Building Predictive Models of Newsworthiness to Support Science Journalism

Sachita Nishal and Nick Diakopoulos
Northwestern University
Corresponding author: nishal@u.northwestern.edu

Introduction

When science journalists begin hunting for news stories, they explore the cutting-edge of scientific work across books, journals, at conference presentations, on preprint servers, etc. However, the scale of scientific publishing continues to grow exponentially today, creating an information overload on readers, since cognitive capacity and time availability remain constant [1].

In this context, our work aims to science journalists effectively identify **newsworthy** scientific research that may warrant development into news articles. This approach draws from **computational news discovery** [2] and leverages crowdsourcing and machine-learning to predict the newsworthiness of individual scientific articles.

We build a predictive model that uses textual and metadata features to suggest promising leads from the thousands of articles published on the arXiv preprint server every month. We validate these recommendations with expert journalists, to provide insight into the model's predictions.

Modelling "Newsworthiness"

News Values.
Journalists' decisions of what material is "newsworthy" are affected by several cultural, organizational, and sociological factors relating to the material itself, which are known as **news values** [3], e.g. controversy, surprise, impact, etc. We surveyed science journalists and thematically coded their responses to understand which news values carried high priority in their work, and could feasibly be crowdsourced from layperson crowd-workers to construct the dataset for a machine-learning model.

Crowdsourcing Newsworthiness.
Based on subsequent pilots on Amazon' Mechanical Turk, we found that given a particular scientific abstract, certain news values could be (1) consistently understand and evaluated by crowd-workers based on the abstracts of science articles, and (2) were moderately correlated to journalists' evaluations of what was newsworthy vs. what was not. These news vales were: **actuality, magnitude of impact, positive vs. negative impact, and the explainability** of an article itself.

Predicting Newsworthiness

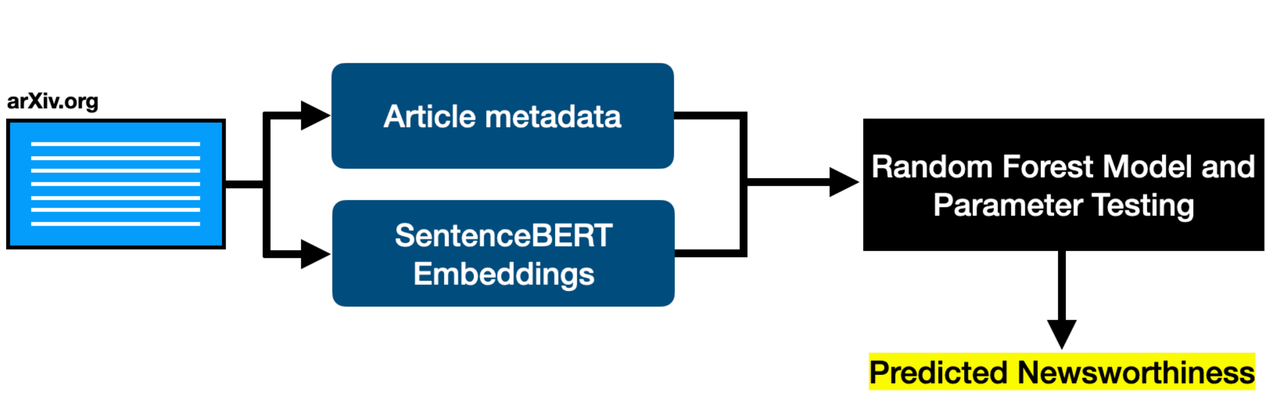

Fig 1. Data Pipeline for Predicting Newsworthiness

- To build a predictive model for newsworthiness, we sampled 500 arXiv articles from various sub-domains in Computer Science from Jan-Feb 2021, and Aug-Sep 2021.
- For these articles, we also collected the full text, as well different metadata features, such as their readability scores, their code availability, their crowdsourced news value ratings, etc.
- We then trained a Random Forest model to predict **newsworthiness scores**, which were derived by summing individual news value ratings.

- For a small validation dataset, we ranked articles by the model's predicted newsworthiness, and also by crowdsourced newsworthiness.
- We then calculated a **precision@K metric** for predicted and crowdsourced ratings, which tells us for the top-K ranked articles, the % that are newsworthy as per the ratings of journalists (ground-truth).
- The model performed visibly better than crowdsourced ratings for this criterion, with **80% of the model's top-10 ranked articles being newsworthy** as per journalists too.

Value of K	Precision@K of Predicted Newsworthiness Ratings	Precision@K Using Crowd Ratings
10	0.80	0.60
15	0.67	0.53
20	0.50	0.50

Table 1. Validating Model Predictions with Experts

Implications and Future Work

- Despite the subjectivity of the newsworthiness prediction task, and the moderate correlations of quantitative ratings between crowd-workers and experts, the predictive model provides a reasonably precise set of ranked recommendations.
- A key limitation hinges on how we define and scope "newsworthiness" for crowdsourced ratings, keeping in mind that there are certain expert-specific criteria (e.g., marketability, nature of audiences) that crowd-workers might not be able to reason about.
- We envision a deployment of the model such that expert science journalists are still empowered as the final judges of what is newsworthy (this deployment study is under review).
- We also aim to build further explainability and transparency into the model, to improve trust, error-checking, and even educational value for end-users.

Link to the CSCW-version of this study here:

Acknowledgements + References

Support for this project came from the National Science Foundation (IIS-1845460).

[1] Tom Hope, Doug Downey, Oren Etzioni, Daniel S. Weld, and Eric Horvitz. 2022. A Computational Inflection for Scientific Discovery. arXiv:2205.02007 [cs] (May 2022). Retrieved May 11, 2022 from http://arxiv.org/abs/2205.02007

[2] Nicholas Diakopoulos. 2020. Computational News Discovery: Towards Design Considerations for Editorial Orientation Algorithms in Journalism. Digital Journalism 8, 7 (August 2020), 945–967. DOI:https://doi.org/10.1080/21670811.2020.1736946

[3] Franziska Badenschier and Holger Wormer. 2012. Issue Selection in Science Journalism: Towards a Special Theory of News Values for Science News? In The Sciences' Media Connection –Public Communication and its Repercussions, Simone Rödder, Martina Franzen and Peter Weingart (eds.). Springer Netherlands, Dordrecht, 59–85. DOI:https://doi.org/10.1007/978-94-007-2085-5_4