts and Humanities and *Regional Studies* are set on Environmental Science and Social Sciences categories. For single journals, fields like Computer Science, Psychology and Engineering are also targets.

Believing such a set of outlets has sufficient in common to be addressed together but also varied interests to be distinguishable, we propose to use their both advertisement section, named *blurb*, and produced contents as objects for publication contents comparison. The current project aims to evaluate the potential of machine learning to assess the similarity between the contents published and the ones described in the blurb section designed to attract authors interest for submission.

Table 1. List of journals.

Short name	Journal	Subject area
AMJ	Academy of Management Journal	Business, Management and Accounting
AMR	Academy of Management Review	Business, Management and Accounting
ASQ	Administrative Science Quarterly	Arts and Humanities; Social Sciences
CJE	Cambridge Journal of Economics	Economics, Econometrics and Finance
HR	Human Relations	Arts and Humanities; Business, Management and Accounting; Social Sciences
ICC	Industrial and Corporate Change	Economics, Econometrics and Finance
IJTM	International Journal of Technology Management	Business, Management and Accounting; Computer Science; Engineering; Social Sciences
JIBS	Journal of International Business Studies	Business, Management and Accounting; Economics, Econometrics and Finance
JMS	Journal of Management Studies	Business, Management and Accounting
MS	Management Science	Business, Management and Accounting; Decision Sciences
OSc	Organization Science	Business, Management and Accounting
OSt	Organization Studies	Business, Management and Accounting
RDM	R&D Management	Business, Management and Accounting
RS	Regional Studies	Environmental Science; Social Sciences
RP	Research Policy	Business, Management and Accounting; Decision Sciences; Engineering
SBE	Small Business Economics	Business, Management and Accounting; Economics, Econometrics and Finance
SMJ	Strategic Management Journal	Business, Management and Accounting
TASM	Technology Analysis & Strategic Management	Business, Management and Accounting; Decision Sciences
TFSC	Technological Forecasting and Social Change	Business, Management and Accounting; Psychology
Tec	Technovation	Business, Management and Accounting; Engineering

Developing a model for blurbs classification

Model conception

This section presents the details of the proposed techniques for matching abstracts from articles published with blurbs sections. The rationales behind these techniques selected are also discussed. Since the accuracy of machine learning algorithms is known to improve with greater quantities of data to train on, articles' abstracts were used to train the model and test it for overall accuracy determination. Later, the algorithm will be applied to match journals blurbs to the related articles abstracts.

In this work, we propose to use a supervised text classification method as we can train the model with the articles' abstracts extracted associated to the outlets titles which published them. The algorithm will learn from the labelled training data to predict outcomes for unforeseen data.

The full process is illustrated in Figure 1 and encompasses five smaller steps which will be addressed in sequence.

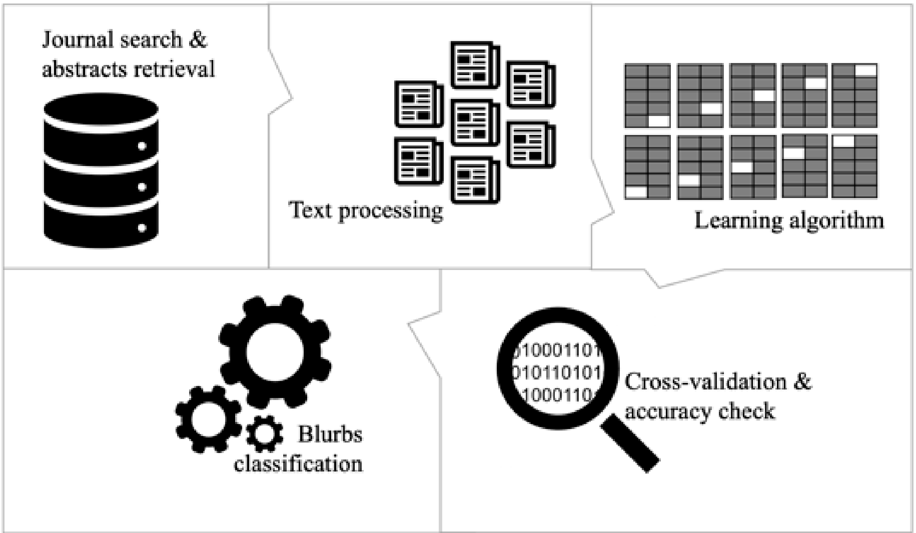

Figure 1. Supervised learning process applied.

With this approach, a training abstracts dataset (previously classified into journals) is used to identify unique patterns that represent each top-tier, and then use these identified patterns to correctly predict the outlet a future instance will belong to.

Journal's search & abstracts retrieval

We searched on Web of Science for the all the titles published in these twenty outlets from 2010 to 2019 (Norris & Oppenheim, 2007). For each one, we extracted the title, the list of authors, the abstract and year of publishing. Only articles with all these contents available were considered. At the end, we realised each journal published different numbers of articles per year and it changed over the years. The Figure 2 shows the number of articles published in each outlet during this ten-years period and the overall volume of articles published for all of them.

Outlets presented different number of published articles and an overall trend to increase over the years. Both TFSC and MS presented a significant increase in the number of accepted manuscripts for publication, publishing less than 150 articles in 2010 and publishing more than 300 in 2019. Some titles were also found with a more modest growth such as AMR and CJE which presented 28 and 64 articles published in 2010 and 34 and 70 in 2019, respectively.
Such increasing trend was anticipated by de Solla Price (1961) who first published quantitative data about the growth of science from 1650 to 1950 with a growth rate of 5.6% per year and forecasted an increasing in journals number reaching one million by 2000. Also interesting are outlets like IJTM which published more in 2010 than in 2019.

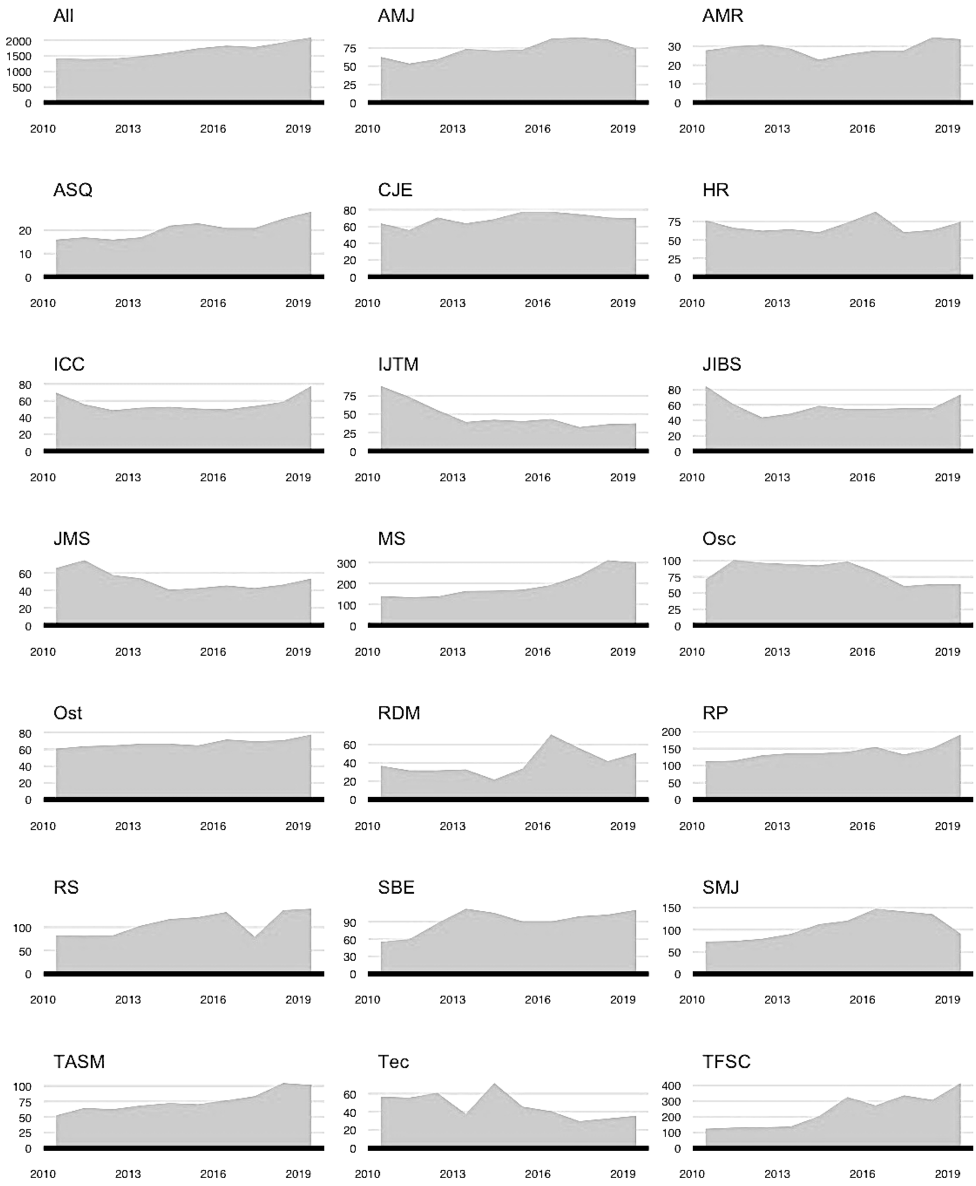

Figure 2. Number of articles published in each outlet from 2010 to 2019.

Text processing

As the classifiers and learning algorithms cannot directly process the text documents in their original form, we need to transform our raw text documents with variable length into numerical feature vectors with a fixed size. Therefore, abstracts collected were pre-processed and converted into a more manageable representation.

This first major step involved four smaller tasks: 1) word tokenization; 2) case transformation; 3) stop-word removal and 4) stemming. Word Tokenization aims to separate words as tokens.

In a simple way to perform tokenization is to assume [space] as separators between words or tokens. Case transformation involves changing all capitalized letters in a lower-case format. Stop-word removal aims to delete the most frequent words that occur in the English language like *the*, *and*, *a*, *is*... In addition, we also removed "paper" as it was quite frequent. These so common words do not bring any meaning to the text which allow us to exclude them without undesired impact. Finally, stemming reduces the inflected words to their root form or stem. In this study, we applied the Porter Stemmer algorithm, one of the most used stemming algorithms (Porter, 1980).
The pre-process step is considered a critical one as it affects the speed and accuracy of a learning algorithm. At the end, it is expected datasets no longer have high levels of noise or unmeaning stems which will be used as features. With this clean dataset, each feature (word) was mapped to the corresponding number of occurrences (term-frequency) within the whole text. This is called "word embedding" process and converts each abstract to a vector of the same length containing the frequency of the words. These vectors will be used to train the model and later, classify blurbs.

Learning algorithm

Among the supervised learning algorithms known, those combining multiple learning models using either boosting or bagging techniques have shown to achieve better results than simple learning models (Caruana & Niculescu-Mizil, 2006). For this work, Random Forests was the one which outperform other learning algorithms (Breiman, 2001). Although it does not consider the sequence of instances to be classified in sequential labelling tasks, it does not present a major problem for our project.
While correcting for decision trees' inherent problem of over-fitting of training examples, Random Forest was operated by constructing a multitude of decision trees during training, then outputting the class that was the mean prediction of the individual decision trees. Like any other supervised learning techniques, the goal of Random Forest model is to identify patterns in a training set and then use these identified patterns to predict future unseen cases.
The main assumption of machine learning is that the distribution of training data is identical to the distribution of test data and future examples. If the learning algorithm accurately classifies the set used for testing, then the machine learning assumption suggests that it will perform as well for future unseen cases. For the training purposes, our classification dataset made of abstracts was split into two disjoint sets and 70% of the observations were used for training and 30% for testing.

Cross-validation & accuracy check

Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.
Training accuracy was determined by a ten-fold cross validation (Arlot & Celisse, 2010; Kohavi, 1995). In a ten-fold cross validation, the input data is divided into ten equal disjoint subsets. Each subset is used as the test set while all the others are used as the training set. For each validation, accuracy is measured by the ratio between the "number of correctly classified cases" and the "total number of cases". The final accuracy will be the average accuracy of the ten different subsets. Depending on the application the learning model is designed for, the minimum acceptable accuracy level may be different. The overall accuracy of our model was 80% which means our model can correctly predict four out of five abstracts. Previous articles have described and used accuracy measures that are widely recognized as the standard way to determine the test and training accuracies of prediction models (Mullen & Collier, 2004).

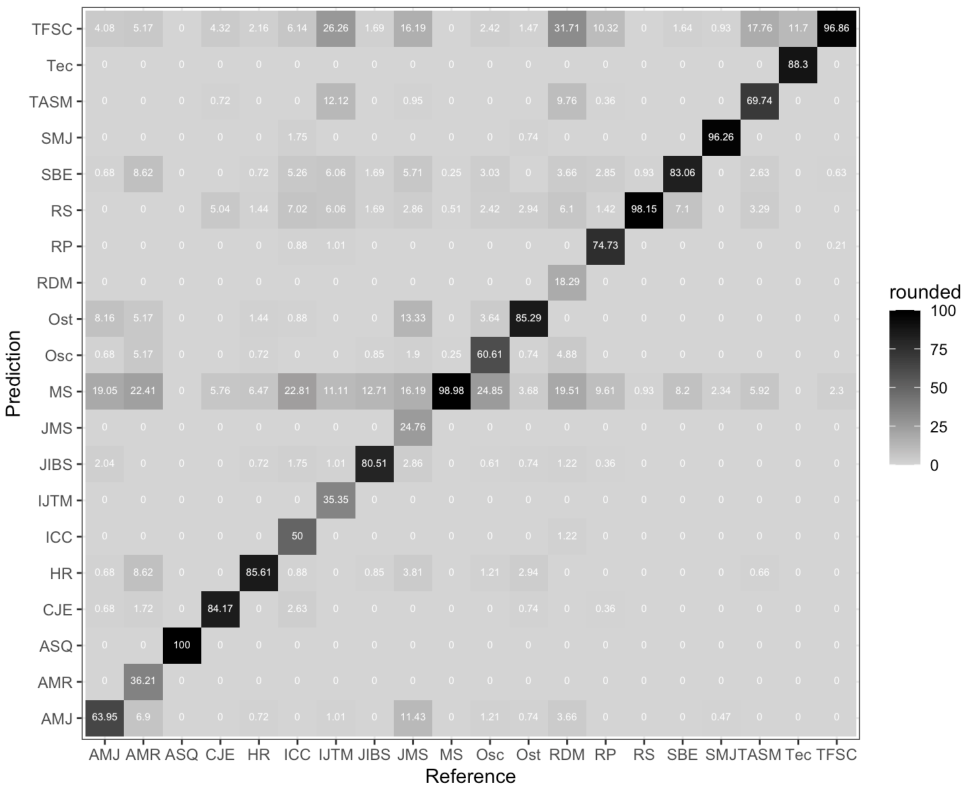

Figure 3. Abstracts' classification.

Although the overall accuracy was 80%, the accuracy achieved by model classifying the abstracts belonging to each specific journal varies greatly. Figure 3 shows all the *Administrative Science Quarterly*'s abstracts are correctly classified. Closer to this figure are the abstracts published on *Management Science* (98,98%), *Regional Studies* (98,15%) and *Technological Forecasting and Social Change* (96,86%). Nevertheless, the model performance was not so good predicting the abstracts published on *R&D Management*, *Journal of Management Studies*, *International Journal of Technology Management* and *Academy of Management Review*.
Such accuracy differences presented by the model are explained by the type of contents published. The more specialized journals are greater the predictable accuracy achieved by the model. Journals publishing a larger number of heterogeneous topics are classified as the publishers of abstracts from other outlets. This happens with *Management Science* and *Technological Forecasting and Social Change*. Both are predicted to have published great numbers of abstracts from other journals. Stating their eclectic and diverse interest, their editorial announcement is not only being defined as generalist but also as direct competitors of other journals.

Blurbs' classification

Our analysis showed that the current model can be efficiently used to match abstracts with the journals which published them. Abstracts previously collected were used to train and test the model, the algorithm was trained again considering this larger dataset. All pre-process steps were repeated before the training. This time the test dataset was the blurbs sections collected from the outlets. The model was asked to associate the blurbs according to the most semantically similar journal abstracts collection. Figure 4 presents the classifications suggested by the model.

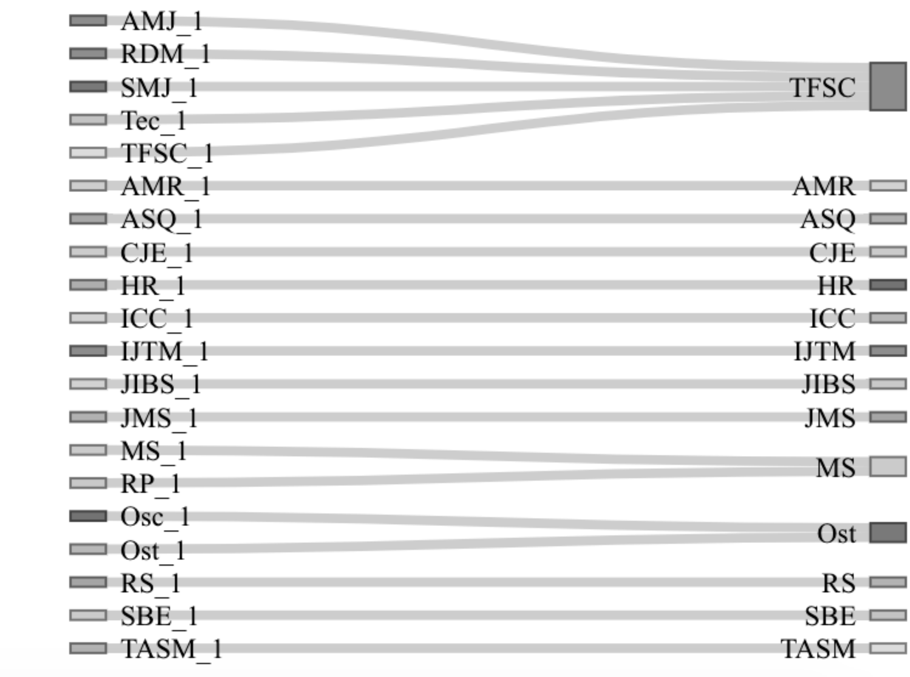

Figure 4. Blurbs' classification.

Within the twenty journals sample, fourteen blurbs were correctly linked to the journals they belong to. *Technological Forecasting and Social Change, Management Science* and *Organization Studies* were associated with blurbs from other journals besides their own's. The model recognized *Management Science* and *Research Policy* blurbs sections like *Management Science's* abstracts from ten years of published articles. The model also recognized blurbs from *Academy of Management Journal*, *R&D Management*, *Strategic Management Journal*, *Technovation* and *Technological Forecasting and Social Change* as being related to the last one. Blurbs from *Organization Science* and *Organization Studies* were also associated with *Organization Studies* published abstracts.
Following such results, an author may also wonder about submit a manuscript designed for *Research Policy* also to *Management Science*. It seems manuscripts from both top tiers are poorly differentiated and confusing to distinguish the outlets which published them. Thus, all manuscripts are classified as part of *Management Science* portfolio. As both are in Q1 according to the IF in their fields, a manuscript in the scientific area of *Research Policy* may be also interesting for *Management Science.* The same could be also true for the five journals which had their blurbs' section associated with the contents of *Technological Forecasting and Social Change.* Considering the contents previously published besides the journals' external metrics available may provide additional opportunities for researcher submissions.

Discussion

Given the constantly growing stream of scientific journals' data, automatic classification of unstructured text data into relevant researcher-defined content categories is likely to continue charming scientists. The results of our analysis both confirms it is possible to recognize journals contents by their blurbs and suggests blurbs are distinctive pieces of information able to judge journals interests.

In terms of the accuracy of determining the web page type using machine learning techniques, Random Forest achieved up to 80% accuracy. The automatic classifier, however, misclassifies articles from *Management Science* journal. The number of abstracts used to train the model cannot explain this behaviour as this journal was pointed out as one the journals with higher number of articles published over the years and with a significant increase. Our suggestion is that *Management Science* publishes a great variety of scientific topics misleading the algorithm when recognising the journals which published a manuscript.

Organization Studies showed to be also the outlet which could also publish the works a research may think about submit to *Organization Science. Technological Forecasting and Social Change* seems to be the outlet which could encompass choices from four different top-tiers: *Academy of Management Journal*, *R&D Management*, *Strategic Management Journal*, *Technovation* and *Technological Forecasting and Social Change*. The sematic similarity between these ones, make *Technological Forecasting and Social Change* an alternative option to the three others.

There are of course limitations to this research. First, we considered only the ten years of data from articles available on Web of Science from the Innovations Studies, a field which emerged a few decades ago. Considering other older scientific fields and larger data published may bring more accurate classifications. Second, for the algorithm training, we used only the abstracts of full articles which may differ from the usage of full articles contents. All these issues are also ideas for further studies to enlarge even more the scientometric approaches able to provide contents sensitive methods for journals comparison.

Concluding Remarks

Text classification, that is computer-aided analysis of textual data, offers a great opportunity to advance journal selection strategies. This motivated us to apply such technique to a real problem: compare published contents with the advertisement section designed to attract further submissions.

However, some limitations should be reported and overcome in the future works. First and foremost, the content period of this project is set as ten years (ranging between 2010 and 2019), which was mainly driven by the consideration of balancing the size of data and the computational capacity.

We applied a machine learning due to its great potential for efficient classification. However, a drawback was noticed for this method. It lacks transparency, which prevents identification of causal factors (Liu et al., 2019).

For future exploration of the potential of machine learning algorithms to compare journal contents, a larger set of annotated training abstracts, blurbs and journals would be necessary. For the current proof of concept, the "Abstracts" and Blurbs sections were manually taken from journals' websites. Automatic extraction will efficiently generate a larger data sets and may bring such methodology to other areas of science knowledge, benefiting researchers from areas beyond innovation studies.

References

Aarssen, L. W., Tregenza, T., Budden, A. E., Lortie, C. J., Koricheva, J., & Leimu, R. (2008). Bang for Your Buck: Rejection Rates and Impact Factors in Ecological Journals. *The Open Ecology Journal*, *1*(1), 14–19. https://doi.org/10.2174/1874213000801010014

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79. https://doi.org/10.1214/09-SS054

Björk, B. C., & Öörni, A. (2009). A Method for Comparing Scholarly Journals as Service Providers to Authors. *Serials Review*, *35*, 62–69. https://doi.org/10.1080/00987913.2009.10765213

Björk, B. C., & Solomon, D. (2013). The publishing delay in scholarly peer-reviewed journals. *Journal of Informetrics*, *7*(4), 914–923. https://doi.org/10.1016/j.joi.2013.09.001

Bornmann, L., & Marx, W. (2015). Methods for the generation of normalized citation impact scores in bibliometrics: Which method best reflects the judgements of experts? *Journal of Informetrics*, *9*(2), 408–418. https://doi.org/10.1016/j.joi.2015.01.006

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), 2215–2222. https://doi.org/10.1002/asi.23329

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Burns, T. (1961). *The management of innovation*. Tavistock Publications.

Calcagno, V., Demoinet, E., Gollner, K., Guidi, L., Ruths, D., & de Mazancourt, C. (2012). Flows of research manuscripts among scientific journals reveal hidden submission patterns. *Science*, *338*(6110), 1065–1069. https://doi.org/10.1126/science.1227833

Callaway, E. (2016). Beat it, impact factor! Publishing elite turns against controversial metric. *Nature*, *535*(7611), 210–211. https://doi.org/10.1038/nature.2016.20224

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *ACM International Conference Proceeding Series*, *148*, 161–168. https://doi.org/10.1145/1143844.1143865

de Solla Price, D. J. (1961). *Science since Babylon*. Yale U.P.

de Solla Price, D. J. (1965). *Littel Science, Big Science*. Columbia University Press.

Evans, J. A. (2013). Future science. *Science*, *342*(6154), 44–45. https://doi.org/10.1126/science.1245218

Fagerberg, J., Fosaas, M., & Sapprasert, K. (2012). Innovation: Exploring the knowledge base. *Research Policy*, *41*(7), 1132–1153. https://doi.org/10.1016/J.RESPOL.2012.03.008

Forrester, A., Björk, B.-C., & Tenopir, C. (2017). New web services that help authors choose journals. *Learned Publishing*, *30*(4), 281–287. https://doi.org/10.1002/leap.1112

Garfield, E. (1972). Citation Analysis as a Tool in Journal Evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science*, *178*(4060), 471–479. https://doi.org/10.1126/science.178.4060.471

Garfield, Eugene. (2006). The History and Meaning of the Journal Impact Factor. *JAMA*, *295*(1), 90. https://doi.org/10.1001/jama.295.1.90

Garg, K. C. (2016). Publication delay of manuscripts in periodicals published by CSIR-NISCAIR. *Current Science*, *111*(12).

Gu, X., & Blackmore, K. L. (2016). Recent trends in academic journal growth. *Scientometrics*, *108*(2), 693–716. https://doi.org/10.1007/s11192-016-1985-3

Journal Selector | Edanz Group. (n.d.). Retrieved July 12, 2019, from https://www.edanzediting.com/journal-selector

Kang, N., Doornenbal, M., & Schijvenaars, B. (2015). Elsevier journal finder: Recommending journals for your paper. *RecSys 2015 - Proceedings of the 9th ACM Conference on Recommender Systems*, 261–264. https://doi.org/10.1145/2792838.2799663

Kochen, M., & Tagliacozzo, R. (1974). Matching authors and readers of scientific papers. *Information Storage and Retrieval*, *10*(5–6), 197–210. https://doi.org/10.1016/0020-0271(74)90059-X

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence*, *14*(12), 1137–1143.

Liu, Y., Chen, P. H. C., Krause, J., & Peng, L. (2019). How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *JAMA*, *322*(18), 1806–1816. https://doi.org/10.1001/jama.2019.16489

Martin, B. (2012). The evolution of science policy and innovation studies. *Research Policy*, *41*(7), 1219–1239. https://doi.org/10.1016/J.RESPOL.2012.03.012

Mullen, T., & Collier, N. (2004). Incorporating topic information into sentiment analysis models. *Empirical Methods in Natural Language Processing*, 412–418.

Nelson, R. R. (1959). The Simple Economics of Basic Scientific Research. *Journal of Political Economy*, *67*(3), 297–306. https://doi.org/10.1086/258177

Norris, M., & Oppenheim, C. (2007). Comparing alternatives to the Web of Science for coverage of the social sciences' literature. *Journal of Informetrics*, *1*(2), 161–169. https://doi.org/10.1016/j.joi.2006.12.001

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130– 137. https://doi.org/10.1108/00330330610681286

Rogers, E. M. (1983). *Diffusion of Innovations* (3rd ed.).

Rowlands, I., & Nicholas, D. (2005). Scholarly communication in the digital environment: The 2005 survey of journal author behaviour and attitudes. *Aslib Proceedings*, *57*(6), 481–497. https://doi.org/10.1108/00012530510634226

Salinas, S., & Munch, S. B. (2015). Where Should I Send It? Optimizing the Submission Decision Process. *Plos One*, *10*(1), e0115451.

Shiffrin, R. M., Börner, K., & Stigler, S. M. (2018). Scientific progress despite irreproducibility: A seeming paradox. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2632–2639. https://doi.org/10.1073/pnas.1711786114

Solomon, D. J., & Björk, B.-C. (2012). Publication fees in open access publishing: Sources of funding and factors influencing choice of journal. *Journal of the American Society for Information Science and Technology*, *63*(1), 98–107. https://doi.org/10.1002/asi.21660

Starbuck, W. H. (2005). How much better are the most-prestigious journals? The statistics of academic publication. *Organization Science*, *16*(2), 180–200. https://doi.org/10.1287/orsc.1040.0107

Wang, D., Song, C., & Barabási, A. L. (2013). Quantifying long-term scientific impact. *Science*, *342*(6154), 127–132. https://doi.org/10.1126/science.1237825