Predicting News Coverage of Scientific Articles

Ansel MacLaughlin
College of Computer and Information Science
Northeastern University
Boston, MA

John Wihbey
College of Arts, Media and Design
Northeastern University
Boston, MA

David A. Smith
College of Computer and Information Science
Northeastern University
Boston, MA

Abstract

Journalists act as gatekeepers to the scientific world, controlling what information reaches the public eye and how it is presented. Analyzing the kinds of research that typically receive more media attention is vital to understanding issues such as the "science of science communication" (National Academies of Sciences, Engineering, and Medicine 2017), patterns of misinformation, and the "cycle of hype." We track the coverage of 91,997 scientific articles published in 2016 across various disciplines, publishers, and news outlets using metadata and text data from a leading tracker of scientific coverage in social and traditional media, Altmetric. We approach the problem as one of ranking each day's, or week's, papers by their likely level of media attention, using the learning-to-rank model lambdaMART (Burges 2010). We find that ngram features from the title, abstract and press release significantly improve performance over the metadata features journal, publisher, and subjects.

1 Introduction

News media are an important source of scientific information for the public in domains such as health, medicine, and climate change research, making the accurate communication of findings – and patterns of misinformation – a vital issue for society and public policy (Geller, Bernhardt, and Holtzman 2002; Boykoff and Boykoff 2004; Brechman, Lee, and Cappella 2009). A substantial amount of scholarly attention has been devoted to studying the mechanisms by which academic research findings move along and through a chain of translation-oriented intermediaries, from journals and university communications offices to reporters and bloggers and finally to citizen groups and individual members of the public.

Journalists' internal processes of selection and framing news are very significant, as they are an antecedent and structural factor that affects all subsequent issues of public attention. Because parsing and interpreting the methods and results of highly technical scientific papers is a difficult task, journalists may use simple heuristics to determine whether research is novel and of high quality and impact, such as inspecting the name of the journal and the article subjects. An even simpler explanation is that journalists may publish on a subset of the press releases they read each day, either found on a press release aggregation website such as EurekAlert! or Science Daily or sent to them by media offices. Discovering what sort of content journalists believe is significant and likely to be popular provides insight into how journalists judge the newsworthiness of different content, uncovering their biases and preferences, and helps shed light on high-profile issues being actively debated relating to the "science of science communication" (National Academies of Sciences, Engineering, and Medicine 2017).

Scholars who have studied the decision-making of journalists in terms of story selection and framing have examined various factors, from institutional and economic incentives to a need for social validation (Donsbach 2004). In the realm of science, practices such as embargoing of scholarly findings and the attendant effects on communication have been examined, suggesting that certain science communications and public relations strategies can influence news coverage (Kiernan 2003b). Research has suggested that journalists seldom choose to report on scientific studies outside of the domains of health and medicine (Suleski and Ibaraki 2009).

Issues of hype and errors of framing and omission have long plagued various forms of news media and science communication, with the cycle of hype sometimes beginning with researchers themselves using exaggerated terms such as "breakthrough" in press materials and public announcements (Bubela et al. 2009). News reporting may then replicate the exaggerations found in such communications materials (Sumner et al. 2014). Further, news stories, particularly those produced by elite outlets, may influence the prestige of, and citations to, scientific research, fueling a hype cycle that is increasingly coming under scrutiny in the research and media communities (Kiernan 2003a; Caulfield et al. 2016).

1.1 Current Work

Using the metadata supplied by Altmetric,[1] we are able to find and crawl thousands of scientific article abstracts, press releases, and related news stories. Previous work started their investigations from a small number of journals, a small number of university press offices, or a small number of me-

[1] http://altmetric.com

dia outlets. This work, however, is the first, to our knowledge, analysis of news coverage of scientific articles across hundreds of disciplines and journals and thousands of news outlets.

1.2 Summary of Results

In summary, we find that the text of a scientific article's abstract, press release, and title, along with metadata on the subjects, journal and publisher are indicative of news coverage. We model journalists' daily or weekly selection process as one of ranking scientific articles to cover. This allows us to deal naturally with the resource constraints in any given outlet, as there are only so many slots for science stories in a given daily or weekly paper. We learn to rank a list of the scientific articles published on a given day or week using lambdaMART (Burges 2010) and optimizing $NDCG$. We find that having a press release is the most important feature for predicting article rank. Running just on the subset of scientific articles with press releases published on EurekAlert! or Science Daily (approximately 22% of the articles), performance decreases on the ranking task. This performance drop is perhaps indicative of the utility of press releases to journalists. In order to cover an article without a press release, journalists must devote significantly more time to reading and researching to determine if and how to write a piece.

2 Related Work

There are three lines of research relevant to the task at hand. First, we discuss research on altmetrics, their reliability, and their utility. Second, by examining surveys and research relating to the changing media ecosystem, we provide vital context for our analysis. Finally, we discuss computational approaches to predicting the popularity of content, scientific and not, on social and mainstream media.

2.1 Altmetrics

In 2010, the term "altmetrics" was coined (Priem et al. 2010). Altmetrics include a wide variety of counts, from coverage in mainstream news and social media (tweets, shares, likes, etc) to citations on Wikipedia. In the past decade, research has examined the utility of these new metrics to authors, universities, journals and repositories (Piwowar 2013; Priem, Groth, and Taraborelli 2012); what data ought to be included by altmetric providers and how to measure them (Bornmann 2014); and the correlation of various new metrics to traditional ones (Thelwall et al. 2013; Costas, Zahedi, and Wouters 2014). One significant problem with altmetrics is that they are easy to manipulate. It is not difficult to use a bot to artificially inflate the number of tweets or mentions of a given article. Although altmetric providers, such as our data source Altmetric, make efforts to identify and remove false coverage, a thorough examination of the data and sources is necessary. As is discussed later, counts of news coverage of scientific articles in our dataset are artificially inflated by non-content-creating websites which copy abstracts and press releases from other sources.

2.2 Media Ecosystem Context

There has also been significant research by scientists and journalists investigating what kind of scientific topics are covered in the news and how journalists select articles and convey information (Tanner 2004; Viswanath et al. 2008). In a study of health journalists, Len-Rios et al. (2009) find that journalists often use public relations materials when reporting on scientific news. Additionally, they find that, when attempting to judge the potential newsworthiness of a scientific article, journalists often consult other news outlets. Woloshin and Schwartz (2002) interview press officers and study press releases of 9 high-profile journals. They find that press offices select articles for press releases based on perceived newsworthiness and that most press releases exaggerate the importance of findings and do not note the limitations of the study. A survey of 99 scientific journalists by Harvard's Shorenstein Center on Media, Politics, and Public Policy (2016) investigates difficulties journalists face when covering medical science articles. Asked what factors limit their efforts to report on medical science issues in a timely and knowledgeable way, 42 journalists list that "insufficient time to do adequate background research" and 32 that "difficulty in determining whether a health/medical claim is valid" are major problems.

The role of the science journalist has been changing with the advent of digital and social media, during which time there has been a dramatic contraction in the numbers of specialized science journalists (Russell 2006). At the same time, some science journalists have become more interactively engaged than ever with scientists and interested audiences, and the online world has witnessed journalists taking a stance that is more critical toward sources and more interpretative as compared to science journalism of the past (Fahy and Nisbet 2011).

Still, journalists often lack formal training in statistics and lack a knowledge-based understanding of issues (Patterson 2013; Wihbey and Coddington 2017). On issues involving uncertainty or controversial science, or in domains that have been highly politicized, this lack of formal training as well as traditional norms of "newsworthiness" – where sensation and controversy are seen as virtues – can make reporters more likely to echo faulty claims. Journalists may be inclined to employ "false balance" on scientific issues, or giving equal credence to two "sides" despite the weight of evidence supporting only one (Friedman, Dunwoody, and Rogers 2010; Boykoff and Boykoff 2004; Clarke 2008).

2.3 Popularity Prediction

Predicting popularity of news and other content on social media such as Twitter or Reddit has been researched extensively (Ji He et al. 2016; Guerini, Strapparava, and Ozbal 2012; Wu et al. 2011; Althoff et al. 2013; Zhao et al. 2015; Tan, Lee, and Pang 2014). Hong, Dan, and Davison (2011) use textual features to predict the popularity of tweets as both a binary prediction problem (retweet or not) and a multiclass classification problem (class breakdown by tweet volume). Bandari, Asur, and Huberman (2012) approach the

problem of predicting the popularity of a news article on Twitter prior to its release, rather than after observing its initial reception. As features, they use the news source, category of the article (subject), subjectivity of article language, and named entities in the article. They predict the number of tweets using various regression and classification algorithms (binning the tweets into categories by count). They find that publication source is the most important feature. Tan, Friggeri, and Adamic (2016) track the flow of information from press releases to news articles to shares and comments on Facebook. They consider press releases of four types: political, technological, finance, and science (limited to MIT, Stanford and Berkeley), and track their coverage through 1800 news outlets. Controlling for the popularity of the news outlet, they predict news article shares on Facebook, using a variety of features, including subjectivity, positivity, and unigram coverage of the press release source.

There has also been research predicting the popularity specifically of scientific content in both the news and academia. Wallace, Paul, and Elhadad (2015) use logistic regression trained on textual and metadata features to identify attributes of health science articles which correlate with issuance of a press release and media coverage. They use two data sets: approximately 1,300 scientific articles with news coverage in Reuters and 27,000 matched sample (same journal and year) articles with no coverage, and approximately 800 scientific articles published in JAMA with press releases along with 10,000 negative samples. Zhang et al. (2016) build on this work, using supervised LDA and augmenting Sumner's dataset of 462 articles, press releases and news stories to this analysis (Sumner et al. 2014). They find various textual features, such as "95% CI" and "drinking" to be predicative of both press release issuance and news coverage, indicating that scientific journals tend to disseminate press releases for articles whose content is likely to be newsworthy.

Guerini, Pepe, and Lepri (2012) analyze popularity as indicated by an article's number of downloads, bookmarks, and citations. They attempt to model non-topical features of the abstract, such as readability, percentage of pronouns and percentage of future tense verbs. Their dataset consists of articles from the fields of physics and astronomy.

Yogatama et al. (2011) analyze popularity of papers in two fields, economics and computational linguistics, also using downloads and citations as a metric of popularity. They find that textual features improve prediction accuracy over just the metadata features of author name, subject and conference venue.

3 Dataset

Altmetric tracks a manually-curated list of over 2000 RSS feeds from news websites in various languages.[2] Given a recently crawled news article, they detect scientific coverage and match it with the original journal article using two methods: 1) search for links to content published in journals or on other academic platforms 2) extract potential journal names, article titles and author names, perform a search on CrossRef in a time window of the 45 days before and after the news report's publication, and link to journal article if CrossRef returns a match. The second, information extraction, method is only applied to English news articles.

Altmetric has provided us with the DOIs of every (tracked) scientific article published in 2016 which received (tracked) news coverage in 2016. This totals 91,997 scientific articles. Querying their database for these DOIs, we are able to retrieve the metadata on each article.[3] This includes the following bibliographic information (note, Altmetric does not have complete bibliographic information for all scientific articles): title, abstract, journal, publisher, subjects. Additionally, we use the CrossRef API to retrieve accurate dates (date on which the publisher deposited metadata) and more subjects for each scientific article.[4] Counts for each feature type are as follows (out of 91,997 possible): Title: 91,990, Abstract: 65,748, Journal: 91,438, Publisher: 52,417, Subjects (at least one type): 86,203.

Altmetric's API also provides the names of and links to each associated news outlet for a given scientific article. There are 640,610 news articles across 2,057 outlets referencing one or more of the scientific articles. In order to find and download press release(s) associated with each article, we search for two prominent press release aggregation/publication websites, EurekAlert! and Science Daily, in Altmetric's list of relevant news articles for each scientific articles. EurekAlert! is a global news service operated by the AAAS. It is a prominent source for journalists to find press releases from universities, medical centers, journals, government agencies, corporations and other organizations. They publish press releases from all areas of science, medicine, and technology. We find and download press release(s) for 18,287 scientific articles. Science Daily, another prominent press release aggregation website, similarly publishes selected press releases submitted by universities and research organizations. We are unable to download any articles from their website, however, as they block scraping.

3.1 What is Real News Coverage?

The presence of EurekAlert and Science Daily in the list of news source indicates that Altmetric does not track solely mainstream news websites, but also press release publication and aggregation websites. Through manual inspection of outlet publication frequencies and content, we discover that a number of news outlets tracked by Altmetric are not real news outlets which create original content. Specifically, we find that a significant percentage of the coverage for a large number of scientific articles is by outlets which simply copy abstracts or press releases from EurekAlert! or are themselves sources of press releases (such as the MIT press office).

3.2 Copy Detection

Since we aimed to predict "real" coverage of new articles where a journalist had taken the time to craft a story, we use two methods to detect outlets which do not produce real

[2] https://www.altmetric.com/about-our-data/our-sources/news/

[3] https://api.altmetric.com/docs/call_fetch.html

[4] https://github.com/CrossRef/rest-api-doc

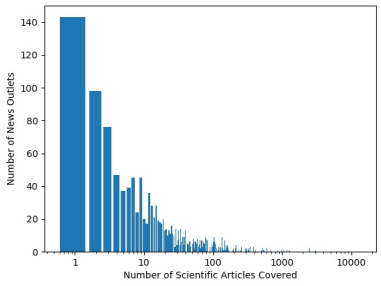

Figure 1: Outlet publication frequency in training set, the first 39 Weeks (276 days) of 2016.

content, finding and removing those which exactly copy abstracts or press releases. One method is based on n-gram overlap, the other on publication frequency. We are able to successfully crawl 474,734 of the related news articles from 1,860 outlets. We then extract the article content using Readability.[5] We lowercase then transform each abstract, press release (from EurekAlert!) and news article into a vector of counts of five-grams. For each article published by each news outlet, we compute its cosine similarity to its parent abstract and press release(s). We identify an outlet as a copier if at least 25% of its news articles have cosine similarity greater than 0.7 to one of the potential sources. This threshold was selected through manual inspection of outlet similarities on our training set (the first 39 weeks, 276 days, of 2016) to be robust to noise from imperfect removal of ads and marginalia in the HTML and outlet-specific boilerplate text. This results in a list of 136 outlets, a mix of press release, university, and special topic websites such as PR Web, MIT News and Seed Daily. Most outlets copy press releases, while a few copy abstracts.

This overlap method, however, does not capture all copiers. Certain science-roundup websites, such as Physician's Briefing, post weekly or monthly review pages which are simply copies of all relevant press releases. The presence of all the extra text lowers the cosine similarity with the original source below our target threshold. Other websites often post only article meta-data. Lastly, nearly two hundred websites, such as Science Daily, either block crawling or removed the news story before we are able to crawl it. In order to identify potentially copying outlets among these, we examine frequency of publication during our training set.

Our intuition is that outlets publishing content on a very large number of articles each day are most likely not generating new content, but simply scraping EurekAlert!, the journal, or some other parent source then posting the content on their website according to a simple rule such as journal name or article subject. Through manual inspection of source frequencies and content, we identify covering 10 scientific articles per day as a reasonable threshold to distinguish real news outlets from copiers. Figure 1 displays outlet publication frequency in the training set. 972 outlets cover fewer than 1 article per week, 1031 between one a week and 10 a day, and 20 outlets more than 10 per day. EurekAlert! is the highest occurring outlet in the training set, covering 13,925 scientific articles. There are 34 news sources in test set not present in the training set. The highest frequency mainstream news outlet in our training set is the Daily Mail, publishing stories referencing approximately 9 scientific articles per day. The Huffington Post is in second at just under 7 per day. This method detects 20 super high frequency outlets, 10 of which are already identified by the five-gram cosine similarity method. Thus we identify a total of 146 copiers and remove these copiers from the list of news coverage for each scientific article.

[5]https://github.com/buriy/python-readability

4 Predicting Scientific Article Popularity

We approach the problem of predicting the popularity of a scientific article as a ranking task. Given all of the articles published on a given day or in a given week, we aim to learn to rank them by a relevance metric corresponding to the number of real news articles. We treat each week or day as a query, then grade each article's relevance according to its real news coverage.

We calculate relevance as follows: for a given period of time, take threshold 1, T_1, as the amount of news coverage of scientific article in the 50th percentile + 1 and threshold 2, T_2, as the amount of news coverage of the scientific article in the 90th percentile + 1. We label an article's relevance, R, as follows:

$$R_0 : 0 \leq c < T_1$$
$$R_1 : T_1 \leq c < T_2$$
$$R_2 : T_2 \leq c$$

where c is the number of news articles published on a given scientific article. At the granularity of a day, on average, the relevance thresholds are $T_1 = 1.93$, $T_2 = 10.46$. At a week, the average thresholds are similar, at $T_1 = 1.98$ and $T_2 = 10.73$. Although coverage varies significantly day-to-day, where a scientific article with 20 news articles could be labeled as relevance 1 or 2 depending on the competing articles that day, at the week level, the thresholds are relatively uniform.

Our intuition for determining relevance is as follows: say a journalist were to be presented with a list of the scientific articles published on a given day or in a given week for which Altmetric had tracked some amount of real coverage. In order to select which articles to write on, the journalist would need to read some number of press releases and abstracts. Reading or skimming all scientific articles covered by Altmetric, at approximately 250 articles per day, is unfeasible. Additionally, half of those articles are relevance 0, covered by 1 or no real news outlets, and thus are of questionable importance. Sorting by predicted relevance allows journalists to read only a small subset of articles a day, knowing these articles are predicted to be widely popular and interesting to a broad audience. Although journalists may be interested exclusively in relevance 2 scientific articles, we include relevance 1 (instead of a binary relevance

Journal	Avg. Coverage
JAMA Internal Medicine	21.2
JAMA: Journal of the American Medical Association	16.9
Pediatrics	16.5
Current Biology	16.1
JAMA Pediatrics	15.0
MMWR: Morbidity & Mortality Weekly Report	14.9
Circulation	14.7
New England Journal of Medicine	14.4
Nature Geoscience	13.0
Nature Climate Change	13.0

Table 1: Top 10 journals by average amount of real news coverage. Only journals with at least 100 articles in the dataset included in analysis.

Subject	Avg. Coverage
health information management	15.4
health, toxicology and mutagenesis	10.7
pediatrics, perinatology, and child health	10.5
environmental science, miscellaneous	10.1
agricultural and biological sciences, miscellaneous	9.7
internal medicine	9.5
general earth planetary sciences	8.9
epidemiology	8.8
archaeology	8.7
social sciences, miscellaneous	8.4

Table 2: Top 10 CrossRef subjects by average amount of real news coverage. Only subjects describing at least 100 articles in the dataset included in analysis.

task), to indicate that a mis-rank of a relevance 2 below a relevance 0 is worse than below a relevance 1.

We use the first 39 weeks (276 days) of 2016 as our training set and evaluate on weeks 40-52 (days 277-365). There are 72,540 scientific articles in the training set. Their coverage ranges from 0 to 368 news mentions, with the scientific articles in the 50th and 90th percentiles receiving 1 and 9 news articles of coverage, respectively. There are 19,457 scientific articles in the test set. Their coverage ranges from 0 to 303, with the scientific articles in the 50th and 90th percentiles receiving 1 and 11 news articles of coverage, respectively.

4.1 Model

We use the learning-to-rank algorithm, lambdaMART (Burges 2010). LambdaMART directly optimizes ranking quality measures such as Normalized Discounted Cumulative Gain (NDCG) and Mean Average Precision (MAP) using gradient boosted decision trees. We use Microsoft's LightGBM to train the model, training it to optimize NDCG. For the daily problem, we optimize and evaluate on $NDCG@10$ since it is a usual search engine metric and we found 10 to be the upper limit on the number of scientific articles a real news outlet would cover in a day. For the weekly problem, we optimize and evaluate on a range of ranks 7, 35, and 70, corresponding to outlets which publish on 1, 5, or 10 scientific articles each day. We use DART (Dropouts meet Multiple Additive Regression Trees) as the boosting type, as previous research has shown that DART overcomes MART's issue of over-specialization to a considerable extent and improves performance on ranking tasks (Rashmi and Gilad-Bachrach 2015).

4.2 Features

We extract and treat as binary features the earlier mentioned bibliographic features: **Journal**, **Publisher**. Table 1 displays the top 10 journals with at least 100 articles by average news coverage (all journals, however, are included as features). As expected, top journals such as JAMA, Pediatrics, New England Journal of Medicine, and Nature, make up most of the top 10.

We also extract and treat as binary features **four different types of subjects** from Altmetric and CrossRef: Medline subject codes for journal (Altmetric), subjects as indexed by SCOPUS (Altmetric), publisher subjects (Altmetric), and CrossRef subjects. Subjects are mostly at the granularity of the journal or higher. Similar journals have subjects in common, with, for instance, the journals Nature Conservation, Conservation Biology, Ecology and Evolution, and many more sharing the CrossRef subject "nature and landscape conservation." Multidisciplinary journals, such as PNAS and Nature, are often tagged with more general subjects such as "science" or "multidisciplinary." Table 2 displays the top 10 CrossRef subjects describing at least 100 articles by average news coverage. As expected, health related subjects top the list, along with those related to the environment. Unfortunately, but as expected, "general computer science" and "computational mathematics" secured the last places, at 0.5 and 0.3 news articles on average, respectively.

We also extract textual features, when available, for each document: **Title**, **Abstract**, **Press Release**. For scientific articles with more than one press release, we randomly select one. For each type of text document, we preprocess using a standard English stop word list, extracting counts of unigrams and bigrams, keeping tokens which appear in more than one document and no more than 80% of documents of the same type in the training set. We keep the 30,000 most frequently occurring features in the training set. We then scale the vectors of counts by tfidf weights fit on the training set.

Finally, we create a boolean feature **hasPR** indicating whether the scientific article received coverage on one of the two prominent press release aggregation websites, EurekAlert! or Science Daily. We hypothesize that mention on one of these would expose the scientific article to more journalists and thus garner more coverage.

		Day	Week		
	Features	NDCG@10	NDCG@7	NDCG@35	NDCG@70
All Articles	Metadata (excluding hasPR)	*0.4114*	*0.4513*	*0.4460*	*0.4068*
	Metadata	*0.5485*	*0.6758*	*0.6148*	*0.5740*
	Metadata + Title	*0.6112*	*0.7615*	*0.6778*	*0.6097*
	Metadata + Abstract	*0.6116*	0.7935	0.6920	*0.6284*
	Metadata + Press Release	*0.5957*	*0.7563*	*0.6943*	*0.6344*
	Metadata + Title + Abstract	*0.6114*	0.8236	0.6944	*0.6280*
	Metadata + Title + Press Release	0.6209	*0.8107*	0.7282	0.6472
	Title + Abstract + Press Release	0.6273	**0.8868**	*0.7076*	0.6416
	All Features	**0.6354**	0.8841	**0.7313**	**0.6621**
Conditioned on Press Release	Metadata	*0.5101*	*0.4775*	*0.4433*	*0.5307*
	Metadata + Title	*0.5362*	*0.4782*	*0.4584*	*0.5259*
	Metadata + Abstract	*0.5611*	0.5629	*0.5085*	*0.5721*
	Metadata + Press Release	0.5841	0.6015	0.5315	0.6009
	Metadata + Title + Abstract	*0.5657*	0.6020	*0.4936*	*0.5597*
	Metadata + Title + Press Release	0.5888	0.6036	0.5350	0.6050
	Title + Abstract + Press Release	0.5902	*0.5526*	0.5251	*0.5904*
	All Features	**0.6065**	**0.6156**	**0.5559**	**0.6106**

Table 3: Daily and weekly popularity prediction: results on all articles then conditioning on press release issuance. Metadata features include hasPR, subjects, journal and publisher. After conditioning on press release issuance, there are some days on which fewer than 10 scientific articles were published - for those days NDCG is calculated up to that number of scientific articles. Best result for each column is in bold. Models in each column that the respective best performing model significantly outperforms are italicized ($p < 0.05$, permutation test).

5 Results

We present results on three tasks. The first two tasks are ranking ones. In task one, we attempt to learn to rank lists of scientific articles grouped by day. In task two, we perform the same task, but on lists of scientific articles grouped by week. For task three, we consider the binary task of differentiating between scientific articles with real coverage and those with coverage only from press release issuers and copiers.

5.1 Daily Prediction

See Table 3 for results for this experiment. In line with previous research, textual bag-of-words features provide baseline performance higher than just metadata features. The model using all features performs the best, but it does not statistically significantly outperform models using all textual features or metadata, title and press release features.

Table 4 lists the top features for the best performing feature set, all features. Due to the nonlinear nature of gradient boosted trees, we are unable to exactly determine the polarity of different features, only their importances, given by the number of times on which they were split. We approximate feature polarity by the difference in average relevance of documents with and without each feature.

To note, interpretation of these features, especially single features, is highly conjectural. Regardless, we provide some discussion and speculation. As expected, coverage on EurekAlert! or Science Daily is very predictive, as are generally popular and unpopular subjects such as medicine (popular) and chemistry (unpopular). As can be seen in table 4,

# Split	Polarity	Feature	Source
100	0.57	hasPR	hasPR
88	1.31	mesothelioma	abstract
85	0.44	patients	press release
83	-0.21	life sciences	scopus subj
70	-0.37	alzheimer	abstract
61	0.16	years	abstract
54	0.65	health information management	crossref subj
48	0.82	Massachusetts Medical Society	publisher
46	0.18	95	abstract
42	0.31	The Royal Society	publisher
41	0.62	people	press release
39	0.59	years	press release
38	0.65	zika virus	title
35	0.20	science	journal subj
33	0.00	agricultural and biological sciences	scopus subj
32	0.61	Science Advances	journal
31	0.19	medicine	journal subj
31	-0.20	chemistry	scopus subj
30	0.14	Elsevier	publisher
30	0.13	evidence	abstract

Table 4: Daily popularity prediction. Features: All features. 20 most important features. Polarity calculated by difference in average relevance of documents with and without each feature.

medically related text and metadata features, such as "patients," "alzheimer," "health information management," and "medicine" are also important and of mostly positive polarity, since medical articles attract greater news coverage than, say, a recently published computer science paper. As we discuss later, the presence of "alzheimer" as an important, negative feature is more an artifact of the dataset than an identification of trends in popularity. We also find an artifact of year (2016) of the dataset with the feature "zika virus," due to the 2015-2016 zika epidemic in North and South America. Three publishers appear in the top 20, the Massachusetts Medical Society, the publisher of the New England Journal of Medicine, one of the most prestigious peer-reviewed medical journals, "The Royal Society," publisher of the less prestigious, but longest-running scientific journal, "Philosophical Transactions of the Royal Society," and "Elsevier," one of the largest science publishers.

5.2 Weekly Prediction

See Table 3 for results for this experiment. Relative performance on the weekly task is similar to its daily counterpart. Understandably, performance is best when optimized for and evaluated on smaller ranks for $NDCG$ since weekly scientific article lists are long and there are more super popular, possibly easier to rank, articles which can fill the top 7 or 35. Segmenting dates by week rather than day may be more realistic, since scientific articles compete for news coverage with articles published around the same time rather than just those published on the same day.

At ranks 35 and 70, as in the daily prediction task, the models using all features perform the best, but not statistically significantly better than models trained on some feature subsets. At rank 7, the model trained on textual features outperforms the model with all features, but not statistically significantly.

Table 5 gives the top features for the model trained to optimize $NDCG@7$ and using textual features. Again, medical and people-related terms, such as "diet," "participants," and "health" are quite important and of positive polarity. Interestingly, "protein" in the press release is of positive polarity on average, whereas it is negative when present in the abstract. We hypothesize that this may be because, on average, very technical scientific papers, such as those mentioning proteins and cells, are less accessible to journalists, but those which garner a press release are much more likely to be read and covered. Similar to Zhang et al. (2016) "95" (for 95% CI) is an important, positive feature, perhaps since this is commonly found in results of medical studies that may be of interest to the general public.

5.3 Condition on Press Release Issuance

Since having a press release on EurekAlert! or Science Daily is, understandably, correlated with increased news coverage, we repeat the same experiments on the subset of scientific articles with press releases posted on EurekAlert! or Science Daily. This subset consists of 20,546 scientific articles. Thresholds between relevances are much higher on this subset, with thresholds at the daily level of $T_1 = 3.64$ and $T_2 = 30.36$, and $T_1 = 3.53$ and $T_1 = 33.04$ at the weekly level, on average.

# Split	Polarity	Feature	Source
103	0.60	study	press release
94	1.29	mesothelioma	abstract
86	0.18	95	abstract
75	-0.35	alzheimer	abstract
62	0.46	patients	press release
60	0.58	use	press release
45	0.24	protein	press release
44	0.21	health	abstract
41	-0.14	cell	abstract
40	0.13	evidence	abstract
39	-0.20	protein	abstract
39	0.82	diet	press release
38	0.61	years	press release
38	0.64	people	press release
38	0.20	participants	abstract
33	0.34	issue	press release
31	0.24	world	abstract
31	0.59	evidence	press release
28	0.16	body	abstract
28	0.16	years	abstract

Table 5: Weekly popularity prediction. Optimizing $NDCD@7$. Features: Title, abstract, press release. 20 most important features. Polarity calculated by difference in average relevance of documents with and without each feature.

As seen in Table 3, this problem is harder, with the best performance on the daily task at 0.6065, using all features, and 0.6156 on the weekly task at rank 7, using all features. However, although harder, this problem may be more realistic since, as noted in the introduction, many journalists start with press materials to inform what science they cover before reading original articles. Table 6 displays the top 20 features for the daily experiment using all features. Important features for these experiments are similar to those listed for the experiments including all scientific articles, except features from the press release are more prevalent. This makes sense as press releases were issued for all articles in this subset (and we have the text of most of them), and press releases are an important source text for journalists. We do not list features from the weekly experiments as they are quite similar.

5.4 Coverage vs No Coverage

In the previous experiments, we treat articles with news coverage from only automatic copying of abstracts or press releases as having coverage count 0. In this section, we also explore the binary prediction problem of differentiating between those 0 count articles and those with any amount of real news coverage. We create a matched sampling problem using the 34,329 scientific articles with no real news coverage as negative examples. For each negative example, we attempt to find a positive example: 1. with the same subject fields, 2. published in the same journal, 3. published within

# Split	Polarity	Feature	Source
55	0.17	people	press release
54	0.30	years	abstract
50	0.13	evidence	press release
39	0.34	Nature	journal
38	0.52	zika	title
33	0.14	planet	press release
33	0.40	American Medical Association	publisher
33	0.09	humans	press release
31	-0.01	patients	abstract
30	-0.22	protein	press release
30	0.33	95	abstract
29	-0.03	patients	press release
28	0.23	published today	press release
27	0.12	study	press release
27	0.13	years	press release
26	0.19	percent	press release
25	0.29	consumption	press release
25	0.36	foods	press release
24	0.13	fossil	press release
23	-0.25	proteins	press release

Table 6: Conditioned on press release publication on EurekAlert! or Science Daily. Daily popularity prediction. Features: All Features. 20 most important features.

Weight	Feature	Source
0.8226	fish	press release
0.8236	fat	press release
0.8307	screening	abstract
0.8803	quantum	press release
0.9044	detection	title
0.9616	Project HOPE - The People-to-People Health Foundation, Inc.	publisher
0.9646	early	title
1.0459	American Chemical Society	publisher
1.1415	time	title
1.2791	Oxford University Press	publisher
-1.1910	protein	title
-1.0599	protein	press release
-0.9404	medicine	title
-0.9334	cell	title
-0.9287	mechanism	title
-0.8788	expression	title
-0.8500	expression	abstract
-0.8041	genes	press release
-0.8025	following	title
-0.7689	domain	title

Table 7: Conditioned on press release publication on EurekAlert! or Science Daily, matched pairs binary prediction problem, controlling for journal, subjects, and time. 10 most positive and most negative features.

7 days of the negative instance, 4. did/did not receive a press release, depending on the negative example. If no such positive example exists, the negative sample is not included.

We find 8,436 matched pairs. We train logistic regression with the same train-test split (days 1-277, days 278-365). Using all features, logistic regression achieves an F1 of 0.6037. Although many of the positive feature are sensible and similar to those found in the ranking experiments and previous research (Wallace, Paul, and Elhadad 2015; Zhang et al. 2016), we find that the negative features are dominated by alzheimers and dementia related terms in the title and abstract. These results are indicative of an oddity in the dataset provided by Altmetric, specifically the website Alzforum. Alzforum is an information resource website for alzheimers researchers and is one of the websites included in Altmetric's news crawl. Most of the Alzforum articles tracked by Altmetric post only article metadata, and the website was marked as a non-news source due to its super high rate of publication. Before sampling, Alzforum covers 8,347 of the 34,329 scientific articles with no real coverage and is the highest frequency outlet on that subset. However, it covers only 667 of the 57,668 scientific articles with real coverage. Of 8,436 matched pairs, Alzforum covers 1,380 negative and 224 positive samples. Although we control for subjects, they are typically more general, such as "neuroscience," "psychology and cognitive sciences," and "geriatrics and gerontology." Thus, we are unable to control for specific topics, such as alzheimers, and logistic regression learns that most alzheimers articles are unpopular.

5.5 Condition on Press Release Issuance

Similar to the ranking experiments, we repeat the matched sample experiment on the subset of articles which received a press release. We find 3,398 matched pairs and confirm that the positive and negative instances do not have disproportionate skew towards any one outlet. Using all features, logistic regression achieves an F1 of 0.5794. Table 7 lists the top features for these results. Top negative features appear to be those related to more technical scientific work, such as "proteins" and "genes" "expression", while positive features are related to health, such as "fat," "screening," and "Project HOPE," a publisher of health policy articles. Surprisingly, the American Chemical Society is a positive feature and "medicine" is a negative feature when occurring in the title.

6 Discussion

In this paper we present an analysis of news coverage of scientific articles across various disciplines, universities, journals, and news outlets. After filtering out press release and abstract issuance and copier websites, we approach the problem as a ranking one. For a given period of time, we learn to rank scientific articles by the amount of news coverage they receive. We find that textual features significantly improve the accuracy of the prediction over metadata features, with abstract and press release features providing the largest boost in accuracy. We find the most important feature to be whether the scientific article has a press release published

on EurekAlert! or Science Daily. Conditioning on receiving press release coverage on at least one of these websites, we repeat these experiments, finding significant performance drop on the daily and weekly tasks. Performance on this task may be more realistic, though, since press releases are a valuable resource to journalists when writing a piece, and many journalists use them as a starting point when crafting an article.

We also examine the problem of predicting whether a scientific article will garner coverage by just press release issuance and copier websites, or whether it will receive real news coverage. We set up a matched sampling problem, controlling for journal, subjects, time and press release publication. On the subset of articles with press releases, we find that health-related terms are positively correlated with coverage, while technical scientific terms, such as protein and cell, are negatively correlated. In other words, use of certain keywords, along with strategic production of press materials, may account for patterns of attention, as well as neglect, for scientific topics across society.

Overall, we provide insights that speak to current issues in science communication, such as the "cycle of hype" and systematic bias in media selection. These dynamics potentially influence a range of downstream policy issues, from research funding potential to public opinion and risk perceptions. Establishing evidence of patterns of subjective media attention furnishes important knowledge relating to the public communication of science.

7 Future work

There are many avenues for future work with this dataset. Using the text of newspaper articles, we could model the process of information diffusion from scientific articles through press releases to news articles. There is a rich literature in tracing information diffusion in mainstream and social media, studying how to identify and track phrases that spread through networks (Leskovec, Backstrom, and Kleinberg 2009; Daniel M. Romero, Meeder, and Kleinberg 2011; Simmons, Adamic, and Adar 2011; Tan, Friggeri, and Adamic 2016). Retrieving reliable dates for the newspaper articles, we could examine the effects of coverage among outlets. For instance, if a major outlet like the New York Times covers a scientific article, does that article receive a bump in coverage the following day or week? Using date data, we could also examine relative publication speed and timing of different outlets, discovering which outlets report on scientific findings the fastest or slowest and which outlets consistently follow or precede others. Without date information, networks of information propagation could be inferred through analysis of patterns of copying among outlets.

Expanding the dataset with scientific articles from other years would also enable further research. With data across multiple years, we could explore historical trends in the publication and reporting of scientific news. Furthermore, with additional years of data, author names might become a useful feature, similar to how Yogatama (Yogatama et al. 2011), using 10 years of economic research papers and 26 years of computational linguistics papers, finds that some authors are more prolific, producing more popular content than others.

References

Althoff, T.; Borth, D.; Hees, J.; and Dengel, A. 2013. Analysis and forecasting of trending topics in online media streams. In *Proceedings of the 21st ACM international conference on Multimedia Pages*.

Bandari, R.; Asur, S.; and Huberman, B. A. 2012. The pulse of news in social media: Forecasting popularity. In *Proceedings of ICWSM*.

Bornmann, L. 2014. Do altmetrics point to the broader impact of research? an overview of benefits and disadvantages of altmetrics. *Journal of Informetrics* 8:895–903.

Boykoff, M., and Boykoff, J. 2004. Balance as bias: Global warming and the us prestige press. *Global Environmental Change* 14:125–136.

Brechman, J.; Lee, C.; and Cappella, J. 2009. Lost in translation?: a comparison of cancer-genetics reporting in the press release and its subsequent coverage in the press. *Sci Commun* 30:453–474.

Bubela, T.; Nisbet, M. C.; Borchelt, R.; Brunger, F.; Critchley, C.; Einsiedel, E.; Geller, G.; Gupta, A.; Hampel, J.; Hyde-Lay, R.; Jandciu, E. W.; Jones, S. A.; Kolopack, P.; Lane, S.; Lougheed, T.; Nerlich, B.; Ogbogu, U.; O'Riordan, K.; Ouellette, C.; Spear, M.; Strauss, S.; Thavaratnam, T.; Willemse, L.; and Caulfield, T. 2009. Science communication reconsidered. *Nature Biotechnology*.

Burges, C. J. 2010. From ranknet to lambdarank to lambdamart: An overview. Technical report, Microsoft Research.

Caulfield, T.; Sipp, D.; Murry, C. E.; Daley, G. Q.; and Kimmelman, J. 2016. Confronting stem cell hype. *Science* 352:776–777.

Clarke, C. E. 2008. A question of balance: The autism-vaccine controversy in the british and american elite press. *Science Communication* 30:77–107.

Costas, R.; Zahedi, Z.; and Wouters, P. 2014. Do altmetrics correlate with citations? extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology* 66:2003–2019.

Daniel M. Romero, D. M.; Meeder, B.; and Kleinberg, J. 2011. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proceedings of WWW*.

Donsbach, W. 2004. Psychology of news decisions factors behind journalists professional behavior. *Journalism* 5:131–157.

Fahy, D., and Nisbet, M. C. 2011. The science journalist online: Shifting roles and emerging practices. *Journalism* 12:778793.

Friedman, S. M.; Dunwoody, S.; and Rogers, C. L. 2010. *Communicating uncertainty: media coverage of new and controversial science*. Routledge.

Geller, G.; Bernhardt, B. A.; and Holtzman, N. A. 2002. The media and public reaction to genetic research. *Journal of the American Medical Association* 287:773.

Guerini, M.; Pepe, A.; and Lepri, B. 2012. Do linguistic style and readability of scientific abstracts affect their virality? In *Proceedings of ICWSM*.

Guerini, M.; Strapparava, C.; and Ozbal, G. 2012. Exploring text virality in social networks. In *Proceedings of ICWSM*.

Harvard Kennedy School Shorenstein Center on Media Politics and Public Policy. 2016. Unpublished survey. https://repository.library.northeastern.edu/files/neu:cj82qp370.

Hong, L.; Dan, O.; and Davison, B. D. 2011. Predicting popular messages in twitter. In *Proceedings of WWW*.

Ji He, J.; Mari Ostendorf, M.; Xiaodong He, X.; Jianshu Chen, J.; Jianfeng Gao, J.; Lihong Li, L.; and Deng, L. 2016. Deep reinforcement learning with a combinatorial action space for predicting popular reddit threads. In *Conference on Empirical Methods in Natural Language Processing*.

Kiernan, V. 2003a. Diffusion of news about research. *Science Communication* 25:313.

Kiernan, V. 2003b. Embargoes and science news. *Journalism & Mass Communication Quarterly* 80:903920.

Len-Rios, M. E.; Hinnant, A.; Park, S.-A.; Cameron, G. T.; Frisby, C. M.; and Youngah, L. 2009. Health news agenda building: Journalists' perceptions of the role of public relations. *Journalism & Mass Communication Quarterly* 86.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of KDD*.

National Academies of Sciences, Engineering, and Medicine. 2017. *Communicating Science Effectively: A Research Agenda*. The National Academies Press.

Patterson, T. E. 2013. *Informing the news: the need for knowledge-based journalism*. NY: Vintage Books.

Piwowar, H. 2013. Altmetrics: Value all research products. *Nature* 493.

Priem, J.; Taraborelli, D.; Groth, P.; and Neylon, C. 2010. Altmetrics: A manifesto.

Priem, J.; Groth, P.; and Taraborelli, D. 2012. The altmetrics collection. *PLoS ONE* 7.

Rashmi, K., and Gilad-Bachrach, R. 2015. Dart: Dropouts meet multiple additive regression trees. In *Eighteenth International Conference on Artificial Intelligence and Statistics*, 489–497.

Russell, C. 2006. Covering controversial science: Improving reporting on science and public policy. Harvard Kennedy School, Working Paper Series.

Simmons, M. P.; Adamic, L. A.; and Adar, E. 2011. Memes online: Extracted, subtracted, injected, and recollected. In *Proceedings of ICWSM*.

Suleski, J., and Ibaraki, M. 2009. Scientists are talking, but mostly to each other: a quantitative analysis of research represented in mass media. *Public Understanding of Science* 19:115125.

Sumner, P.; Vivian-Griffiths, S.; Boivin, J.; Williams, A.; Venetis, C.; Davies, A.; Ogden, J.; Whelan, L.; Hughes, B.; Boy, F.; and et al. 2014. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *BMJ* 349.

Tan, C.; Friggeri, A.; and Adamic, L. 2016. Lost in propagation? unfolding news cycles from the source. In *Proceedings of ICWSM*.

Tan, C.; Lee, L.; and Pang, B. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. In *Proceedings of ACL*.

Tanner, A. 2004. Agenda building, source selection, and health news at local television stations. *Science Communication* 350–363.

Thelwall, M.; Haustein, S.; Larivire, V.; and Sugimoto, C. R. 2013. Do altmetrics work? twitter and ten other social web services. *PLoS ONE* 8(5).

Viswanath, K.; Blake, K.; Meissner, H.; Saiontz, N.; Mull, C.; Freeman, C.; Hesse, B.; and Croyle, R. 2008. Occupational practices and the making of health news: A national survey of u.s. health and medical science journalists. *Journal of Health Communication* 13:759–777.

Wallace, B. C.; Paul, M. J.; and Elhadad, N. 2015. What predicts media coverage of health science articles. In *The International Workshop on the World Wide Web and Public Health Intelligence (W3PHI)*.

Wihbey, J., and Coddington, M. 2017. Knowing the numbers: Assessing attitudes among journalists and educators about using and interpreting data, statistics, and research. In *ISOJ International Symposium of Online Journalism*.

Woloshin, S., and Schwartz, L. M. 2002. Translating research into news. *JAMA* 287:2856–2858.

Wu, S.; Tan, C.; Kleinberg, J.; and Macy, M. 2011. Does bad news go away faster? In *Proceedings of ICWSM (short paper)*.

Yogatama, D.; Heilman, M.; O'Connor, B.; Dyer, C.; Routledge, B. R.; and Smith, N. A. 2011. Predicting a scientific communitys response to an article. In *Conference on Empirical Methods in Natural Language Processing*.

Zhang, Y.; Willis, E.; Paul, M. J.; Elhadad, N.; and Wallace, B. C. 2016. Characterizing the (perceived) newsworthiness of health science articles: A data-driven approach. *JMIR Medical Informatics* 4.

Zhao, Q.; Erdogdu, M. A.; He, H. Y.; Rajaraman, A.; and Leskovec, J. 2015. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of KDD*.