Where Will Your Paper Go? Predicting Journal Submission and Citations with Topic Models

Zhe Guan[a,1]

This manuscript was compiled on January 6, 2025

Academic impact prediction is an important task for academic scholars, funding agencies, and journal editors to guide funding arrangements, research priorities, and editorial decisions. This study presents a framework that combines topic modelling techniques, including structural topic modelling (STM) and latent Dirichlet allocation (LDA), with machine learning classifiers while integrating multiple clustering algorithms, such as K-means and hierarchical clustering, to predict both the number of citations a paper will receive and the most suitable journal for submission. Additionally, an interactive dashboard tool is provided to allow users to input paper abstracts and obtain predictions using different models.

Topic Modeling | Machine Learning | Citation Prediction | Journal Classification

Introduction

The exponential growth in the number of academic papers in different journals and on various topics brings unignorable challenges for researchers and policymakers who arrange funding and research priorities in institutions(1). There is an increasing need for intelligent tools that can analyse and predict both citation potential and appropriate target journals for academic papers(2). To address this, we focus on three journals within the field of operations research and systems analysis: the "Journal of the Operational Research Society", "Health Systems", "Journal of Simulation". The journals were selected for their relevance to decision analytics, healthcare systems, and simulation methodologies.

1. Dataset and Cleaning

The analysed dataset in this work includes the 4141 records. The initial features are shown in Table 1.

Table 1. The initial features in dataset from journals

Features	Description
Title	Title of the paper
Journal	Name of the journal
Year	Published year
Pages	Number of pages in the paper
Authors	Names of the authors
Views	Number of views the paper received
Citations	Number of times the paper was cited
Altmetric Score	Altmetric score indicating social and public engagement
Abstract	Summary of the paper's content

Since the title name may reveal useful information for topic modelling, the two variables, "Title" and "Abstract," are combined into a single input variable "Title Abstract". And "Authors" are transformed into "Author_Counts" which can provide more useful information. Different techniques are applied to fill the missing 'Views' values in the train and test data, with details provided in the regression section. Considering that one of our targets is predicting journal names, the two variables, the "Views" and "Almetric Score," are excluded from the model from the classification model, as these metrics are not available prior to submission. A unique identifier, "row_id", was added to each record. "Title_Abstract" was transformed into lemmatisation to standardise word forms first. Next, stop words and numeric tokens were removed to reduce noise in the dataset.

2. Topic Modelling

Bag of words and Comparisions.
The combined text was tokenised into individual words and bigrams using a bag-of-words(BoW) approach (3). The top 100 frequencies of unigrams and bigrams in the dataset are shown in Figure 1. The bigrams can provide more information about differences among journals compared to common unigrams.

Fig. 1. Top 100 frequencies of unigrams and bigrams from datasets

The BoW frequency of bigrams is analysed from two perspectives: its variation across different journals and over four time periods, as shown in Figure 2 and Figure 3.

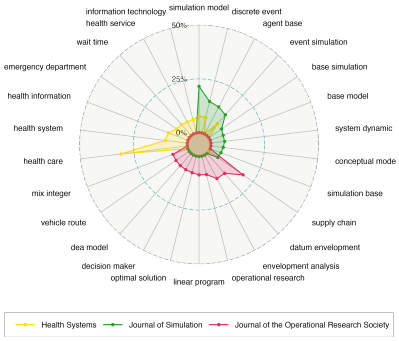

Fig. 2. The top 10 most frequent bigrams across three journals.

Different areas are focused on across journals, and the high-frequency bigrams show obvious distinctions, which can help

Author affiliations: [a]The University of Southampton

[1] E-mail: zg2u24soton.ac.uk

in developing future classification models to identify the most appropriate target journal. Some common bigrams still exist across journals, such as “simulation model” and “event simulation” appearing between the journal *Health Systems* and *Journal of Simulation*, as well as “supply chain” shared between *Journal of Simulation* and *Journal of Operational Research Society*.

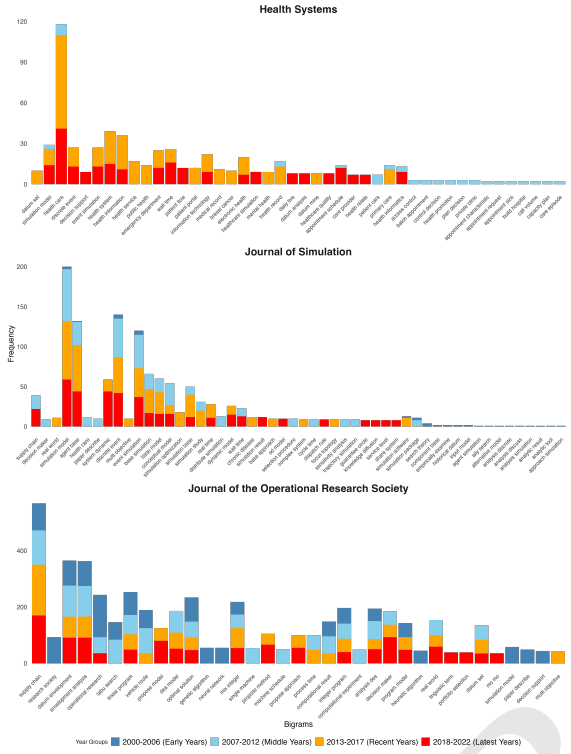

Fig. 3. The top 30 most frequent bigrams from the dataset across four different time periods: 2000–2006, 2007–2012, 2013–2017, and 2018–2022 for three journals.

Some interesting phenomena appear over time across three journals. Firstly, there is no paper published in the journal *Health Systems* before the Middle period (2007-2012), and “health care” is the main topic of the journal and shows increasing popularity across years. Secondly, the key bigrams in the *Journal of Simulation* are relatively concentrated and increasingly popular across years, such as “simulation model”, “discrete event”, and “event simulation”, whereas the *Journal of the Operational Research Society* shows more flat probabilities across bigrams, with more focus on terms like ”supply chain.”

Latent Dirichlet Allocation. Latent Dirichlet Allocation (LDA) (4) is a widely used topic modelling algorithm that identifies hidden topics in a collection of documents based on word co-occurrence patterns. It assumes that each document is a mixture of topics, and each topic is a distribution over words. LDA requires specifying a value of *k* (the number of topics) in the corpus as shown in Figure 4. Determining an appropriate value for *k* is significant, as it probably impacts the quality and interpretability of the topic modelling results.

Perplexity is a common metric used to evaluate the fit of topic models and provides a reference for selecting the optimal *k* value in LDA.

Lower perplexity generally indicates a better model fit. As shown in Figure 5, the *2000-2022* year group has a much higher

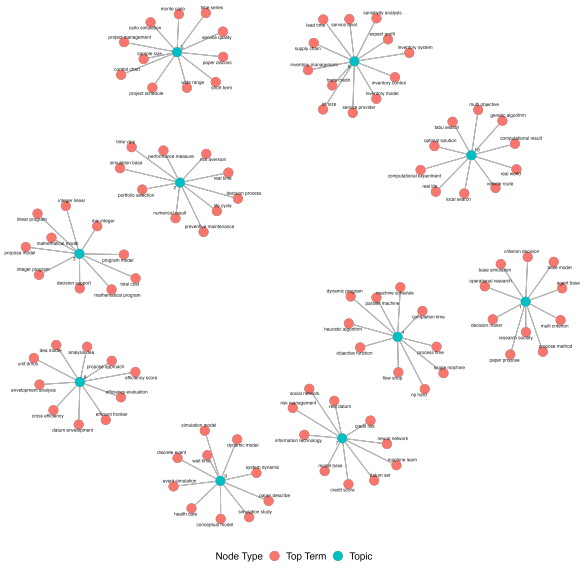

Fig. 4. The identified topics using LDA with the number of topics $k = 10$.

perplexity compared to other year groups, indicating that over a longer time span, the data is more diverse, and the model finds it harder to capture consistent topics across the entire period. Similarly, the *Journal of the Operational Research Society* has a much higher perplexity compared to other journals, indicating that the journal covers a broader and more diverse range of topics, making it harder for the model to find consistent patterns.

Furthermore, Figure 5 also shows that considering dataset features can help reduce perplexity and the optimal number of topics.

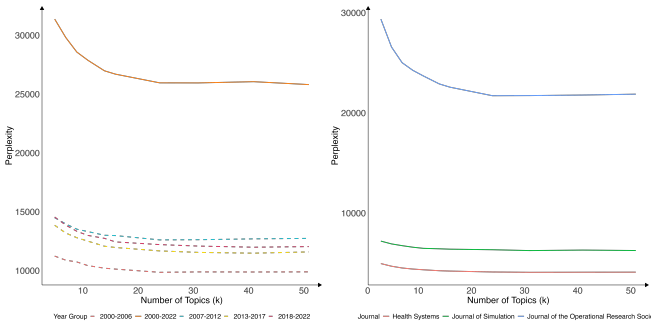

Fig. 5. The perplexity with different numbers of topics across years and journals.

Structural Topic Modelling.

Structural Topic Modelling (STM) (4) provided a method to incorporate document-level metadata (e.g., publication year, journal name) into the topic modelling process, which builds on the foundation of LDA. Several metrics can be used to evaluate the optimal k (5), such as *Exclusivity*, *Heldout Likelihood*, *Residuals*, and *Semantic Coherence* as shown in Table 2.

The STM metrics suggest that $k = 10$ is a reasonable choice, as the majority of the metrics remain stable at this point compared to the stable case of $k = 20$ using LDA shown in Figure 5.

Table 2. Comparison of STM Metrics for Different K Values

K	Exclus	Semcoh	Heldout	Residual
2	7.93	-165.02	-8.90	61.93
4	8.82	-184.20	-8.92	47.10
6	9.41	-236.94	-8.97	36.15
8	9.53	-230.01	-8.94	29.66
10	9.67	-219.33	-8.86	24.15
12	9.76	-219.71	-8.84	19.81
14	9.76	-216.37	-8.81	16.59
16	9.80	-223.70	-8.88	14.75
18	9.79	-213.05	-8.84	12.42
20	9.73	-223.25	-8.98	11.30

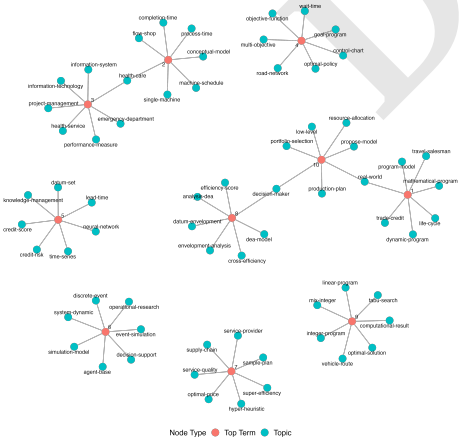

Fig. 6. The identified topics using STM with the number of topics $k = 10$.

The similarity S of topics i and j between two models can be evaluated by Formula 1 (6):

$$S_{ij} = \frac{\sum_{w=1}^{n} A_{wi} B_{wj}}{\sqrt{\sum_{w=1}^{n} A_{wi}^2} \times \sqrt{\sum_{w=1}^{n} B_{wj}^2}} \quad [1]$$

Where A_{wi} and B_{wj} are the probabilities $\beta_{wi} = P\left(\text{word}_w \mid \text{topic}_i\right)$ and $\beta_{wj} = P\left(\text{word}_w \mid \text{topic}_j\right)$ of the w^{th} word from the LDA model and STM model.

The similarity matrix shown in Figure 7 suggests that 50% of topics have a similarity greater than 0.7, and 80% of topics have a similarity greater than 0.4 between the two models with the same number of topics $k = 10$.

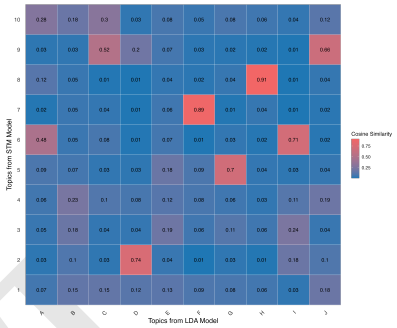

Fig. 7. The similarity matrix between LDA and STM models with topic number $k = 10$.

One of the benefits of the STM model is that it can easily reveal the topic trend with data features like time. As shown in Figure 8, topics 1, 10, 7, and 8 demonstrate an increasing trend in popularity over the years. In contrast, topics 2, 3, 4, and 5 suggest a potential decline in popularity, although with wide uncertainty bands. Notably, topics 6 and 9 show clear and consistent decreases in popularity over time.

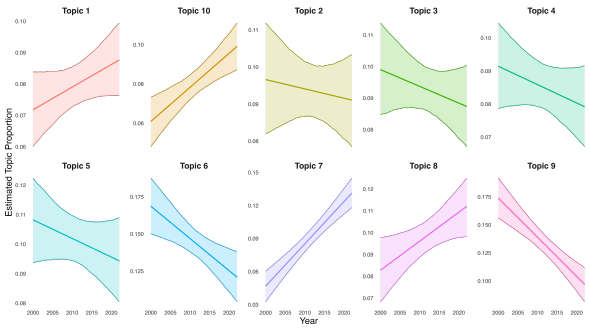

Fig. 8. Topic trends over time with the STM model and topic number $k = 10$.

Citations Prediction with Regression

Preprocessing.

The topic probabilities generated by the two models were combined with the original data as features for regression analysis. The dataset was split into 80% training data and 20% test data to evaluate model performance. Additionally, the Journal Name column was transformed using one-hot encoding to convert categorical values into binary features.

The missing 'Views' values in the training data are filled using predictions from a linear regression model based on 'Citations,' while in the test data, they are replaced with mean values.

Linear Regression.

Although two topic models, LDA and STM, have been trained, the similarity may lead to redundancy in the features generated from both models with the same k value. The redundancy can negatively affect regression models (7) due to multicollinearity. Principal Component Analysis (PCA) (8) is applied to reduce the dimensionality of the combined topic distributions and transform correlated features into a set of uncorrelated principal components. As shown in Figure 9, the first 23 PCA components are selected to train the linear regression model.

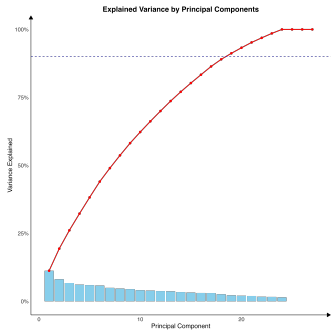

Fig. 9. Explained and cumulative variances by principal components.

XGBoost. eXtreme Gradient Boosting (XGBoost) (9) uses gradient boosting, minimising a loss function by adding new trees that predict the residual errors of the previous trees. However, the effectiveness of this process heavily depends on the selection of appropriate hyperparameters. In this work, a grid search is applied, as shown in Table 3 for optimal choice.

RandomForest. Compared to XGBoost, Random Forest (10) takes a different approach by employing an ensemble of independent decision trees. Similar to XGBoost, the performance of Random Forest is sensitive to hyperparameter selection. A grid search is similarly conducted, as shown in Table 3, to identify the best combination of hyperparameters for Random Forest.

Table 3. The grid scan for XGBoost and Random Forest

Model	Parameter	Description	Scan range
XGBoost	nrounds	Number of boosting rounds	100, 300, 500
	η	Learning rate	0.005, 0.01, 0.03, 0.1
	max_depth	Maximum depth of trees	5, 8, 10
	γ	Minimum loss reduction	0, 1
	colsample_bytree	Column subsample ratio	0.8, 1.0
	min_child_weight	Minimum sum of instance weight	1, 5, 10
	subsample	Row subsample ratio	0.8, 1.0
Random Forest	mtry	Number of variables tried at each split	2, 5, 10
	splitrule	Splitting rule	"variance", "extratrees"
	min.node.size	Minimum size of terminal nodes	1, 5, 10

The three models are evaluated using four metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), and Residuals, as shown in Figure 10. The linear regression model shows the poorest performance due to the non-linear relationships between the target variable and the predictors.

3. Classification

Clustering for more information.

Clustering (11) helps capture hidden patterns or latent groupings in the data that are not immediately obvious from the original features. These patterns can provide useful additional information for classification models to make better predictions.

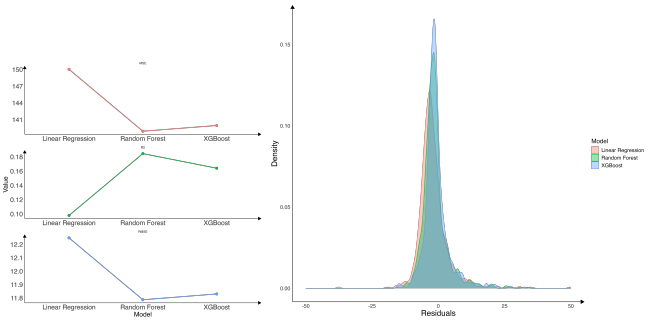

Fig. 10. MSE, RMSE, R^2, and residual density for regression model evaluations.

In this work, a combined clustering approach is applied. The Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) model (12) is first used to generate initial clusters. The cluster labels obtained from HDBSCAN are then used as an input feature for the KMeans method (13) to build a second clustering model. The top three words with the highest mean Term Frequency-Inverse Document Frequency (TF-IDF) (14) scores for each cluster are presented in Table 4. Additionally, the corresponding journal names are displayed in Table 5, which shows that Cluster 3 is associated with the "Journal of the Operational Research Society," while Clusters 1 and 2 provide useful insights for the other two journals.

Table 4. Top Words with Highest TF-IDF Scores by Cluster

Cluster	Word	TF-IDF Score
1	supply chain	0.0105
1	operational research	0.0108
1	datum envelopment	0.0077
2	supply chain	0.0102
2	operational research	0.0090
2	simulation model	0.0073
3	research society	0.0290
3	operational research	0.0275
3	supply chain	0.0092

Table 5. Number of Articles by Journal and Cluster

Cluster	Health Systems Journal	Journal of Simulation	Journal of the Operational Research Society
1	122	164	963
2	86	217	1281
3	0	23	1285

Before classification, the training data is created by selecting 80% of the original dataset, combined with cluster labels, STM and LDA topic predictions, and a binary variable indicating whether the document contains journal-specific frequent bigrams, as shown in Figure 2.

K-Nearest Neighbours. K-Nearest Neighbours (KNN) (15) is a non-parametric, distance-based algorithm commonly used for classification tasks. The class label of a test sample is determined by a majority vote among its 20 nearest neighbours from the training data. The Euclidean distance metric was used to measure the similarity between samples.

Multinomial Logistic Regression. Multinomial Logistic Regression (LG) (16) is a supervised learning algorithm used for multi-class

classification. The model predicts the class label by estimating the probability of each possible class. In this study, the predictions were trained by selecting the class with the highest probability for each test sample.

Random Forest. The Random Forest model used for classification was trained with 500 trees and 3 features randomly selected at each split.

XGBoost. For the classification task, the XGBoost model was trained with the same parameter settings as in the regression task, including a maximum tree depth of 6, a learning rate (η) of 0.3, and 100 boosting rounds. The objective function was set to multi-class classification (softmax) to handle the multi-class journal prediction task.

SVM. Support Vector Machine (SVM) (17) is a supervised learning algorithm that finds an optimal hyperplane to separate different classes in a high-dimensional feature space. In this study, a radial basis function (RBF) kernel was employed to allow for non-linear decision boundaries, enhancing the model's ability to capture complex relationships in the data.

Neural Network. A neural network (NN) (18) model is a supervised learning algorithm that models complex relationships between input features and target labels through interconnected layers of neurones. The network architecture in this work consisted of two hidden layers, with 10 neurones in the first layer and 5 neurones in the second layer. The activation function used was sigmoid, ensuring non-linear transformations of the input features.

Evaluations. The RF, SVM, and XGBoost models have the highest accuracy and precision scores, both exceeding 95%, compared to the other three models as shown in Figure 12 and Figure 11. While SVM demonstrated the highest precision across all three models, it showed poorer recall performance for minority samples in the *Health Systems* and *Journal of Simulation categories*.

References. References should be cited in numerical order as they appear in text; this will be done automatically via bibtex, e.g. (**?**) and (**? ? ? ?**). All references cited in the main text should be included in the main manuscript file.

Acknowledgments

Please include your acknowledgments here, set in a single paragraph. Please do not include any acknowledgments in the Supporting Information, or anywhere else in the manuscript.

1 MA Hanson, PG Barreiro, P Crosetto, D Brockington, The strain on scientific publishing. *Quant. Sci. Stud.* **5**, 823–843 (2024).
2 K Kousha, M Thelwall, Factors associating with or predicting more cited or higher quality journal articles: An annual review of information science and technology (arist) paper. *J. Assoc. for Inf. Sci. Technol.* **75**, 215–244 (2024).
3 WA Qader, MM Ameen, BI Ahmed, An overview of bag of words;importance, implementation, applications, and challenges in *2019 International Engineering Conference (IEC)*. pp. 200–204 (2019).
4 P Kherwa, P Bansal, Topic modeling: A comprehensive review. *EAI Endorsed Transactions on Scalable Inf. Syst.* **7** (2019).
5 ME Roberts, BM Stewart, D Tingley, stm: An R package for structural topic models. *J. Stat. Softw.* **91**, 1–40 (2019).
6 J Han, M Kamber, J Pei, 2 - getting to know your data in *Data Mining (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems, eds. J Han, M Kamber, J Pei. (Morgan Kaufmann, Boston), Third edition edition, pp. 39–82 (2012).
7 R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria), (2024).
8 KP F.R.S., Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, Dublin Philos. Mag. J. Sci.* **2**, 559–572 (1901).
9 T Chen, C Guestrin, Xgboost: A scalable tree boosting system in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. (Association for Computing Machinery, New York, NY, USA), p. 785–794 (2016).
10 L Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001).
11 M Maechler, P Rousseeuw, A Struyf, M Hubert, K Hornik, *cluster: Cluster Analysis Basics and Extensions*, (2024) R package version 2.1.8 — For new features, see the 'NEWS' and the 'Changelog' file in the package source).
12 RJGB Campello, D Moulavi, A Zimek, J Sander, Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Trans. Knowl. Discov. Data* **10** (2015).
13 X Jin, J Han, *K-Means Clustering*, eds. C Sammut, GI Webb. (Springer US, Boston, MA), pp. 563–564 (2010).
14 C Sammut, GI Webb, eds., *TF–IDF*. (Springer US, Boston, MA), pp. 986–987 (2010).
15 A Mucherino, PJ Papajorgji, PM Pardalos, *k-Nearest Neighbor Classification*. (Springer New York, New York, NY), pp. 83–106 (2009).
16 , *Multinomial Logistic Regression*. (John Wiley Sons, Ltd), pp. 109–124 (2016).
17 M Hearst, S Dumais, E Osuna, J Platt, B Scholkopf, Support vector machines. *IEEE Intell. Syst. their Appl.* **13**, 18–28 (1998).
18 WS McCulloch, W Pitts, A logical calculus of the ideas immanent in nervous activity. *The bulletin mathematical biophysics* **5**, 115–133 (1943).

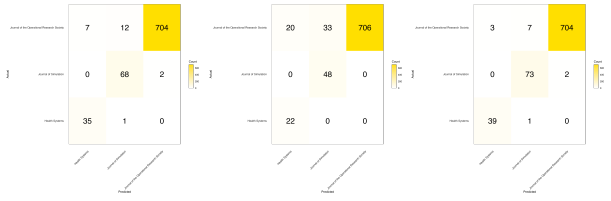

Fig. 11. Confusion matrices for RF, SVM, and XGBoost models.

by

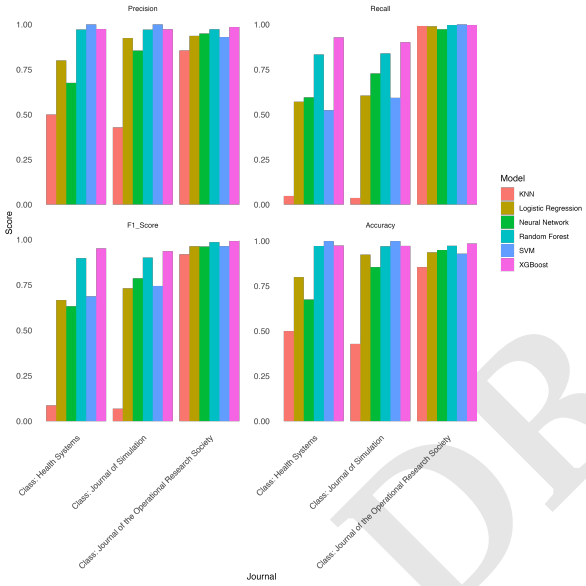

Fig. 12. Metrics for classification evaluation.