What Predicts Media Coverage of Health Science Articles?

Byron C Wallace
School of Information
University of Texas at Austin
byron.wallace@utexas.edu

Michael J Paul
Dept. of Computer Science
Johns Hopkins University
mpaul@cs.jhu.edu

Noémie Elhadad
Dept. of Biomedical Informatics
Columbia University
noemie.elhadad@columbia.edu

Abstract

An important aspect of health science is communicating research findings to the public. The media is a critical instrument in disseminating research. Yet the process by which a scientific article becomes "newsworthy" is not well understood. In this study, we use large-scale text analysis to characterize the content features of articles that are predictive of newsworthiness. We experiment with two novel corpora: (i) 28,910 articles from a diverse range of biomedical and health journals, of which 1,343 were covered by the news agency *Reuters*, and (ii) 10,760 articles from the *JAMA* journals, of which 846 were given press releases by the journal editors. We show that media coverage can be predicted reasonably well: logistic regression achieves mean AUCs of 0.783 and 0.882 on the Reuters and JAMA datasets, respectively. We present and discuss interesting findings concerning the most predictive content features.

Introduction

Public understanding of emerging health science requires timely and accurate reporting of new findings. Journalists play a critical role in disseminating biomedical findings to the public. Media coverage of health science has been studied from several viewpoints, from the impact of media coverage on individual and population health behaviors (Walsh-Childers and Brown 2009), health service utilisation (Grilli, Ramsay, and Minozzi 2002; Evans et al. 2014), and scholarly influence (Kiernan 2003) to the use of media as a health communication instrument (Wallack 1990), the ethical implications for researchers and journalists (Snyder, Mayes, and Spencer 2009), and the issues in media reporting research accurately and responsibly (Klaidman 1991; Shuchman and Wilkes 1997; Yavchitz et al. 2012).

Given the overwhelmingly large volume of scientific articles published every day, media outlets are constrained to select only a handful of "newsworthy" articles for coverage. This selection process is thus inherently biased. Prior studies have investigated specific factors that drive media coverage of health research, including press releases issued by scientific journals (Woloshin and Schwartz 2002), engagement of individual researchers with the media (Tsfati, Cohen, and Gunther 2010), and external events such as public figure disclosures of a health condition and disease awareness months (Konfortion, Jack, and Davies 2014). But the general selection process of a health scientific discovery for news coverage remains largely opaque. Similarly, factors explaining which scientific articles are selected for press releases by the scientific journals themselves are not well understood. Identifying these factors can help illuminate the biases inherent to news coverage of health research.

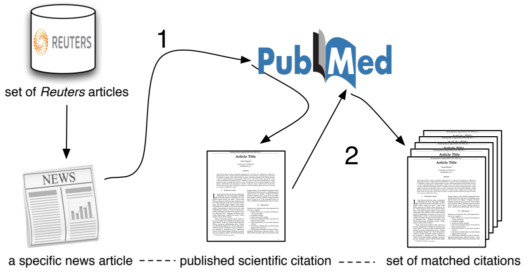

Figure 1: A schematic depicting corpus construction for the media coverage prediction task. For each Reuters news story, we: **1** retrieve from PubMed the corresponding scientific article (i.e., positive instance), and, **2** identify a *set* of similar articles (published in the same journal in the same year) that were *not* picked up by Reuters (i.e., negative instances).

In this work, we explore what predicts media coverage of health scientific articles. We aim to answer the following questions: (i) is it possible to predict whether a scientific article is likely to be picked up for a press release and/or for media coverage? and (ii) which article's features are associated with being picked up? Using two novel datasets, we show that information about an article, such as MeSH headings, and content from the title and abstract of articles have predictive power for both prediction tasks, and we identify several factors suggestive of coverage.

Methods

We conducted two sets of independent experiments: (i) given a corpus of scientific articles from several journals, predict which article(s) will be covered by a news agency; and (ii) focusing on a high-impact journal, predict which articles will be given press release by the journal editors.

Datasets

We constructed two novel datasets: one for each classification task (media coverage and press release prediction). These corpora are publicly available at: `https://github.com/bwallace/w3phi-2015`. For our tasks we need a large dataset of articles to learn from along with labels indicating presence of a press release or media coverage. To construct this dataset, we relied on a large collection of Reuters[1] health news stories and a collection of press releases issued by the JAMA editors.

For both datasets, we were faced with the challenge of constructing a set of 'negative' instances, i.e., articles that might have garnered media coverage (or received a press release) but that did not. To this end, we constructed a 'matched' set of negative examples for each positive article, in the same spirit as the 'matched sampling' approach (Rosenbaum and Rubin 1985), in which we are attempting to isolate predictors that correlate with garnering media attention. We next describe the two datasets in more detail.

Reuters corpus The Reuters corpus comprises health news stories that report on particular biomedical and health research study, as published by Reuters news agency. In each story, Reuters journalists cite and link to the original scientific article on which the story reports. Thus the Reuters stories and their corresponding scientific articles provide us with positive instances for the media coverage prediction task. In practice, the reference to the original scientific article was resolved from the Reuters story to a unique Digital Object Identifier (DOI), which was then used to retrieve citation and content information in PubMed, the open repository of biomedical literature.[2]

The corpus of Reuters health news stories was downloaded via the news aggregator Factiva for the period of January 1st, 2012 to September 1st, 2014. It resulted in 1,343 pairs of news stories and corresponding scientific articles, i.e., positive instances in the media coverage prediction task.

Negative instances were collected using a 'matched sampling' approach that attempted to control for several factors. Specifically, for each positive article, we sampled another 20 articles published in the same journal in the same year that did not receive coverage in the Reuters corpus.[3] We used several filtering heuristics to include only full-length original research articles in the corpus. The aim of this matched sampling was to identify articles that were just as likely to have been covered by Reuters, but were not. We were left with 27,567 articles, representing our negative instances.

JAMA corpus For the press release prediction task, we focused on a single, high-impact journal JAMA (Journal of the American Medical Association). The JAMA corpus comprises 846 positive instances, defined as articles for which JAMA editors created a press release.[4]

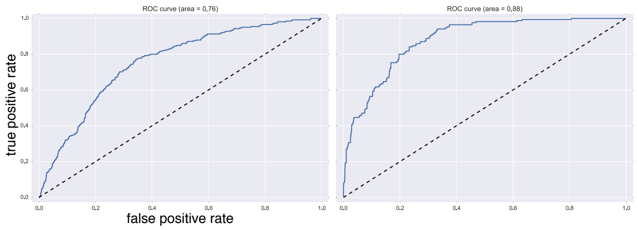

Figure 2: ROC curves illustrating classification performance achieved on the Reuters (left) and JAMA (right) datasets.

Like for the Reuters corpus, negative instances were constructed via matched sampling, focusing on articles from the same journal and year but for which no press release was issued. After removing duplicates, this corpus comprised 9,914 'negative' articles. This collection is exhaustive, containing all press releases available on the JAMA web archive (from October 1st, 2012 to October 1st, 2014).

Learning

For both prediction tasks we used standard logistic regression with a squared $\ell2$ norm penalty on the weights for regularization. We tuned the parameter encoding the tradeoff between regularization and predictive performance on the training dataset. When generating predictive features for inspection we used the entire available corpora; when assessing predictive performance we used 10-fold cross-validation.

We extracted citation features from both datasets, i.e., journal name, institution of first author, where we use email domains as a readily available proxy, and content features i.e., uni- and bi-grams extracted from titles, abstracts and MeSH terms.[5] We used a standard English stopword list. We kept tokens observed in more than 100 articles in the combined corpus. Ultimately this resulted in 14,614 unique features.

Results

Predictive Performance

Figure 2 shows the ROC curves for the two tasks. The predictive performance is reasonably good. We assess this via 10 fold cross validation. For the Reuters corpus, we achieve a mean AUC of 0.783, range: (0.746, 0.811). On the JAMA corpus, we observed a mean AUC of 0.882 (0.853, 0.918).

Predictive Content Features

We report the top fifty highest weighted positive and negative features for each model in Tables 2 and 3. The former set of features are predictive of mainstream media coverage of an article, while the latter are features predictive of an editor issuing a press release for an article. Note that none of the top 100 most predictive features for either of the corpora correspond to a journal indicator. This is likely due to our design: the journal is 'marginalized' out because for any

[1] `http://www.reuters.com`

[2] We used the NLM's API: `http://www.ncbi.nlm.nih.gov/books/NBK25501/`

[3] We eliminated duplicate instances of 'negative' articles.

[4] `http://media.jamanetwork.com/past-releases/`

[5] MeSH is the NLM's controlled vocabulary theasurus: `http://www.nlm.nih.gov/pubs/factsheets/mesh.html`.

term	*Reuters* - *JAMA* weight
weight	0.09
exercise	0.09
mh-adult	0.08
virus	0.07
mh-effects	0.06
influenza	0.06
mh-humans	0.06
mh-female mh-humans	0.06
mh-child	0.05
mh-aged	0.05
intake	-0.04
incident	-0.04
consumption	-0.03
mh-numerical mh-data	-0.03
mh-data	-0.03
smoking	-0.02
mh-numerical	-0.02
years	-0.01

Table 1: Features with the largest magnitude of difference (with respect to their normalized estimated coefficients) between the Reuters and JAMA datasets. We show the 10 features with larger coefficients in the Reuters compared to the JAMA model. We show only 7 features with estimated coefficients larger in the JAMA model, because all other (normalized) feature weights were smaller. We are not sure why this is the case.

given relevant article sampled from a specific journal, we sample an additional 20 from the same journal that are (intentionally) negative instances. Thus, journal indicators are balanced across negative and positive examples.

Any interpretation of these features is obviously speculative and we would caution against over-interpretation. But some interesting – if suggestive – trends are apparent. As per Table 2, articles reporting on 'exercise', 'intake', 'smoking', 'pregnancy', and 'cancer' all seem to be more likely to garner attention in the press. These are topics that affect large numbers of people, and may be of particular interest because of their association with personal behavior.

Interestingly, too, the MeSH terms 'mh-data' and 'mh-numerical' are highly predictive. These seem to correspond to articles conducting exploratory statistical analyses that report correlations. This effect is similarly visible in the JAMA corpus (Table 3). It may be the case that such (apparently) secondary analyses are generally more likely to receive a press release than primary studies. This would explain the consistently negative coefficients associated with words such as 'patients', 'clinical' and 'dosing'. We do not yet have an alternative interpretation of this observation. Another property shared by these two datasets is an apparent preference for results relevant to women: 'women' ranks very highly in both lists.

We were intrigued by the '000' token ranked highly in the Reuters list, so we inspected some examples. This seems to be capturing a specific style of reporting results where authors state odds in concrete numbers. For example: "Having more than 2 dermatologists per 100 000". This may simply correlate with the types of numerical analyses that tend to receive attention, or it may suggest that this editing style makes the article more attractive as a press piece (note that this is not a highly ranked feature in the JAMA corpus).

Finally, we note that one feature strikingly apparent from the JAMA coefficients is the importance of statistical significance: the single best predictor of a press release being issued for an article is the mention of a (95%) confidence interval (CI).

Conclusions

This paper presents our initial experiments with characterizing what makes a scientific article newsworthy. The primary contributions of the work are the construction of the two novel corpora, which enable us to study this question through two specific prediction tasks. There is much future work involved in further analyzing the predictive power of the identified content features and extending the prediction tasks to other health journals. More generally, this line of work presents a novel approach to characterizing the biases of media reporting to health science.

References

Evans, D.; Barwell, J.; Eccles, D.; Collins, A.; Izatt, L.; Jacobs, C.; Donaldson, A.; Brady, A.; Cuthbert, A.; and et al, R. H. 2014. The angelina jolie effect: how high celebrity profile can have a major impact on provision of cancer related services. *Breast Cancer Research* 16(5):442.

Grilli, R.; Ramsay, C.; and Minozzi, S. 2002. Mass media interventions: effects on health services utilisation. *Cochrane Database Syst Rev* CD000389.

Kiernan, V. 2003. Diffusion of news about research. *Science Communication* 25(1):3–13.

Klaidman, S. 1991. *Health in the headlines: The stories behind the stories*. Oxford University Press New York.

Konfortion, J.; Jack, R.; and Davies, E. 2014. Coverage of common cancer types in uk national newspapers: a content analysis. *BMJ open* 4(7):e004677.

Rosenbaum, P. R., and Rubin, D. B. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39(1):33–38.

Shuchman, M., and Wilkes, M. S. 1997. Medical scientists and health news reporting: a case of miscommunication. *Annals of Internal Medicine* 126(12):976–982.

Snyder, P.; Mayes, L.; and Spencer, D., eds. 2009. *Science and the Media: Delgado's Brave Bulls and the Ethics of Scientific Disclosure*. Academic Press.

Tsfati, Y.; Cohen, J.; and Gunther, A. C. 2010. The influence of presumed media influence on news about science and scientists. *Science Communication*.

Wallack, L. 1990. *Mass communication and public health*. SAGE. chapter Mass media and health promotion.

Walsh-Childers, K., and Brown, J. 2009. *Media Effects: Advances in Theory and Research*. Routledge. chapter Effect of Media on Personal and Public Health.

Woloshin, S., and Schwartz, L. 2002. Press releases: Translating research into news. *JAMA* 287(21):2856–2858.

Yavchitz, A.; Boutron, I.; Bafeta, A.; Marroun, I.; Charles, P.; Mantz, J.; and Ravaud, P. 2012. Misrepresentation of randomized controlled trials in press releases and news coverage: A cohort study. *PLoS Med* 9(9):e1001308.

negative		*positive*	
-1.102	patients	0.617	exercise
-0.507	clinical	0.610	mh-data
-0.457	2012	0.604	mh-numerical mh-data
-0.393	survival	0.586	mh-numerical
-0.364	therapy	0.536	intake
-0.337	complications	0.531	mh-adult
-0.323	surgical	0.508	cancer
-0.321	response	0.500	mh-effects
-0.308	plasma	0.492	years
-0.300	pediatric	0.461	mh-child
-0.285	diagnostic	0.459	virus
-0.281	imaging	0.455	mh-aged
-0.275	2013	0.443	smoking
-0.267	management	0.442	influenza
-0.265	expression	0.428	mh-female mh-humans
-0.258	factors	0.418	consumption
-0.252	outcomes	0.407	incident
-0.250	score	0.407	women
-0.248	range	0.407	weight
-0.246	treatment	0.391	mh-humans
-0.246	function	0.389	exposure
-0.245	diabetes	0.383	asd
-0.242	review	0.381	pregnancy
-0.236	os	0.380	year
-0.235	protein	0.378	mh-studies
-0.231	mice	0.372	mh-female
-0.231	serum	0.359	effect
-0.227	values	0.355	95
-0.223	model	0.354	age
-0.223	mm	0.349	mh-male
-0.222	shunt	0.347	physical
-0.217	care	0.346	injuries
-0.217	tumor	0.341	intervention
-0.214	safety	0.338	physical activity
-0.213	strategies	0.336	cardiovascular
-0.210	treated	0.336	reported
-0.204	activation	0.335	children
-0.203	role	0.330	000
-0.201	biopsy	0.316	mh-humans mh-male
-0.199	cell	0.315	mh-factors
-0.198	ti-in	0.314	trials
-0.196	hr	0.313	beverages
-0.196	growth	0.311	trend
-0.196	ventricular	0.298	fatigue
-0.195	correlated	0.298	cancers
-0.193	prognostic	0.289	mh-control
-0.192	relevance	0.289	2008
-0.192	lesions	0.287	ad
-0.190	insulin	0.282	men
-0.189	resection	0.279	cognitive

Table 2: Top fifty features and associated weights for *Reuters* corpus, ranked by magnitude. The ‘mh’ prefix indicates a MeSH term.

negative		*positive*	
-0.615	patients	0.852	ci
-0.473	clinical	0.838	95
-0.357	dosing	0.808	95 ci
-0.356	sbp	0.750	women
-0.349	evidence	0.476	cancer
-0.318	injury	0.447	increased
-0.312	ezetimibe	0.433	mh-numerical
-0.310	functional	0.430	breast
-0.304	management	0.429	years
-0.302	review	0.425	mh-data
-0.294	patient	0.410	vs
-0.293	handover	0.407	mh-numerical mh-data
-0.290	schizophrenia	0.404	prevalence
-0.287	resection	0.373	men
-0.286	information	0.366	states
-0.281	mechanical	0.341	pregnancy
-0.277	aortic	0.341	insurance
-0.276	days	0.336	person-years
-0.274	hospitalization	0.325	tobacco
-0.274	acupuncture	0.319	rates
-0.273	score	0.316	breast cancer
-0.257	scores	0.311	maternal
-0.252	faculty	0.306	costs
-0.251	relapse	0.305	health
-0.245	bacteremia	0.303	chd
-0.244	gastric	0.303	cvd
-0.242	studies	0.283	smoking
-0.242	hcv	0.277	drinking
-0.239	continuity	0.266	child
-0.238	brain	0.262	age
-0.235	severity	0.261	main outcome
-0.235	treatment	0.260	intake
-0.235	pci	0.259	medicaid
-0.232	weight loss	0.258	associated
-0.222	engagement	0.258	hr
-0.219	surgical	0.256	associated increased
-0.217	mm	0.251	united
-0.215	veterans	0.241	black
-0.212	outcomes	0.240	copd
-0.208	15-year	0.240	spending
-0.206	warfarin	0.231	mh-health
-0.201	group	0.231	united states
-0.201	search	0.227	exposure
-0.201	preventable	0.227	hearing
-0.200	areas	0.225	mh-risk
-0.194	plasma	0.225	mh-factors
-0.193	health information	0.224	services
-0.192	connectivity	0.223	likely
-0.192	genetic	0.222	increased risk
-0.192	120	0.221	association

Table 3: Top fifty features and associated weights for the *JAMA* corpus, ranked by magnitude.