From Crowd Ratings to Predictive Models of Newsworthiness to Support Science Journalism

SACHITA NISHAL, Northwestern University, USA.
NICHOLAS DIAKOPOULOS, Northwestern University, USA

The scale of scientific publishing continues to grow, creating overload on science journalists who are inundated with choices for what would be most interesting, important, and newsworthy to cover in their reporting. Our work addresses this problem by considering the viability of creating a predictive model of newsworthiness of scientific articles that is trained using crowdsourced evaluations of newsworthiness. We proceed by first evaluating the potential of crowd-sourced evaluations of newsworthiness by assessing their alignment with expert ratings of newsworthiness, analyzing both quantitative correlations and qualitative rating rationale to understand limitations. We then demonstrate and evaluate a predictive model trained on these crowd ratings together with arXiv article metadata, text, and other computed features. Based on the crowdsourcing protocol we developed, we find that while crowdsourced ratings of newsworthiness often align moderately with expert ratings, there are also notable differences and divergences which limit the approach. Yet despite these limitations we also find that the predictive model we built provides a reasonably precise set of rankings when validated against expert evaluations (P@10 = 0.8, P@15 = 0.67), suggesting that a viable signal can be learned from crowdsourced evaluations of newsworthiness. Based on these findings we discuss opportunities for future work to leverage crowdsourcing and predictive approaches to support journalistic work in discovering and filtering newsworthy information.

CCS Concepts: • **Human-centered computing** → **HCI design and evaluation methods**; **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: Newsworthiness, News Values, Science Journalism, Crowdsourcing

ACM Reference Format:
Sachita Nishal and Nicholas Diakopoulos. 2022. From Crowd Ratings to Predictive Models of Newsworthiness to Support Science Journalism. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 441 (November 2022), 28 pages. https://doi.org/10.1145/3555542

1 INTRODUCTION

From its inception more than a hundred years ago as the "Gee-Whiz" reporting of new scientific findings to more modern conceptions of "telling the whole complicated story" [11] science journalism occupies an important role in society, serving both to translate and critique scientific findings that have important bearing on a range of issues, from climate change and global pandemics to the rise of artificial intelligence in social systems. Much like other domains of journalism, science journalism is confronted with the opportunities and challenges presented by a changing media ecosystem adapting to algorithmic distribution, connected social media audiences, and the proliferation of misinformation [10, 15, 17, 29]. Scientific productivity has exploded over the past 60 years

Authors' addresses: Sachita Nishal, nishal@u.northwestern.edu, Northwestern University, USA.; Nicholas Diakopoulos, nad@northwestern.edu, Northwestern University, USA.

2573-0142/2022/11-ART441 $15.00
https://doi.org/10.1145/3555542

[12], with newer publication avenues such as conferences and open-archives fueling increases in the volume of published research [47]. A recent report from the U.S. National Science Board indicates that the volume of global research output published in the Scopus database in 2018 alone was 2.6 million articles[1]. Yet despite the growing terrain science journalists must traverse, resource limitations [2, 3] threaten to constrain their ability to accurately and comprehensively cover their beat.

Our work explores how to help science journalists effectively grapple with monitoring the growing scale of scientific information available in the present environment. To do so we develop a socio-technical approach leveraging crowdsourcing and a machine-learned model for predicting newsworthiness.

This approach draws on the notion of *computational news discovery* [23] and is particularly geared towards facilitating the identification of newsworthy scientific research that may warrant development into news articles. More specifically, our objective is to build a news discovery process that can help science journalists identify potentially newsworthy research by suggesting the most promising leads from the thousands of articles published on the arXiv preprint server every month.

In order to do so, we operationalize a set of *news values* [34] that we use to capture the newsworthiness of any given abstract from arXiv. We strive to do this in a way that aligns with professional evaluations. We collect expert evaluations of newsworthiness from professional science journalists, for scientific articles on arXiv, and then conduct a thematic analysis to identify specific criteria that experts utilize that could be feasibly crowdsourced from laypersons. The goal is to speak to the manifold ways in which experts may assess newsworthiness, with an eye towards elaborating the nuances and potential limitations associated with our method of crowd-sourcing and modeling newsworthiness. We then collect, compare and contrast crowdsourced ratings with the expert evaluations to understand the ways in which crowds and experts align (and mis-align) in their evaluations of newsworthiness, and ultimately use the crowd ratings on a sample of articles to build a predictive model for newsworthiness. The results are validated against the expert ratings, and qualitative expert responses are analyzed to provide insight into the model's predictions. Two key research questions underlie the development and evaluation of our approach:

> **RQ1**: To what extent do crowd-worker and domain expert ratings of newsworthiness of scientific articles align?
>
> **RQ2**: To what extent can a model trained on crowdsourced ratings of newsworthiness of scientific articles predict domain expert ratings of newsworthiness?

With respect to the first question, we contribute findings indicating that crowdsourced ratings of newsworthiness often align with expert opinions on newsworthiness along the axes of certain empirically established news values in the literature of science journalism [8]. In response to our second question, we find that, despite some discrepancies between the crowd and expert criteria for evaluating newsworthiness, a predictive model trained on the research article metadata and its corresponding crowdsourced ratings is capable of ranking newsworthy articles with high precision, when validated by expert ratings. Together these findings suggest that there is a valid signal of newsworthiness in lay-person ratings of articles, and suggest an opportunity for further development of predictive models that capitalize on the collective intelligence of crowd workers for lead discovery in science journalism by helping to filter the information space down to more potentially newsworthy articles.

[1] https://ncses.nsf.gov/pubs/nsb20206/publication-output-by-region-country-or-economy

2 RELATED WORK

The conception and design of this research project builds on and contributes to related work in (1) the domain of computational journalism, within which we speak to the automation of news discovery, and the tools which support it for science journalists; and (2) the evaluation of newsworthiness of new information, via news values that represent potentially interesting facets of that information.

2.1 Computational News Discovery for Science Journalism

Computational tools and approaches are prevalent at various stages of the news production pipeline, ranging from knowledge discovery [24, 67] and fact-checking [32], to automated news generation [38], the creation of informative news visualizations [35], and news dissemination [50]. The research presented here focuses on the task of news discovery and contributes to the broader domain of computational journalism, which is primarily aimed at "using algorithms to transform information and data for journalistic purposes" [23].

We focus on Reich's conception of "news discovery" as the reporter's "first contact with the first source" [61] where they first uncover a potentially newsworthy story, which is distinct from the broader "news gathering" process of data collection and information verification, which "supplies the building blocks of the news item". Our goal is to help reporters in the news discovery phase, as they wrangle with the question of "How do I become aware as quickly (and as exclusively) as possible of a potential new item in order to start the news process?".

This process incurs substantial time and material costs for journalists [54, 68]. "Commercial news criteria" i.e. factors extraneous to a potentially newsworthy event (e.g. costs such as the budget allocations, staff shortages, time requirements) can often play a significant role in the selection and the shaping of the stories that ultimately become news [3]. Such costs can often have a drastic impact on the amount and the kinds of stories that receive coverage in the mainstream media [40]. In their study of the mainstream media's coverage of scientific research from prestigious journals such as *Nature*, *Science*, *NEJM*, and *JAMA*, Suleski and Ibaraki [70] conclude that:

> Overwhelmingly, scientific research is not making it beyond the borders of the scientific community, and an increasing amount is failing to gain attention from researchers outside the specialized fields. Though scientific output continues to rise, its appearance in news media is less than 0.013% of total articles published, a mere 66 unique papers appearing in Time and on NBC News out of the 508,795 papers published in 1990. (p. 122)

They ultimately traced the broader issue of under-reportage of scientific research to: (1) the extremely high volume of research output that science journalists have to cover, and (2) a lack of responsibility for the communication of interesting results to lay audiences within the modern scientific community. In this work we focus on the first of these issues.

In recent years the science media ecosystem has undergone considerable upheaval. With the advent of an online environment where scientists, advocates, and laypersons all contribute to news production with their varied expertise and avenues (e.g., blogs, social media, etc.), science journalists have taken on a variety of new roles, such as those of the curator, the civic educator, the public intellectual, the watchdog, etc. [30] adding to the intensification of their jobs. And, while research blogging and engaged scholarship [44] have emerged to help facilitate the communication of research to broader audiences, the advent of preprint servers has contributed to the skyrocketing volume of published scientific research. For instance, the arXiv preprint server that we consider in this paper receives thousands of submissions every month: in fact, it received more than 15,000 submissions for the month of June 2021 alone. Articles on preprint servers and at many conferences

are also not typically published with press releases, the presence and the framing of which have been found to be important predictors of news media coverage for research [51, 70]. Not only is the volume of information overwhelmingly high, but in a day and age where all of it is simply a click away for news audiences, time is of great essence to science journalists, as highlighted by Dunwoody [29]: "Building science news stories for Internet consumption presents many challenges, among them the need for constant updating, managing the speed with which information must be turned into narrative and maximising the brevity of those narratives, so critical to audiences with only seconds to spare."

In this work we explore computational approaches to news discovery applied in the domain of science journalism, with an eye towards reducing the cost of monitoring the growing array of scientific preprint articles to discover potentially newsworthy leads. We define this computational news discovery as "the use of algorithms to orient editorial attention to potentially newsworthy events or information prior to publication." [23]. The idea of computational news discovery has precedent in some of the earliest conceptions that considered the role of computational tools in journalism and hypothesized about automated monitoring systems that could detect and alert journalists to the occurrence of anomalous events [41]. Today, this vision is embodied in the various tools that have been developed in the news industry to monitor and detect anomalous and newsworthy information from live social media feeds, as well as from official textual documents from organizations and institutions.

The CityBeat system, for instance, was developed and tested in collaboration with New York City newsrooms to find potentially newsworthy, real-time events using Instagram data, and automatically assess the accuracy and the public interest for the detected information [67]. The Reuters Tracer tool [49] provided a more elaborate and comprehensive version of such a detection system, by entirely automating the monitoring of Twitter feeds for potentially newsworthy events; the filtration and clustering of events; the contextualization, summarization, and newsworthiness assessment of each story; and ultimately the dissemination of selected news events. In addition to gathering information from social media, computational news discovery tools are also designed to ingest data from institutional data sources, and suggest promising avenues for journalistic research and reporting [27, 69]. This automated form of watchdog journalism is exemplified by the Marple system that monitors criminal offences in Sweden across all municipalities, and detects anomalous events that could lead to interesting news stories [52]. Another example of such a system is the Lead Locator developed at the Washington Post, which assists political reporters by analyzing national voter data in the US to rank counties in terms of how interesting their voting patterns could be for journalistic reporting [25].

Our work adds to this previous research that monitors and evaluates publicly available documents for newsworthiness, while distinguishing itself by focusing specifically on the science journalism beat. Similar to the computational news discovery systems highlighted above, the process we developed aims to help journalists surface potential leads in a high-velocity, high-volume information landscape. But here we focus on monitoring scientific information — the arXiv preprint server in particular — by computationally operationalizing news values applicable to the science journalism domain, which we expand on next.

2.2 From Crowdsourced to Predicted Newsworthiness

When asked to justify their decisions regarding the newsworthiness of certain events and occurrences, journalists have often been known to offer their "gut feeling" as an explanation [16, 66]. However, academics studying the process of news selection by journalists have posited the existence of a specific set of cultural, organizational, and sociological factors, termed *news values*, that affect journalistic decision-making and manifest within published news stories. First put forth by Galtung

and Ruge in their landmark analysis of foreign news coverage in Norwegian newspapers [34], these news values have since been identified and updated via further scholarly analyses of published news stories [37, 42] and ethnographic interviews with journalists [28, 66]. Some examples of these news values include controversy, surprise/unexpectedness, actuality, magnitude, reference to the power elite, good news, bad news, continuity etc. [8, 42, 43]. The relative importance of individual news values to journalists and news organizations varies due to the influence of several contextual, sociological, and other chance-based factors, and it is not strictly necessary for all possible news values to be present in a potential story for it to be newsworthy. Ultimately, the news values simply serve to inform a highly domain-specific and diverse editorial process that journalists undertake [43].

Existing systems for computational news discovery have operationalized varying news values, in pursuit of their own diverse ends. The CityBeat system, for instance, assesses the "deviation" of real-time Instagram activity for a given region in New York City, from past time-series data, in order to discover candidate news stories [67], and provides hourly statistics to optimize for the quick discovery of recent events. The Reuters Tracer system formalizes the broader dimensions of *novelty*, *scope/impact*, and *localization* for each potential news story that it detects and contextualizes from social media [49]. The Lead Locator system operationalizes *novelty*, *political relevance*, and *magnitude* in its specialized pipeline for political reporting [25]. The selection of relevant news values to operationalize newsworthiness is foundational to our endeavor. In contrast to these prior efforts, in this work we develop and evaluate crowdsourced and computational operationalizations of news values as they are applied specifically to the domain of science journalism [8].

The prediction of newsworthiness in our research is supported by crowdsourcing ratings of individual research articles along the axes of the selected news values. Crowdsourcing techniques have been successfully applied in journalism at various stages, including lead-discovery [1, 67], article-writing [46], and even news verification and fact-checking [73]. In the context of science journalism, The Guardian experimented with the StoryTracker system, where they published a piece of science journalism, and then invited readers, scientists, and other concerned citizens to submit follow-up analyses and reactions - e.g. subsequent press coverage, reported retractions, reactions from the wider scientific community, blog posts etc. - which would keep getting added to the story to flesh out a more intricate picture of the broader impact it had had since the publication of the news story [45]. In the domain of administrative document news discovery, crowdsourced ratings have been collected for news values including *negative impact*, *magnitude of impact*, *controversy*, and *surprise*, and these ratings were then evaluated with expert journalists to assess their receptivity to this information [27].

The current work builds on such prior efforts by also directly crowdsourcing ratings (and qualitative rationale) for news values that we believe are both relevant to science journalism, and also plausible to crowdsource based on our analyses of expert judgements, including *actuality*, *unexpectedness / surprise*, *impact* (magnitude and valence), and *controversy*. These values and their operationalization are detailed further in Section 4. We further advance the prior work by developing a computational model that is trained on the crowdsourced evaluations of news values, validating the crowd ratings and model with expert journalists' evaluations of newsworthiness, and then predicting the newsworthiness of previously unseen research articles using their text and metadata, demonstrated in Section 5.

3 DATA

In order to build a model to predict newsworthiness ratings, we collect and preprocess our training and validation datasets using the open-access arXiv API[2]. For the purposes of this study, we focus on papers from the field of Computer Science, as classified in the arXiv Category Taxonomy. This is a design decision we take due to the heightened relevance of computational science across scientific disciplines, but we recognize that this may also limit the scope of our findings. In future work, the procedure of data collection and preprocessing specified below could be extended to other categories of journalistic interest, both on arXiv and on the myriad other domain-specific preprint servers available online.

Using the API we created two independent datasets: (1) a validation dataset used to hone the crowd-sourcing methodology, compare crowd-sourced to expert ratings, and validate our predictive model, and (2) a training dataset used to develop our predictive model. To create our validation dataset, we considered the arXiv papers published in the month of November, 2020, and collected newsworthiness ratings for these papers from both Amazon Mechanical Turk (AMT) workers and our recruited domain experts (science journalists). Over the course of a few months, we piloted several iterations of surveys on the AMT platform in order to refine our crowd-sourcing survey, and once we decided on a final survey design that we believed was well-suited to the task at hand, we proceeded with the collection and rating of our training dataset. Since *actuality* i.e. contemporary relevance, is a news value we aim to measure, it was vital to gather ratings for a recent set of research papers at time of crowd-sourcing. Our training dataset was collected in two separate phases as we built and expanded our model, and included arXiv papers published in the months of January-February, 2021 and August-September, 2021.

In creating these datasets we filtered down to eleven arXiv categories within Computer Science, because we believe these are more likely to contain papers with the potential for newsworthiness. In our assessment, these categories tended to contain research papers with higher practical applicability and were sometimes associated with general public interest topics (e.g. computer vision for facial recognition). These categories include: cs.AI (Artificial Intelligence), cs.CL (Computation and Language), cs.CV (Computer Vision and Pattern Recognition), cs.CY (Computers and Society), cs.HC (Human-Computer Interaction), cs.IR (Information Retrieval), cs.LG (Machine Learning), cs.MM (Multimedia), cs.NI (Networking and Internet Architecture), cs.RO (Robotics), and cs.SI (Social and Information Networks). By retaining papers that contain only these categories in their list of author-assigned categories, we aimed to filter out highly technical or theoretical research that might have limited practical applicability or ease of comprehension (by journalists and crowd-workers alike). We recognize that this filtering process is also accompanied by a reduction in the scope of applicability of our work, a limitation we elaborate upon in Section 6.

Each research paper uploaded to arXiv is published under a "Primary Category", which is identified by the authors of the publication as one category out of all the ones assigned to the paper that they consider most relevant. The four most popular primary categories in our training dataset encompass nearly three quarters of the papers: cs.CV (29.2%), cs.CL (21.0%), cs.LG (18.6%), , and cs.RO (9.2%). These are also the four most popular categories in our validation dataset, and with similar proportions.

Such a heavily skewed frequency distribution over primary categories in both samples could potentially correspond to the amount of research activity in those respective fields of study, but it may not necessarily reflect the potential for newsworthiness in those fields. As a result, we decided to select an evenly balanced sample of research papers from each of the eleven potential primary categories to train and validate our model. We randomly sampled 5 papers from each

[2]See: https://arxiv.org/help/api/basics

Table 1. Evolution of the Training and Validation Samples Through the Filtering Process

Sample Type	Total Uploads to arXiv	Total Uploads to arXiv in Computer Science	Total Papers after Category-based Filtering	Total Papers After Random Sampling
Training Set (Jan-Feb. & Aug-Sep. 2021)	55731	20000	4157	500
Validation Set (Nov. 2020)	15130	5330	1559	55

primary category to create our validation set, and collected domain expert and crowd-worker newsworthiness ratings for these papers. To create a training dataset, we collected a larger sample of 50 research papers from each primary category, and discovered that one of the selected categories - cs.MM (Multimedia) - had an insufficient number of papers uploaded to it in our specified timeframe. As a result, we did not include any papers from this category in our training dataset. This filtering and sampling process drastically reduces the size of both our initial dataset, and Table 1 reflects these changes in sample size, as well as the final number of papers in the training and validation sets.

The arXiv API provides detailed metadata for each paper, of which we collect: arXiv ID, title, abstract, author list, publication date, primary arXiv category, and other arXiv categories (all categories being within Computer Science, and specified by the authors). We supplement this arXiv metadata by adding a *readability* score for the abstract text of each research article. This score is calculated using the publicly available implementation of the science De-Jargonizer, that is based on a corpus of 90 million words published on the BBC news website during the years 2012–2015 [60]. The De-Jargonizer *readability* helps us assess the interpretability of a paper with a generalized method that is easily scalable.

4 RQ1: EXPERT AND CROWDSOURCED RATINGS OF NEWSWORTHINESS

This section describes the mixed-methods approach we adopted to address RQ1, where we investigate the extent to which crowd-worker and domain expert ratings of the newsworthiness of scientific articles are aligned. After collecting and augmenting abstracts for our training and validation datasets, we collected Likert ratings for newsworthiness and corresponding free-text rationale, from crowd-workers and experts, for all the papers in our validation set. This was an iteratively piloted data collection process. Section 4.1 dives into the specific details of these survey methods for domain experts, as well as into the results of the qualitative thematic analysis we conducted on their responses. In Section 4.2, we use these results to help justify the specific aspects of newsworthiness that might be plausibly addressed via crowdsourcing, and in Section 4.3, we present the resulting survey design for crowd-workers. In Section 4.4, we provide a quantitative description of the crowd-ratings, along with their relationship to expert ratings, which motivated us to scale-up the endeavor and build a model to predict newsworthiness as described in Section 5.

4.1 Expert Ratings of Newsworthiness

We constructed our validation set by collecting assessments of newsworthiness from two domain-experts, i.e. science journalists. These professional journalists were recruited through personal networks and were brought on as consultants because of their experience and expertise in the

field of journalism, specifically writing about science and technology. Given their high level of expertise we paid these consultants $50 / hour. Section 4.1.1 describes the design of the surveys our experts completed, and Section 4.1.2 discusses our qualitative analysis of the results, wherein we infer latent themes that underlie expert judgments of newsworthiness.

4.1.1 Survey Design for Expert Newsworthiness Ratings. To set up the rating task for the experts, we provided them with the title, the abstract, and the arXiv URL of each of the 55 papers in our validation sample. We believed that this basic information would be the most relevant to their assessment of newsworthiness, but we also provided a link to the original paper to provide any additional context, information, or clarification that might be necessary. We then asked our experts to provide a single rating for each paper, which reflected the extent to which they agreed with the statement: "The research described could be interesting for journalists to either report on directly, or to develop into a news article for a science or tech-focused publication". To ensure that journalistic ratings of newsworthiness could derive from a wide set of criteria, this question was as broadly scoped as possible, without prompting for explicit news values. The rating for this question was collected along a five-point Likert scale from 1 (Strongly Disagree) to 5 (Strongly Agree), with a higher rating being indicative of greater potential *newsworthiness*. We also asked our experts to provide a 2-3 sentence justification for this rating, that pointed to specific aspects of the research that they found interesting (or uninteresting). We computed the correlation between the two experts to understand inter-rater reliability, and analyzed the experts' written explanations to better understand the criteria they were applying in their newsworthiness judgements, and to help justify how the crowdsourcing survey was structured to prompt for specific criteria.

4.1.2 Thematic Analysis of Expert Rationale. In this section, we describe the criteria that our experts employ to judge an article's newsworthiness, which emerge from our qualitative thematic analysis [13] of their respective free-text rationale. The themes were coded by the first author, and discussed and refined with the second author. This process was conducted inductively, where we sequentially examined each rationale to see if it fit into any of the existing themes that had been coded thus far, or if it could constitute a new theme in the data. We also maintained a set of memos that served to expound on the themes, establish connections between them, and make note of illustrative examples. Once a set of potential themes had been identified, we conducted a round of axial coding to integrate the codes and establish a broader thematic 'map' of the criteria that undergird expert judgements of newsworthiness.

For each theme, we present any sub-themes under it, and illustrate them with examples. Three main themes emerge from our analysis: Research Characteristics, Story Actualization, and Story Reception.

Research Characteristics. Domain experts strongly rely on certain attributes of the research article itself - which could be a function of its field, premise, methods, impacts, etc. - to judge newsworthiness. The criteria in this section often correspond closely to news values that have been previously established in the literature discussed in Section 2.2. All of these criteria involve reasoning around the information provided in the article's abstract, but domain experts also occasionally contextualize this information within the broader scientific field, and this affects their ratings. We point out these cases as and where they occur.

Actuality. We observe that our experts weigh the actuality (i.e. relevance to contemporary issues) of research articles quite highly in their explanations, often recognizing topics such as COVID-19, privacy, or hate speech, to be of current interest to their audiences. For instance, for one article on session-based recommendation systems, an expert comments: "The study tackles an interesting

and timely problem—how to configure user recommendations under much more stringent data privacy restrictions."

Range. The experts are frequently mindful of the "number of people participating in an event or affected by the event" [8], especially while discussing the applications of technological developments. For instance, an article on machine translation elicits the following response: "Interesting in that it has obvious implications for expanding machine translation for low resource languages, but the authors seem to acknowledge their method has limited utility. Would be hard to get readers jazzed about that."

Unexpectedness/Surprise. This criterion is another one we find a direct parallel for in the literature of traditional news values, although it manifests in two different ways. One aspect of surprise is rooted in the amazement an article's unexpected or futuristic imagery might generate. For instance, an expert responded to an article on cross-lingual word embeddings: "This is a cool result, I like thinking of the potential sci-fi application of it being able to possibly help interpret alien languages." The second aspect of surprise is rooted in the scientific practices described in an article, which may be novel innovations, unexpected combinations, or innovative applications. For instance, one journalist commented on the use of VR for a tutoring system: "There is nothing particularly unique about using VR for anatomy education, especially when the area of study (the base of the skull) is so limited". We note that such comments pointing to the scientific decisions, the research design, or the status quo in the scientific field, require some degree of expert knowledge about the field.

Societal Impacts. The extrapolation of potential societal impacts is a prominent part of expert rationales. This occurs by way of two mechanisms: by making assumptions about the *timeframe* of societal impacts, and by identifying the broader *types* of societal impacts. Assumptions about the *timeframe* of social impacts specifically constitute the speculative aspect of the news value that has traditionally been termed as *timeliness* and is exemplified by any event "that happened the day of or day before publication or ... that is due to happen in the immediate future" [64]. The experts often explicitly state these assumptions, and factor in the time it typically takes for academic innovation to translate to widespread adoption, when making their judgements. For instance, one journalist comments on an article about a novel city exploration interface as follows: "The system sounds interesting, but it'll be more interesting if it's ever actually built and deployed."

The second aspect involves the *types* of social impacts of technologies. Discussions abound for topics such as privacy considerations, the ethics of algorithms, the economy, job-automation, unemployment, disaster management, and align well with a recent thematic analysis of broader impact statements about ML and AI research [57]. The experts also recognize and comment upon the *valence* of these impacts, as seen in this comment for an article about automated team formation in the workplace: "While the concept is an interesting and relatable one, it almost feels like one of those "creepy" applications of AI. I think coverage of this could be negative."

Scientific Impacts. Due to their domain expertise, science journalists are often in a suitable position to comment upon the scientific relevance of certain articles i.e. "the importance of an event for the scientific progress" [8]. Our expert journalists mainly describe two aspects of this impact on the scientific community: the *size* of the impact, and the *location* of the impact in the research pipeline. The size points to how incremental its progress is toward the scientific question it seeks to answer, whereas the location of impact points to whether the article entails any notable improvements to the methods, metrics, performance, etc.

We see the reasoning around the size of impact in an expert's comment on an article about probabilistic graph models: "While it's interesting to note that Alibaba researchers are advancing this area of basic AI research, the study on its own is too incremental ..."

Story Actualization. The process of developing a potential lead into an actual news story involves certain logistical and editorial considerations, which also factor into our experts' assessment of whether a lead may be worth pursuing in the first place. The criteria in this section describe three such important considerations that are part of the process of the story's realization.

Explainability of Article. The extent to which an article is easy to explain and communicate to the audience directly affects the time a journalist spends on it, which is an important constraint in the literature surrounding commercial news values [3]. Expert rationale in the validation set do take cognizance of this constraint, and often trace an article's ease of explainability to its scope i.e. its domain or technical specificity, as can be seen in this expert's rating for an article on self-imitation learning: "Focuses on the simple concept of an AI "learning from its errors", can almost be humanized–we learn from our errors as well!–seems not too difficult to explain"

Jumping-Off Point. The distinction between the "news discovery" and the "news gathering" processes for journalists [61] is demonstrated by experts' views of some research articles as jumping-off points for larger, more general stories. For instance, a survey article on deep learning for recommender systems elicits the following response: "In the way that the authors intend this study to be used—as a cookbook for exploiting user data to generate point-of-interest recommendations—I don't think it would be terribly interesting to a general audience. But a survey of a whole new field of invasive uses of user data could be a good jumping off point for a story about algorithms clawing ever deeper and more precisely into our digital lives."

Framing and Format Possibilities. Journalists often rely on a set of news angles, defined as "conceptual criteria that are used both to assess whether something is newsworthy and also to shape the structure of the resulting news item" [56] in their development of a news story for publication. Factors surrounding the framing of a news story to set up certain news angles, or even general story formats (eg. short piece vs. deep dive), are especially visible in expert rationale. An expert directly cites the *human-interest* news angle and articulates a potential thematic framing in their response for an article on latent embeddings for videos: "Very relatable and clear application, has the added human interest of 'movies' and cinema, IMDB."

Story Reception. The final step of the news pipeline involves publishing and marketing a story to an organization's audience, and we observe themes corresponding to these concerns in the expert rationale as well. These themes broadly correspond to the set of commercial news values [3] and the concept of share-worthiness of research on social media [72] - factors which require expertise in journalism for their assessments.

Potential Audience/Publication. A frequent theme we observe pertains to the type of audience that might enjoy and understand a story. From the responses of our experts, we gather that this could be a highly general audience (e.g. Popular Mechanics), a mathematics focused audience (e.g. Quanta), a tech focused audience (e.g. MIT Technology Review), or even an industry focused audience (e.g. TechCrunch). The suitability of a story to a certain audience is related to its Research Characteristics (e.g. a highly technical and specific story would likely be unsuitable for Popular Mechanics), but the likelihood of a story itself being deemed newsworthy may also take into account the popularity and the breadth of its best-suited publication.

Marketability. We note that our experts also factor in the click-worthiness and the social-media share-worthiness of a news story. They point to the potential for snappy headlines and quick clicks, and while the absence or the presence of these criteria is not critical to their judgment on the newsworthiness, discussions of marketability do work in tandem with some other criteria discussed before. An expert, for example, simultaneously examined the surprise factor, the social impacts, and the marketability in a potential story on using reinforcement learning for human detection and tracking: "I find this rather dystopian, but that doesn't hurt its chances of being a story. There

are two obvious, buzzy applications that come to mind: Amazon style cashier-less retail, and state surveillance. You could easily see a headline about facial recognition that you can't hide from."

4.2 Identifying the Potential for Crowdsourcing

In this work we pursue an approach where we cue crowd-workers with specific questions to engage them on individual newsworthiness criteria with potential for crowdsourcing. We then combine these responses into an aggregate newsworthiness construct. While this approach implies that the aggregate newsworthiness construct will reflect only a subset of the construct as measured directly from professionals (since they consider a wider array of criteria), the approach allows us to provide more explicit guidance to better direct crowd responses, which is an important consideration in crowdsourcing protocols [19, 74]. In this section we elaborate our rationale for the selection of each criterion for crowdsourcing, which often hinge upon the amount of domain expertise necessary to make plausible assessments about it. The final set of dimensions of newsworthiness (i.e. news values) we chose to crowdsource is presented in Table 2 with their definitions or approximations from the literature.

The Research Characteristics set of criteria broadly map to the traditional literature on news values [8, 43]. We observe that the expert assessments for most of them rely either partially, or almost entirely, on the information provided in the abstract, which is explicitly designed to communicate the relevance, impacts, novelty, applicability etc. of the research article in question [5, 58]. However, there are some potential caveats. While laypersons might be able to imagine the futuristic uses of technology that lead to the amazement-related aspect of *surprise*, their unfamiliarity with recent technical developments in the relevant field would make it harder for them to infer the element of *surprise* associated with the scientific novelty of a given research article. Even the timeframe of *societal impacts* requires expert knowledge about practical considerations in the translation of science to industry. We believe that crowdworkers can thus reason about *surprise* and *societal impacts* of the research, but potentially to a limited extent.

Based on this analysis of the experts' rationale we arrive at four major news values that we argue can be meaningfully crowdsourced based on the abstract of the research article: *actuality*, *surprise*, *impact magnitude*, and *impact valence*. The *impact* dimension in particular was mapped to two aspects - *magnitude* and *valence* - which are respectively intended to encompass the *range* criterion, as well as the valence element of *societal impacts*. In addition to these news values, we also decided to collect ratings for *controversy* from our crowd-workers. Though we did not identify it at a coarser level in our thematic analyses, this news value has been observed to be important to multiple stakeholders - be they journalists, news editors, or scientists - in empirical studies in science journalism [8]. It is plausible to measure crowd perceptions for *controversy* [27], and we believe that it could potentially further illuminate certain aspects of an article's *societal impacts*. On the other hand, we argue that *scientific impacts* from our thematic analyses would be extremely difficult for crowd-workers to evaluate, since they are unlikely to have a macroscopic view of the scientific field that such an evaluation demands. Consequently, we did not attempt to crowdsource this dimension of impact in this study.

We also acknowledge that most of the criteria that are part of Story Actualization and Story Reception would be substantially harder for crowd-workers to consider, given their likely lack of expertise in researching, writing, and publishing material in a newsroom. For instance, knowing about specific audience interests, framing or format options, and whether something could still be a reasonable jumping off point all speak to particular professional activities that we do not expect crowds to be able to rate. One exception in this regard is the explainability of the article. While crowd-workers cannot directly estimate the labor required to translate jargon-heavy articles for public consumption, a possible proxy for this could simply be how well the crowd-workers

Table 2. Operationalized News Values and Their Definitions

News Value	Definition
Actuality	Relevance of research to the present moment "coming from the general news situation, or the research operation, or both" [8]
Surprise	"Stories that have an element of surprise, contrast and/or the unusual about them" [43]
Impact Magnitude	"Stories perceived as sufficiently significant in the large numbers of people involved or in potential impact, or involving a degree of extreme behaviour or extreme occurrence." [43]
Impact Valence	The positive or negative nature of the impact of the research on society, such as to individuals, organizations, politics, or the economy (A combination of the traditional news values "Good News" and "Bad News" [43])
Controversy	"Contrasting of differences in opinions", as a result of the research [8]
Understandability	Ease of understanding the abstract presented, intended to approximate practical considerations of explainability [3]

understand the abstract presented to them: lower average ratings of an article's *understandability* might indicate that there are more barriers to overcome for a journalist who seeks to simplify this content for their audience. We therefore decided to crowdsource *understandability* for our validation sample.

4.3 Crowd Ratings of Newsworthiness

4.3.1 Survey Design for Crowd Newsworthiness Ratings. Using AMT, we next collected crowd-worker ratings for each of the 55 research abstracts in our validation sample. A single human intelligence task (HIT) in our case consisted of an individual research abstract and its corresponding survey. We collected ratings from multiple unique crowd-workers for each HIT. Restrictions were also set up to limit HIT completions to workers in the United States (to ensure knowledge of cultural context, which we thought relevant to the news value of *actuality*), who had completed at least 500 HITs with an overall acceptance rate of 98%. In order to ensure our survey was effective at capturing crowd judgements, we incorporated findings and best-practices from the literature on survey and crowdsourcing task design [1, 20, 53, 71, 74] and ran a series of pilots implementing these practices and assessing their effectiveness. Based on the median time taken during early pilots and adjusted based on the length of the survey as it evolved we offered workers $1.25 per rating task, with the goal of offering at least minimum wage in the U.S. state where the research was undertaken.

Each worker was provided with the title, the abstract, and the arXiv URL for an individual research paper, as in the case of the domain experts. The arXiv URL was provided in case the workers required additional information or context to improve their answers, but its usage was specified to be optional., corresponding The response for each question in the survey was collected along a five-point Likert scale, going from 1 (Strongly Agree/Strongly Positive) to 5 (Strongly Disagree/Strongly Negative).

Each Likert question also required a qualitative response to be written in 2-3 complete sentences that were original and detailed articulations of the rationale behind the rating. We also conducted manual screenings of all qualitative responses to ensure they fit the criteria provided in the instructions [71], including being written in complete sentences, being original (i.e. not copy-pasted from the task), and being specific and detailed in the explanation of the rating. If they did not satisfy these requirements, the responses were not included in our data.

Iterations on the survey led to changes in the length of instructions to balance clarity with appeal [74], and to the introduction of ex-ante attention-checks [71] to encourage high response quality. Appendix A contains the final version of the complete survey provided to AMT workers.

For each research paper, we collected ratings from multiple workers to include as diverse an array of opinions as possible, as well as to average out the effect of individual worker subjectivity. In an initial pilot of the survey, we collected responses from five crowd-workers per research article. However, in an effort to reduce the cost of collecting and processing these ratings and their textual explanations, we compared them to the results from collecting three crowd-worker ratings per article instead. Based on the similarity of the mean ratings in both pilots, across individual dimensions of newsworthiness, we decided to collect ratings from three crowd-workers per article, and hence reduce resource cost. However, future work could also explore the effects of a larger amount of crowd-worker ratings on the analyses described in our study, in the absence of resource constraints.

4.4 Comparative Analyses of Expert and Crowd Ratings

After collecting survey responses on our validation set, we conducted a set of statistical tests to assess whether the crowd-worker ratings provided a valid signal that aligns with how experts view newsworthiness. To this end, we drew from empirical studies that have demonstrated the feasibility of using aggregated ratings from experts and non-experts to compare agreement across these two groups on content analysis tasks [27, 48]. We did not compute inter-rater reliability within the crowd-workers because the ratings task focuses on the subjective interpretation of latent (i.e. non-manifest) constructs [65] with the end-goal of learning a predictive model where reliability coefficients are poor indicators of suitability for machine learning applications [6]. The findings from our statistical analyses help motivate the use of crowdsourced ratings for our predictive model, the evaluation of which in Section 5 provides a validation of the overall approach.

We first averaged the individual ratings provided by our domain experts to obtain an overall *expert newsworthiness* rating for each article in the data (again, the range is from 1 to 5, with higher indicating more newsworthy). The Likert responses for the news values from our crowd survey required a slightly more involved aggregation to facilitate quantitative comparisons. Hence, we constructed a *crowd newsworthiness* rating for each article, while adhering to practices in index construction that are typically employed in the social sciences [7]. To this end, we first averaged the scores across the three crowd-worker ratings for each survey question, for each research article. This culminated in six crowdsourced dimensions for each article: *actuality*, *surprise*, *impact magnitude*, *impact valence*, *controversy*, and *understandability*. We transformed the *impact valence*, converting the diverging scale from "Strongly Negative" to "Strongly Positive" in our survey to a sequential scale from 1 to 5, where a higher value indicates higher intensity of positive *or* negative valence. We next summed up the six news values to obtain the new overall *crowd newsworthiness* rating for each article, which is scaled from 1 (least newsworthy) to 5 (most newsworthy). The summation implies that we weighted them equally during index construction - we also evaluated more complex aggregation strategies such as by using regression weights, but these did not appear to offer any improvement when we conducted the correlation analyses described in the next section.

4.4.1 Correlation of Expert and Crowd Newsworthiness Ratings. In order to provide context to judge the agreement between crowd-workers and experts, we first measure the degree to which our expert evaluators agree with each other. The Spearman correlation of the newsworthiness ratings provided by each of the experts show a weak association ($r(55) = 0.2998$, $p = 0.026$), which underscores the variability and context-sensitivity of ratings of newsworthiness even amongst professionals. If experts have difficulty reaching consensus on newsworthiness, this supports the assertion that "who is applying news values can be as important as what news values are being applied" [43].

We next conduct a non-parametric Mann-Whitney test to compare the distributions of the *expert newsworthiness* (Mean=2.74, Median=2.5, SD=0.86) and the *crowd newsworthiness* (Mean=3.18, Median=3.14, SD=0.46), since the former does not exhibit a normal distribution. The test indicates that the *crowd newsworthiness* ratings were typically greater than *expert newsworthiness* ratings (U=2079.0, p=6.9e-04). To further distinguish the selectivity exhibited by crowd-workers and experts when judging newsworthiness, we compare the percentage of papers each group rates as potentially newsworthy, i.e. rates at greater than a value of 3, which is the intermediate value on our Likert scale in the survey. As per the *expert newsworthiness*, 15 papers (27.3%) of the 55 in the validation set are likely to be newsworthy, whereas according to the *crowd newsworthiness*, this figure is 33 papers (60.0%) of the 55. Such a visible gap - the experts are less than half as likely to rate a paper as newsworthy than crowd-workers - points to lesser selectivity among the crowd-workers.

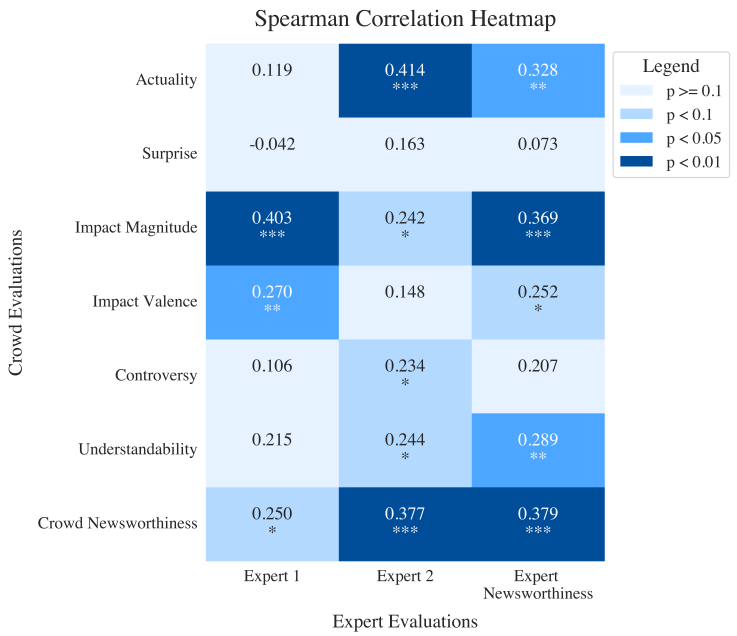

Fig. 1. Spearman correlations between Crowd Evaluations and Expert Evaluations. * => p < 0.1, ** => p < 0.05, *** => p < 0.01 for correlations

We also calculate the Spearman correlations between the different dimensions of the crowd-sourced ratings and the expert ratings, as presented in Figure 1. We observe a moderately positive association between the overall *crowd newsworthiness* and *expert newsworthiness* scores ($r(55) = 0.379$, $p = 0.004$), which compares well against the measured associations among the experts themselves.

This association provides some support to the idea that there is a signal in the multi-dimensional crowdsourced data that aligns to some extent with expert evaluation.

We also observe that both our experts tend to lean on different news values when evaluating overall newsworthiness, by examining the correlations between crowd-workers' ratings for individual news values and the expert ratings for overall newsworthiness. For instance, Expert 1's ratings correlate moderately with the crowd-rated *impact magnitude* ($r(55) = 0.403$, $p = 0.0023$) and *impact valence* ($r(55) = 0.270$, $p = 0.04$), whereas Expert 2's align moreso with the crowd-rated *actuality* ($r(55) = 0.414$, $p = 0.0017$). This is further evidence of the subjective and context-specific nature of the news selection task, and supports our approach of measuring a diverse set of news values and from multiple raters to try to capture some degree of this variation.

We further observe that *surprise* is the only news value that does not correlate sufficiently well with either of the expert evaluations, or the *expert newsworthiness* score derived from their average. Consequently, we experimented with using subsets of the six news values to aggregate for *crowd newsworthiness*, and discovered that dropping the *understandability* and *surprise* variables slightly improves the association between the *crowd newsworthiness* and *expert newsworthiness* scores ($r(55) = 0.407$, $p = 0.002$).

Further, the crowd-rated *understandability* and the algorithmically evaluated *readability* ratings are also moderately correlated at a statistically significant level ($r(55) = 0.524$, $p = 3.98e\text{-}05$). This points to the potential to approximate the crowd-workers' *understandability* of a paper in an automated and cost-effective manner, using the *readability*.

4.4.2 Takeaways from Comparative Analyses. We use the insights derived from our quantitative and qualitative analyses to make modifications to our final survey for the training set, as well as to establish the broader limitations of crowdsourcing judgements of newsworthiness and how they align with expert evaluations.

Broadly, we do observe that the recurrent themes uncovered within expert rationale align with the literature on traditional news values, commercial news values, and even news angles. However, we also observe that in order to reason about these concepts effectively, journalists sometimes rely on several contextual cues e.g. the status-quo in an article's broader discipline, common applications involving similar technology, etc. They also make inferences about the actualization and marketability of the story, in the context of different types of publications, audiences, and news frames. Such modes of reasoning require domain knowledge that our crowd-workers do not always possess, which is an important limitation of our approach towards building a model based on crowdsourced ratings of newsworthiness.

The most obvious manifestation of this incongruence is seen in the *surprise* ratings provided by crowd-workers, where their quantitative responses clearly do not align with either the individual or averaged expert opinions. As a result, we decide to drop *surprise* in the final iteration of the survey, and we exclude it in our calculation of aggregate *crowd newsworthiness* for the validation set.

We also see that the using *understandability* in the calculation of aggregate *crowd newsworthiness* actually reduces correlation to expert newsworthiness, as compared to without it. This occurs despite a moderate correlation between the *understandability* and the *expert newsworthiness* scores themselves ($r(55) = 0.289$, $p = 0.032$), leading us to deduce that while the *understandability* can be informative about an article's appeal to experts, the *crowd newsworthiness* rating itself does not benefit from it. Instead, *understandability* could be useful as an independent variable in the predictive model that we build in the next section. Also, because *readability* is a viable computational proxy for *understandability*, we drop *f* from our survey, and use *readability* in our model for efficiency reasons.

Overall, in terms of RQ1 and the alignment between crowd-worker ratings and expert ratings of newsworthiness, we find a moderate, statistically significant correlation between aggregate *crowd newsworthiness* and aggregate *expert newsworthiness* ratings. This alignment was similar to the alignment between different experts, suggesting that the crowdsourced ratings can act as a scalable and cost-effective proxy for expert responses that can be used to produce a larger corpus of rated material for training a predictive model. At the same time, the findings from the qualitative analysis indicate that there are important dimensions of newsworthiness that won't be captured by the crowd ratings (or model) and which therefore indicate that such a model should only be used to augment and accelerate expert evaluations, rather than attempt to replace them.

We apply the final piloted survey (See Appendix A), minus questions pertaining to *surprise* and *understandability*, to the entire training dataset. In the next section, we describe the predictive model we built to predict crowd newsworthiness for the training dataset, and discuss its performance on the validation dataset.

5 RQ2: PREDICTING NEWSWORTHINESS FROM CROWD-SOURCED RATINGS

5.1 Methods

We now use the broader conceptual findings from the previous section to address RQ2, by building and validating a prediction model that is trained on the crowd-worker ratings we collected for the training dataset. Like the process described in the previous section, we aggregate our selected news values: *actuality*, *impact magnitude*, *impact valence*, and *controversy* to an overall *crowd newsworthiness* which we aim to predict.

Our quantitative analyses in the previous section pointed to high average newsworthiness score on the part of the crowd-workers as compared to experts. We attempted to adjust for this bias by framing the prediction task as a classification problem: we convert the continuous crowd-worker ratings to binary values that indicate if an article could be potentially newsworthy (1) or not (0). To conduct this binarization, we attempted to find a meaningfully similar threshold of newsworthiness for crowd-worker ratings as compared to what we have for the experts (which is a Likert score of 3, i.e. the midpoint of the scale). To this end, we experimented with several methods: using the midpoint of the Likert scale itself, using the z-score of the midpoint from expert ratings to find a corresponding value for crowd-workers, and by regressing expert ratings on crowd ratings to find a meaningful equivalent. We found that our model gave the best performance on the training data when we simply set a threshold at 3 (i.e. the midpoint of the Likert scale) for the crowd-worker ratings: any article rated strictly above this qualified as potentially newsworthy for the purposes of training the classifier.

In order to train our model, we used a set of textual and metadata features derived from the arXiv API and the De-Jargonizer. We use the following features, with certain transformations as specified in parentheses: the De-Jargonizer readability score (re-scaled from 0 to 1), the arXiv author-assigned primary category (one-hot encoded), and the full text of the article. To featurize the text, we do some initial pre-processing to remove author data, references, hyperlinks, special characters, etc., and then use a pre-trained Sentence-BERT model [62] to generate 768-dimensional word embeddings per article. We experimented with using Term Frequency-Inverse Document Frequency (TF-IDF) to generate term weights from the text as well, but this yielded worse performance compared to the BERT-based model. We also generate a binary feature to indicate whether the arXiv article makes any data or code available to the public, which could potentially signal scientific replicability and trustworthiness to domain experts.

We explored a set of feature selection and classification algorithms, including Logistic Regression, Complement Naive Bayes [63], Random Forests [14], and Extra Trees [36], and discovered that the

Table 3. Precision@K of Predicted and Crowd Newsworthiness, with respect to Ground-Truth Expert-Annotated Newsworthiness

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

Extra Trees classifier consistently performed better than the other alternatives. We used Recursive Feature Elimination to select the optimal set of features for this classification pipeline.

An important design consideration we made in this process was that we weighed the precision of the predicted newsworthiness higher than the recall. Since science journalists are exposed to a massive set of potential leads from not only arXiv, but also scientific journals, Twitter threads, and so on, we believed that it was more important to provide them with a likely newsworthy set of recommendations i.e. reduce the false positives, than to exhaustively recommend all newsworthy articles i.e. reduce false negatives. We envisage the model as ultimately providing a ranked list of likely newsworthy scientific articles to journalists periodically, following which they conduct their reporting and exercise their own judgment to determine what could develop into a potential story.

To optimize our model for such a use-case, we tuned the parameters of the entire classification pipeline using 5-fold cross validation on the training dataset, and selected the set that provided us with the best F0.5 score, an adjusted version of the F1 score which weighs the precision more than the recall. We also ranked the final set of predictions by their likelihood of being "newsworthy" i.e. belonging to the positive class, and used these rankings to measure the precision of the top K ranked articles, for varying values of K (i.e. the P@K metric). The benefit of such a ranked list is that professionals can use the outputs as they see fit such as by viewing only some chosen top-K ranked items or by varying the probability thresholds of what is classified as "newsworthy" to increase selectivity.

5.2 Findings

In this section, we present and contextualize the predictions of our classification pipeline, by discussing its performance on the expert-annotated validation set. The code and data needed to replicate these findings is publicaly available at our GitHub repository[3].

As per the averaged expert ratings, 15 (27%) of 55 articles are labeled as "newsworthy" i.e. have a Likert rating greater than 3. Because we envision a deployment of the model such that expert science journalists are still empowered as the final judges of what is newsworthy, we compute a ranked list of potentially newsworthy articles for their perusal and follow-up. The rankings consist of the research articles ordered by decreasing predicted class probabilities for the positive, i.e. newsworthy class. We judge the quality of this list by measuring the precision of its predictions (i.e. the proportion of articles which are newsworthy as per experts) presented in its top K rankings, for varying values of K. Table 3 presents this Precision@K, and compares it to that obtained from using the aggregated crowdsourced newsworthiness rating to rank articles.

These evaluations indicate that the model performs substantially better than randomized selection (i.e. 27%), with 80% of its recommended top 10 newsworthy articles also being rated as newsworthy according to the averaged *expert newsworthiness* score. This is an indicator of our

[3]https://github.com/comp-journalism/predicting_newsworthiness

model's ability to capture expert sentiment on newsworthiness, and is further bolstered by the fact that its performance is better than or equal to the Precision@K of the aggregated crowd ratings for all the tested values of K in Table 3.

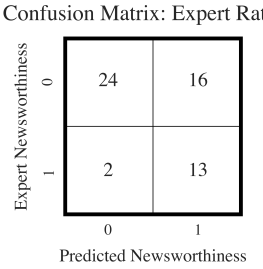

Fig. 2. Confusion Matrix of Expert-Annotated Newsworthiness vs. Predicted Newsworthiness of Model

We further examine the primary arXiv categories of the articles predicted to be highly newsworthy, to gain insight into how the model discriminates between different areas of study. The top twenty most newsworthy articles as per the model's predictions come from the following arXiv categories: Computers and Society (4), Human Computer Interaction (4), Machine Learning (3), Social and Information Networks (2), Computation and Language (2), Computer Vision (2), Networking and Internet Architecture (2), and Artificial Intelligence (1). Articles with the primary category being either Robotics or Information Retrieval do not find mention in the top twenty most newsworthy predictions of the model.

We also present the confusion matrix of the predictions with respect to *expert newsworthiness*, and qualitatively examine the abstracts that yield false positives or false negatives to derive insights about these deviations. Figure 2 indicates that the model has a precision of 0.45 i.e. slightly less than half of the model's recommendations are actually newsworthy, as defined by experts. The baseline precision in this case would be 0.27 i.e. correct predictions if all the items were predicted as "newsworthy", and our model performs visibly better. Additionally, the model exhibits a recall of 0.87 i.e. it correctly recovers 13 of the total 15 articles that averaged expert ratings deemed as "newsworthy". When it comes to the individually subjective ratings of our experts, the model has a recall of 0.60 for Expert 1's newsworthy-rated articles, and a recall of 1.0 for Expert 2's newsworthy-rated articles.

We first qualitatively examine the two article abstracts that experts rate as newsworthy but the model does not predict to be so, to trace potential reasons for such false negatives. We find that expert rationale for these abstracts were often thematically coded for themes such as *the article as a jumping-off point*, *potential audience/publication*, *scientific impact*, and *timeframe of social impacts*. In other words the errors appear to relate to themes that we identified might require an expert's level of domain knowledge, and which would be difficult to crowdsource. For these abstracts, the Likert ratings of newsworthiness as per experts (Mean 3.5, SD: 0.0) were in close proximity to the threshold of 3 that we defined for binarizing newsworthiness. We further look at the sixteen article abstracts that are predicted as potentially newsworthy by the model, but not by the experts. Experts often rationalize their rejection of these abstracts on account of the novelty aspect of *surprise*, the *framing and format possibilities* of a story, the *potential audience/publication*, and even the *marketability* of the story. Quantitatively, true expert ratings for these abstracts are once again, on average, relatively near the threshold of newsworthiness (Mean: 2.38, SD: 0.43), implying that

they might closely miss the cut. This further supports the idea for deploying the model as a ranking tool to make suggestions to journalists rather than a strict filtering tool using a threshold.

While this is a small sample of ratings and rationale, its evident variations in recall over individual expert ratings, thematic patterns of the false positives and false negatives, as well as the quantitative proximity of false positives and false negatives to classification thresholds, all reinforce certain ideas we have traced throughout this work, including that: (1) experts and audiences are sensitive to different themes in the scientific articles, and (2) expert judgements are also prone to high variance, making them harder to predict. Given the high precision of this model to recognize newsworthy articles via its rankings, in the face of the aforementioned challenges, it constitutes an important step forward in the literature on computational news discovery, specifically in the domain of science journalism. Our use of a minimal amount of metadata from the arXiv articles to predict *crowd* and *expert newsworthiness* also ensures the scalability and efficiency of the classification pipeline that we have built. The next section provides a discussion of the limitations of our approach, as well as the potential avenues for future work it could lead to.

6 DISCUSSION

In this work, we sought to examine the effectiveness of crowd-sourced evaluations of newsworthiness at approximating the opinions of domain experts, by measuring the extent to which they are aligned (RQ1). We also created a predictive model of newsworthiness based on crowdsourced ratings, in an endeavor to enable resource-constrained science journalists to discover potential leads at scale and with efficiency (RQ2).

6.1 Crowdsourcing Newsworthiness

In Section 4, we qualitatively analyzed expert ratings of scientific abstracts to uncover specific dimensions of newsworthiness that were amenable to crowdsourcing. We identified that journalists relied on certain inherent characteristics of the research article, as well as on the specific logistical factors involved in news production, in order to judge newsworthiness. Establishing that the former category of criteria would require relatively less domain expertise, we next collected crowd ratings by cuing workers for a selected set of news values, so as to help them consider what an abstract term like "newsworthiness" could tangibly mean.

The observed correlations between experts and crowd-workers suggest that crowd-worker ratings are moderately associated with averaged expert ratings. Crowd ratings for some news values (e.g., actuality, impact magnitude) exhibit greater correlation with expert ratings than some others (e,g., surprise). This finding indicates that aggregated crowdsourced ratings of newsworthiness, based on examining the abstracts of scientific articles, can operate as a proxy for aggregated expert ratings. These correlations also vary over individual experts, which aligns with the scholarly understanding of how different news values are contextually applied by different journalists [43]. The different correlations of crowd ratings for each of our experts also indicate the potential usefulness of a news discovery recommendation strategy that is personalized according to journalists' interests, audiences, and constraints. Given the contextual variations in newsworthiness evaluations, such a recommender might be bootstrapped based on crowd-ratings, but then adapt to the unique interests and context of each journalist. For instance, a journalist working for *Wired* may have different organizational and audience constraints in comparison to one working for *NPR* or *Scientific American*, which a recommender model might be tuned towards based on feedback from individual journalists.

One potential improvement on this work could then involve priming crowd-workers to not only rate an article more effectively along traditional news values, but to actively expand their notion of *newsworthiness* itself, so that they contemplate some of the expert-exclusive themes and

criteria as well. For instance, one could encourage crowd-workers to consider the timeframe of potential impacts (i.e. the immediacy of impacts) in their evaluations of relevance. They could also be prompted to consider not only the article presented, but also other ideas in its proximity, that could make the article a jumping-off point for related big-picture stories. We observed that crowd-workers in their rationale already exhibit a tendency to extrapolate meso-scale scientific results to their macro-scale societal outcomes, however it would require future work that was outside the scope of the current study to understand how this ability to recognize context could be channeled more effectively in a crowd task. Accomplishing this could lead to greater utility for an end-user system, since journalists often see leads as a starting point and stepping stone to broader stories [27]. Finally, one could imagine transforming article abstracts (e.g. using generative models such at GPT-3 fine-tuned on science news headlines) such that crowds might effectively weigh in on the dimension of marketeability that experts considered.

6.2 Predicting Newsworthiness

Based on the general (though clearly not complete) alignment of expert and crowd-worker opinions that we observed across our analyses, we endeavored to train a classification model in Section 5 to predict the likelihood of newsworthiness of individual articles, based on arXiv categories, textual content, an independent readability score, and the corresponding crowd-worker ratings. Despite the subjectivity of the newsworthiness prediction task and the moderate correlations of quantitative ratings between crowd-workers and experts, our predictive model provides a reasonably precise set of ranked recommendations, when measured against expert ratings of newsworthiness. Indeed, the precision of the model output based on the top K rankings was higher than that of the crowd-worker ratings for the same dataset. This demonstrates our model's ability to glean information from the textual features we provide in the form of Sentence-BERT embeddings. A qualitative analysis of the false negatives and the false positives also reveals that they were coded most frequently with expert-specific criteria we uncovered in Section 4.1.2, which are factors we now know create differences in judgment between the model as trained on crowd ratings and the expert ratings.

Predictive models of news coverage for science journalism in the past have significantly relied on the existence of press releases and information subsidies to stimulate coverage [51]. Our modeling methodology supports the idea of a *computational* information subsidy insofar as it enables a scalable and effective approach for identifying preprints that are more likely to be interesting for development into news items. At the same time, any subsidy whether it be a press release or a computational model, comes with caveats such as potential biases [23]. Variance in expert judgement, combined with the biases and blind spots of the model driven by the limitations in our crowdsourcing approach and/or the crowd itself, suggest that larger scale deployments and evaluations of such a model will be necessary to truly assess professional utility. We envision our model being deployed in such a way that it ranks potential leads based on their metadata and textual content, and experts remain the ultimate decision-makers about what becomes news. We do also acknowledge that any such deployment of this model to real journalists would need to be accompanied by substantial transparency information into how the model works, is trained, and what it might miss as a result [26, 55], such as leads that are newsworthy for reasons that crowds are unable to recognize.

Another question worth considering in future work is whether crowd-sourced ratings and a model's predictions can still provide professional value despite mis-alignments. We validated our crowd and predicted ratings under the assumption that by aligning them with expert ratings, they would be more acceptable to incorporation into the workflows of those experts. But it is also worth considering that instead we might have the audience i.e. the crowd-workers, act as the arbiters of what is "newsworthy", and consequently formulate a model that aims to predict

the interests of the crowd: What is it that they would like to hear about from the journalists they trust? Particularly in light of appropriate transparency information that journalists might use to inform their understanding of a model's biases, they might come to put those biases in dialogue with other professional goals such as around 'engaged journalism' or the close incorporation of the information needs of the community they serve into their practice [31]. Future work along these lines could probe how the false positives of a model such as ours might be received by journalists, and whether they might convey some meaningful signals about the stories that an audience cares about. Of course, this inclusion of the audiences as gatekeepers in the newsroom, by proxy of readership metrics and social media engagement, has received heightened attention in the past decade [4, 59], but whether this could be a positive development for science journalism is worth contemplating, and perhaps evaluating in future deployments of models such as the one we developed. While crowd-sourced ratings of potential newsworthiness could point journalists to novel scientific research that their audiences would care to learn and read about, this pursuit could also encourage the tabloidization of the news media that existing readership metrics may contribute to [33].

In this vein, it is also worth considering how the widespread deployment of such a system might impact the presentation of scientific abstracts that it assesses for newsworthiness. Bucher has theorized that the content recommendation algorithms on online platforms like Facebook construct a "regime of visibility" that "imposes a perceived 'threat of invisibility' on the part of the participatory subject." [18]. Cotter has empirically demonstrated how digital creators on platforms such as Instagram thus "play the visibility game" as a response, strategizing and refining the presentation of their content on the basis of their understanding of what the algorithm rewards and what it doesn't [21]. Unlike these online platforms however, our recommendation system lacks specific engagement metrics that might inform scientists about which abstracts received what kind of attention from journalists. Further, such a system for computational news discovery would ultimately be part of a complex socio-technical environment within the newsroom, where journalists are often cognizant of how algorithmic tools apply powerful filters on what is visible to any end-user [59], and are thus careful to prioritize editorial voices [39] and even make explicit calls for impartiality and accountability from such tools [9]. In practice, it would thus be difficult for researchers to gain tangible insights into which specific features of abstracts garnered attention, and leverage this to "play the visibility game". We do however recognize that the lack of such concrete feedback loops would not necessarily prevent folk-theorization about what the recommender rewards [22], and could still result in the adoption of practices to optimize the language and metadata of research articles for newsworthiness. A simple instance of modifying language could be to decrease the use of jargon to enhance understandability, whereas adjustments to metadata could mean that an article is strategically assigned to less technical primary categories on arXiv, and instead to something more inter-disciplinary such as Computers and Society. Ultimately, some features such as an article's primary category or keyword-based understandability would be easier to control than others, such as the embedding vectors generated from its text. As such, future research could also delve into how scientists view and adjust for such algorithmic gatekeepers in their endeavor to communicate their work to the larger public.

6.3 Limitations

Finally, it's worth underscoring a few limitations of our work here. A key limitation is that we conducted our experiments for a specific set of arXiv categories within the domain of Computer Science, based on their general understandability and wide-spread applicability. While this choice enabled us to collect data and draw inferences in a streamlined manner, it's also important to note that this limits the generalizability of our results. For instance, disciplines that may require greater

domain knowledge from crowd-workers to interpret article abstracts and their contributions, such as Mathematical Physics (math-ph), Control Systems (eess.SY), or Econometrics (econ.EM) may not be as amenable to crowdsourcing. While we demonstrate the feasibility of our approach in the domain of Computer Science, future work will need to replicate the approach to assess the external validity in other domains.

Another limitation hinges on how we define and scope "newsworthiness" for crowd-worker ratings, keeping in mind that there are certain expert-specific criteria (e.g., marketability, nature of audiences) that they might not be able to reason about. This limits the output of our predictive model, since it can only capture a part of an article's appeal to journalists. Our proposed deployment scenario accounts for this constraint, where the model ranks leads for experts as part of their "news discovery" process, following which they conduct their own follow-ups and decide upon a lead's suitability for a news story. However, future work could involve expanding the crowd's conception of newsworthiness, beyond the characteristics of the research presented to them in the article itself, in order to assess how this may impact the performance of the classifier.

Finally, our work measures the alignment between experts and crowd-workers based on ratings aggregated from three crowd-workers for each article. This was the consequence of iterative pilots where we compared sample sizes in the range of three to five workers, and also aimed to lower the cost of collecting and processing survey responses. However, future work could explore how sample sizes larger than five impact the agreement between crowd-workers and experts, in a situation where more resources are available.

7 CONCLUSIONS

Science journalists today face considerable information overload by virtue of the sheer volume of scientific publication, across different avenues and outlets. At the same time, their role as the interpreters, communicators, and critics of new scientific findings is crucially important in society. We have approached this research with the intent of crowdsourcing potential leads for science journalists from arXiv articles, in order to reduce the informational burden they might face in confronting this corpus otherwise. We chose to operationalize a set of news values for which we collected ratings from crowd-workers, as well as the rationale behind those ratings.

Over the course of our analysis, we discovered that aggregated crowdscourced ratings of newsworthiness moderately align with aggregated journalistic ratings of newsworthiness, albeit with a few caveats. For one, we did not cue crowd-workers to examine articles from the point-of-view of their potential as *publication-worthy content* that could emerge from the news production pipeline of writing, framing, marketing to audiences, and so on, since we assume they lacked the expertise to do so. For some dimensions that we did cue them to consider (e.g. surprise), we found that associations between crowd-workers and experts were negligibly small. Another challenge we faced pertains to the aggregation of expert opinion itself, in that journalists' assessments of newsworthiness exhibit substantial variance already.

Despite these factors, the predictive model we train on crowd-sourced ratings offers a reasonably precise set of ranked recommendations of potential leads, when validated against expert opinion. This model is also resource-efficient, in that it featurizes basic arXiv metadata, and combines it with an automated readability score that has been validated using crowd-worker ratings. This expert-focused formulation of predicting newsworthiness of scientific articles, along with our qualitative investigations into the factors that affect performance, constitutes a novel contribution to the existing literature on computational news discovery for resource-constrained journalists. We also contribute to the literature on crowd-sourcing and its validity as a method for capturing latent themes in data, for the benefit of building general computational tools for journalists and understanding the alignment between crowd and expert evaluations of news values.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation via award IIS-1845460. We also thank the journalists who volunteered their time to participate in our research and provided their valuable insights!

REFERENCES

[1] Tanja Aitamurto. 2016. Crowdsourcing as a Knowledge-Search Method in Digital Journalism: Ruptured ideals and blended responsibility. *Digital Journalism* 4, 2 (Feb. 2016), 280–297. https://doi.org/10.1080/21670811.2015.1034807

[2] Stuart Allan. 2011. Introduction: Science journalism in a digital age. *Journalism* 12, 7 (Oct. 2011), 771–777. https://doi.org/10.1177/1464884911412688

[3] Sigurd Allern. 2002. Journalistic and Commercial News Values: News Organizations as Patrons of an Institution and Market Actors. *Nordicom Review* 23, 1-2 (Sept. 2002), 137–152. https://doi.org/10.1515/nor-2017-0327 Publisher: Sciendo Section: Nordicom Review.

[4] Cw Anderson. 2011. Between creative and quantified audiences: Web metrics and changing patterns of newswork in local US newsrooms. *Journalism* 12, 5 (July 2011), 550–566. https://doi.org/10.1177/1464884911402451

[5] Chittaranjan Andrade. 2011. How to write a good abstract for a scientific paper or conference presentation. *Indian journal of psychiatry* 53, 2 (April 2011), 172–175. https://doi.org/10.4103/0019-5545.82558 Publisher: Medknow Publications.

[6] Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34, 4 (dec 2008), 555–596. https://doi.org/10.1162/coli.07-034-R2

[7] Earl R. Babbie. 2008. *The basics of social research* (4th ed ed.). Thomson/Wadsworth, Belmont, CA.

[8] 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.). Vol. 28. Springer Netherlands, Dordrecht, 59–85. https://doi.org/10.1007/978-94-007-2085-5_4 Series Title: Sociology of the Sciences Yearbook.

[9] Mariella Bastian, Natalie Helberger, and Mykola Makhortykh. 2021. Safeguarding the Journalistic DNA: Attitudes towards the Role of Professional Values in Algorithmic News Recommender Designs. *Digital Journalism* 9, 6 (July 2021), 835–863. https://doi.org/10.1080/21670811.2021.1912622

[10] Rena Kim Bivens. 2008. The Internet, Modile Phones and Blogging: How new media are transforming traditional journalism. *Journalism Practice* 2, 1 (Feb. 2008), 113–129. https://doi.org/10.1080/17512780701768568

[11] Deborah Blum. 2021. Science journalism grows up. *Science* 372, 6540 (April 2021), 323–323. https://doi.org/10.1126/science.abj0434 Publisher: American Association for the Advancement of Science.

[12] Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications* 8, 1 (Oct. 2021), 224. https://doi.org/10.1057/s41599-021-00903-w

[13] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[14] L. Breiman. 2004. Random Forests. *Machine Learning* 45 (2004), 5–32.

[15] Aengus Bridgman, Eric Merkley, Peter John Loewen, Taylor Owen, Derek Ruths, Lisa Teichmann, and Oleg Zhilin. 2020. The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media. *Harvard Kennedy School Misinformation Review* (June 2020). https://doi.org/10.37016/mr-2020-028

[16] Paul Brighton and Dennis Foy. 2007. *News values.* SAGE Publications, London ; Thousand Oaks, Calif. OCLC: ocn153557904.

[17] Dominique Brossard and Dietram A. Scheufele. 2013. Science, New Media, and the Public. *Science* 339, 6115 (Jan. 2013), 40–41. https://doi.org/10.1126/science.1232329

[18] Taina Bucher. 2012. Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New Media & Society* 14, 7 (Nov. 2012), 1164–1180. https://doi.org/10.1177/1461444812440159

[19] Jesse Chandler, Gabriele Paolacci, and Pam Mueller. 2013. Risks and Rewards of Crowdsourcing Marketplaces. In *Handbook of Human Computation*, Pietro Michelucci (Ed.). Springer New York, New York, NY, 377–392. https://doi.org/10.1007/978-1-4614-8806-4_30

[20] Scott Clifford, Ryan M Jewell, and Philip D Waggoner. 2015. Are samples drawn from Mechanical Turk valid for research on political ideology? *Research & Politics* 2, 4 (Oct. 2015), 205316801562207. https://doi.org/10.1177/2053168015622072

[21] Kelley Cotter. 2019. Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram:. *New Media & Society* 21, 4 (April 2019), 895–913. https://doi.org/10.1177/1461444818815684

[22] Michael A. DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. "Algorithms Ruin Everything": #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the 2017 CHI Conference on Human*

Factors in Computing Systems (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3163–3174. https://doi.org/10.1145/3025453.3025659

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

[24] Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. 2012. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2451–2460. https://doi.org/10.1145/2207676.2208409

[25] Nicholas Diakopoulos, Madison Dong, Leonard Bronner, and Jeremy Bowers. 2020. Generating Location-Based News Leads for National Politics Reporting. In *Proc. Computation + Journalism Symposium*.

[26] Nicholas Diakopoulos and Michael Koliska. 2016. Algorithmic Transparency in the News Media. *Digital Journalism* 5, 7 (08 2016), 1–20. https://doi.org/10.1080/21670811.2016.1208053

[27] Nicholas Diakopoulos, Daniel Trielli, and Grace Lee. 2021. Towards Understanding and Supporting Journalistic Practices Using Semi-Automated News Discovery Tools. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 406, 30 pages. https://doi.org/10.1145/3479550

[28] Murray Dick. 2014. Interactive Infographics and News Values. *Digital Journalism* 2, 4 (Oct. 2014), 490–506. https://doi.org/10.1080/21670811.2013.841368

[29] Sharon Dunwoody. 2012. Science journalism. In *Routledge Handbook of Public Communication of Science and Technology*. Routledge. https://doi.org/10.4324/9780203483794.ch3

[30] Declan Fahy and Matthew C. Nisbet. 2011. The science journalist online: Shifting roles and emerging practices. *Journalism* 12, 7 (Oct. 2011), 778–793. https://doi.org/10.1177/1464884911412697

[31] Patrick Ferrucci, Jacob L Nelson, and Miles P Davis. 2020. From "Public Journalism" to "Engaged Journalism": Imagined Audiences and Denigrating Discourse. *International Journal of Communication* 14, 0 (02 2020), 19.

[32] Richard Fletcher, Steve Schifferes, and Neil Thurman. 2020. Building the 'Truthmeter': Training algorithms to help journalists assess the credibility of social media sources. *Convergence* 26, 1 (Feb. 2020), 19–34. https://doi.org/10.1177/1354856517714955 Publisher: SAGE Publications Ltd.

[33] Silke Fürst. 2020. In the Service of Good Journalism and Audience Interests? How Audience Metrics Affect News Quality. *Media and Communication* 8, 3 (Aug. 2020), 270–280. https://doi.org/10.17645/mac.v8i3.3228

[34] Johan Galtung and Mari Holmboe Ruge. 1965. The Structure of Foreign News: The Presentation of the Congo, Cuba and Cyprus Crises in Four Norwegian Newspapers. *Journal of Peace Research* 2, 1 (March 1965), 64–90. https://doi.org/10.1177/002234336500200104

[35] Tong Gao, Jessica R. Hullman, Eytan Adar, Brent Hecht, and Nicholas Diakopoulos. 2014. NewsViews: an automated pipeline for creating custom geovisualizations for news. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Toronto Ontario Canada, 3005–3014. https://doi.org/10.1145/2556288.2557228

[36] P. Geurts, D. Ernst, and L. Wehenkel. 2006. Extremely randomized trees. *Machine Learning* 63 (2006), 3–42.

[37] Peter Golding and Philip Ross Courtney Elliott. 1979. *Making the news* (1st ed ed.). Longman, London ; New York.

[38] Andreas Graefe. 2016. Guide to Automated Journalism. (2016). https://doi.org/10.7916/D80G3XDJ Publisher: Columbia University.

[39] Marisela Gutierrez Lopez, Colin Porlezza, Glenda Cooper, Stephann Makri, Andrew MacFarlane, and Sondess Missaoui. 2022. A Question of Design: Strategies for Embedding AI-Driven Tools into Journalistic Work Routines. *Digital Journalism* (March 2022), 1–20. https://doi.org/10.1080/21670811.2022.2043759

[40] James Hamilton. 2004. *All the News That's Fit to Sell.* Princeton University Press. https://doi.org/10.2307/j.ctt7smgs

[41] James Hamilton and Fred Turner. 2009. Accountability Through Algorithm: Developing the Field of Computational Journalism. *Center For Advanced Study in the Behavioral Sciences Summer Workshop* (2009). http://web.stanford.edu/~fturner/Hamilton%20Turner%20Acc%20by%20Alg%20Final.pdf

[42] Tony Harcup and Deirdre O'Neill. 2001. What Is News? Galtung and Ruge revisited. *Journalism Studies* 2, 2 (Jan. 2001), 261–280. https://doi.org/10.1080/14616700118449

[43] Tony Harcup and Deirdre O'Neill. 2017. What is News?: News values revisited (again). *Journalism Studies* 18, 12 (Dec. 2017), 1470–1488. https://doi.org/10.1080/1461670X.2016.1150193

[44] Andrew J. Hoffman. 2021. *The engaged scholar : expanding the impact of academic research in today's world.* Stanford University Press.

[45] Alok Jha. 2010. What is a story tracker? https://www.theguardian.com/science/blog/2010/jun/09/science-story-trackers

[46] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E. Kraut. 2011. CrowdForge: crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*. ACM Press, Santa Barbara, California, USA, 43. https://doi.org/10.1145/2047196.2047202

[47] Peder Olesen Larsen and Markus von Ins. 2010. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics* 84, 3 (Sept. 2010), 575–603. https://doi.org/10.1007/s11192-010-

0202-z

[48] Fabienne Lind, Maria Gruber, and Hajo G. Boomgaarden. 2017. Content Analysis by the Crowd: Assessing the Usability of Crowdsourcing for Coding Latent Constructs. *Communication Methods and Measures* 11, 3 (July 2017), 191–209. https://doi.org/10.1080/19312458.2017.1317338

[49] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Sameena Shah, Robert Martin, and John Duprey. 2017. Reuters Tracer: Toward Automated News Production Using Large Scale Social Media Data. *arXiv:1711.04068 [cs]* (Nov. 2017). http://arxiv.org/abs/1711.04068 arXiv: 1711.04068.

[50] Tetyana Lokot and Nicholas Diakopoulos. 2016. News Bots: Automating news and information dissemination on Twitter. *Digital Journalism* 4, 6 (Aug. 2016), 682–699. https://doi.org/10.1080/21670811.2015.1081822

[51] Ansel MacLaughlin, John Wihbey, and David Smith. 2018. Predicting News Coverage of Scientific Articles. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (June 2018). https://ojs.aaai.org/index.php/ICWSM/article/view/14999 Section: Full Papers.

[52] Måns Magnusson, Jens Finnäs, and Leonard Wallentin. 2016. Finding the news lead in the data haystack: Automated local data journalism using crime data. (2016), 4.

[53] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44, 1 (March 2012), 1–23. https://doi.org/10.3758/s13428-011-0124-6

[54] Melinda McClure Haughey, Meena Devii Muralikumar, Cameron A. Wood, and Kate Starbird. 2020. On the Misinformation Beat: Understanding the Work of Investigative Journalists Reporting on Problematic Information Online. *Proc. of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–22. https://doi.org/10.1145/3415204

[55] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. (2019), 220–229.

[56] Enrico Motta, Enrico Daga, Andreas L. Opdahl, and Bjornar Tessem. 2020. Analysis and Design of Computational News Angles. *IEEE Access* 8 (2020), 120613–120626. https://doi.org/10.1109/ACCESS.2020.3005513

[57] Priyanka Nanayakkara, Jessica Hullman, and Nicholas Diakopoulos. 2021. Unpacking the expressed consequences of AI research in broader impact statements. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 795–806.

[58] Nature. 2019. How to Write a Nature Summary. https://www.nature.com/documents/nature-summary-paragraph.pdf

[59] Chelsea Peterson-Salahuddin and Nicholas Diakopoulos. 2020. Negotiated Autonomy: The Role of Social Media Algorithms in Editorial Decision Making. *Media and Communication* 8, 3 (2020), 12.

[60] Tzipora Rakedzon, Elad Segev, Noam Chapnik, Roy Yosef, and Ayelet Baram-Tsabari. 2017. Automatic jargon identifier for scientists engaging with the public and science communication educators. *PLOS ONE* 12, 8 (Aug. 2017), e0181742. https://doi.org/10.1371/journal.pone.0181742

[61] Zvi Reich. 2006. The Process Model of News Initiative: Sources lead first, reporters thereafter. *Journalism Studies* 7, 4 (Aug. 2006), 497–514. https://doi.org/10.1080/14616700600757928

[62] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3980–3990. https://doi.org/10.18653/v1/D19-1410

[63] Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, and David R. Karger. 2003. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003), Washington, D.C.* (proceedings of the twentieth international conference on machine learning (icml 2003), washington, d.c. ed.). https://www.microsoft.com/en-us/research/publication/tackling-poor-assumptions-naive-bayes-text-classifiers/

[64] Carole Rich. 2016. *Writing and reporting news: a coaching method* (8th ed ed.). Cengage Learning, Boston, MA. OCLC: ocn908085863.

[65] Daniel Riffe, Stephen Lacy, Brendan R. Watson, and Frederick Fico. 2019. *Analyzing Media Messages: Using Quantitative Content Analysis in Research* (4th ed.). Lawrence Erlbaum, Mahwah, NJ, USA.

[66] Ida Schultz. 2007. The Journalistic Gut Feeling: Journalistic doxa, news habitus and orthodox news values. *Journalism Practice* 1, 2 (June 2007), 190–207. https://doi.org/10.1080/17512780701275507

[67] Raz Schwartz, Mor Naaman, and Rannie Teodoro. 2015. Editorial Algorithms: Using Social Media to Discover and Report Local News. *Proceedings of the International AAAI Conference on Web and Social Media* 9, 1 (April 2015). https://ojs.aaai.org/index.php/ICWSM/article/view/14633 Number: 1.

[68] C. Estelle Smith, Xinyi Wang, Raghav Pavan Karumur, and Haiyi Zhu. 2018. [Un]breaking News: Design Opportunities for Enhancing Collaboration in Scientific Media Production. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173955

[69] Alexander Spangher, Nanyun Peng, Jonathan May, and Emilio Ferrara. 2021. Modeling "Newsworthiness" for Lead-Generation Across Corpora. *arXiv:2104.09653 [cs]* (April 2021). http://arxiv.org/abs/2104.09653 arXiv: 2104.09653.

[70] Julie Suleski and Motomu Ibaraki. 2010. Scientists are talking, but mostly to each other: a quantitative analysis of research represented in mass media. *Public Understanding of Science* 19, 1 (Jan. 2010), 115–125. https://doi.org/10.1177/0963662508096776

[71] Kyle A. Thomas and Scott Clifford. 2017. Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior* 77 (Dec. 2017), 184–197. https://doi.org/10.1016/j.chb.2017.08.038

[72] Damian Trilling, Petro Tolochko, and Björn Burscher. 2017. From Newsworthiness to Shareworthiness: How to Predict News Sharing Based on Article Characteristics. *Journalism & Mass Communication Quarterly* 94, 1 (March 2017), 38–60. https://doi.org/10.1177/1077699016654682

[73] Johanna Vehkoo. 2013. Crowdsourcing in Investigative Journalism. *Reuters Institute for the Study of Journalism* Oxford: Report (2013).

[74] Meng-Han Wu and Alexander Quinn. 2017. Confusing the crowd: Task instruction quality on amazon mechanical turk. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 5. 206–215.

A APPENDIX

Instructions

- You will read some information about a recently published research paper, and then answer a survey where you rate that paper on a few dimensions.
- As you read, please consider the goals, methods, results, and implications of the research paper.
- For your ratings:
  - You will indicate your degree of agreement or disagreement with several statements about the research. Because these are subjective evaluations we will not reject your work on the basis of these ratings.
  - You will also provide a 2-3 sentence justification for each rating. Your justifications should satisfy the following criteria for your work to be accepted: (1) **Complete**: written in complete sentences; (2) **Original**: written by you and not copy-pasted from the task; (3) **Specific**: including a detailed explanation of why you rated it the way you did.

Task Information

Please read the following title and abstract for a recently published research paper:

Title: ${title}

Abstract:

${summary}

Additional Information: If you feel that you require additional information, you may consult the original research paper: ${arxiv_url}

Survey Questions

Now, please answer each of the following questions, and provide the appropriate explanations:

1. Please type in the fourth letter of the abstract of the paper in the text box, so we know that you are paying attention:

2. Please rate the potential impact of the research described on society, such as to individuals, organizations, politics, or the economy.

1	2	3	4	5
Strongly Negative	Negative	Neither Negative nor Positive	Positive	Strongly Positive

Please explain and justify your rating in 2-3 sentences:

3. Please rate how strongly you agree or disagree with the following statement:

"The research described has the potential to impact a significant number of people."

1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

Please explain and justify your rating in 2-3 sentences:

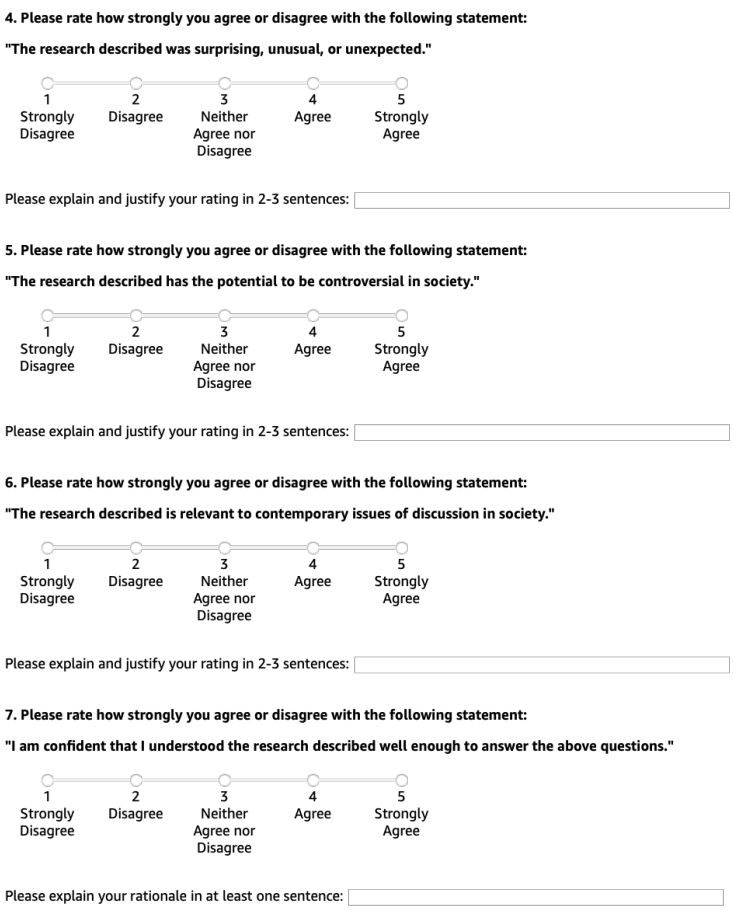

Fig. 3. Instructions and layout for the crowdsourced newsworthiness assessment task.

Received January 2022; revised April 2022; accepted May 2022