Surfacing Newsworthy Public Documents As Leads

Alexander Spangher[a], Emilio Ferrara[a], Ben Welsh[b], Nanyun Peng[c], Serdar Tumgoren[d], Jon May[a]
[a] University of Southern California [b] Reuters News, [c] University of California, Los Angeles, [d] Stanford University

ABSTRACT

Newsworthiness prediction means trying to assess whether a particular story lead will get covered or not. In this work, we build *newsworthiness predictors* to find stories in voluminous city council meetings, aiming to reduce the time and effort it takes journalists to find stories. Building training datasets for this task is challenging: it is hard, ex post facto, to prove that a certain policy *has* or *has not* covered. We address this problem by implementing a novel *probabilistic relational modeling* framework, which we show is a low-annotation methodology that outperforms other, more state-of-the-art baselines. We scale this linking methodology across 13k city council policies from San Francisco Board of Supervisor meetings and 200k articles from the *San Francisco Chronicle* over 10 years of public policy meetings, finding about 7% of policies get covered. Finally, we used this linked dataset to fine-tune language models to consume policy text, transcribed video, public discussion, and other features, and predict the likelihood of coverage. We perform human evaluation with expert journalists and show our systems identify newsworthy policies with 68% F1-score and our coverage recommendations are rated as helpful, with an 84% win-rate against baseline.[1]

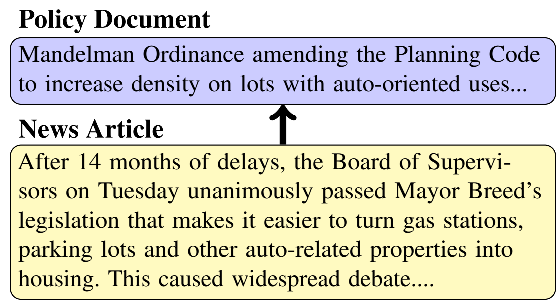

Figure 1: In this paper, we treat *newsworthiness prediction* as a supervised learning task. We (1) train models to infer when public policy items have been covered in the news before and (2) use these to predict if new items *will* be covered.

1 INTRODUCTION

We wish to build newsworthiness predictors to help journalists find stories. Prior work has relied on hand-encoded features, like salience or surprisal, to surface interesting stories [2]. In this work, we want to avoid relying on such features. *News values* vary widely between local [5] and national outlets [3], subject matter specific outlets, etc. With enough training data, a model should *learn* salient features in text that define newsworthiness, itself. So, ideally, we want to train supervised models on a dataset of story leads that are labeled as *newsworthy* or *not newsworthy*.

It is infeasible to collect such data via hand-labeling. So instead, we seek to build a training dataset on a proxy: what *has* and *has not* been covered in the past. Such an approach is conceptually simple and easily generalizable. We can, theoretically, take any news outlet and observe their news values by simply observing the story leads they cover and the ones they do not. We implicitly assume that what *has* been covered in the past is a good predictor of what *will* be covered in the future. For most local beat coverage, this is a relatively safe assumption [5], and we will test that further in this work.

However, determining that a story leads have covered in media, as shown in Figure 1, is a surprisingly challenging task. Unlike related tasks, like *citation prediction* [13], determining policy coverage requires us to establish links between documents in two different linguistic domains, with no pre-existing labels. To illustrate, suppose as in Figure 1, we wish to determine whether a specific SFChron article covers a specific SFBOS policy. Not all SFChron articles are about the SFBOS, and not all SFBOS items are legislative policies (e.g. some are administrative nominees, appointments, etc.). Such variety confounds unsupervised linking models. This motivates us to use a *probabilistic relational model* (PRM) framework [4] to solve this problem. PRM models break down link prediction into smaller, easier-to-supervise subproblems. Learning a PRM helps us outperform other state-of-the-art embeddings-based baselines. We use PRM to create a large linked corpus: 10 years of SFBOS policy items and SFChron articles covering them.

Next, we seek to use this linked dataset to predict if a *new* policy will get covered. We fine-tune language models consume 13k policy items, 3,200 hours of transcribed meetings and thousands of public comment sessions to make newsworthiness predictions. We find that policy items that get covered in news media get discussed slightly longer in meetings and have more members of the public addressing them during public comment periods. However, we find that incorporating these discussions into our predictive models barely yields any performance improvement, indicating that most characteristics that actually *makes* a policy newsworthy might not get discussed during meetings. Finally, our models are helpful to journalists, beating baseline 84% of the time and surfacing relevant items.

2 POLICY ITEM ↔ ARTICLE LINKING

We describe our methods to assess the likelihood a link l exists between an article, a, and a specific policy item, p, or $P(l|a, p)$.

In PRM, we learn conditional attributes $h_1, ...h_t$ of either the article, policy, or both and marginalize over them, as shown in Figure 2. (Note that the model $p(h_i|a, p) = p(h_i|a)$ if the attribute h_i is *only* dependent on the article, a.) Each h_i is chosen after conducting error analyses to determine which areas the previous learned attributes, $h_{<i}$, fell short. We work with two journalists and we annotate data for each hidden attribute, h_i. We have interannotator agreement these tasks $\kappa > .8$.

(1) h_1: "a covers SFBOS". We annotate 100 articles on whether they cover SFBOS, specifically, and train a classifier $p(h_1|a)$.

[1] We release all code and data to our work here: https://github.com/alex2awesome/newsworthiness-public

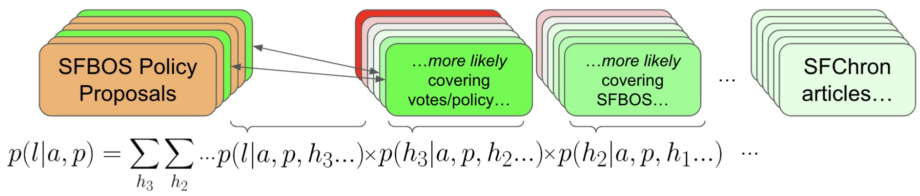

Figure 2: Our probabilistic relational modeling (PRM) process for whether an article a covers a city council proposal, p, i.e. are linked, l. PRM works by introducing auxiliary marginal variables $h_1, ...h_n$ that refine the link model, $p(l|a,p)$ through conditioning. In the diagram, moving from right-to-left, each step shows another variable h_i being applied in the PRM-chain: e.g. h_2 =*"covering SFBOS"*, h_3 =*"covering SFBOS votes and policy"*. h_2, h_3, etc. can be learned separately, and we learn supervised models for each step.

(2) h_2: "a covers votes/policy". We label an additional 100 articles on whether they mention votes and policy. We train a classifier $p(h_2|h_1, a)$.
(3) h_3: "a covers recent policy from SFBOS". We use GPT3.5 with a 10-shot prompt to determine whether a mentions votes occurring less then a month prior to publication. We use logits for "yes"/"no" as $p(h_3|h_2, h_1, a)$.
(4) l: "a covers policy p." We match articles to city-council meeting minutes using cosine similarity over the vector space.

All hidden attributes, h_i are binary variables, taking values "yes" or "no". We learn them by training TF-IDF [11] and Logistic Regression classifiers. Each hidden attribute, we find, can be learned with F1 > .8 and less than 100 annotations. To learn the final linking model, $p(l|a, p)$, we test different representations for articles and policies: TF-IDF, SBERT[2] [12] and OpenAI's `text-embedding-ada-002` embeddings. Finally, we choose a threshold, λ above which items will be considered a match using an evaluation set of 100 true pairs.

2.1 Corpora: SFBOS Policy-Proposals and SFChron Articles

We gather HTML of all SFChron articles published between 2013–2023 and via the Common Crawl. We deduplicate based on text, and ultimately are left with a set of 202,644 SFChron articles[3]. We also scrape the public meeting calendar on the SFBOS website to collect all SFBOS meetings between 2013-2023[4]nd then collect the proposal text for 13,089 SFBOS policy proposals that were discussed a total of 27,371 times in 410 public meetings. Each policy is, on average, discussed in 3 separate SFBOS meetings.

2.2 Linking Results

Our attribute-based model, as shown in Table 1, helps us retrieve $(a, p) \in S_{gold}$ with 68% f1. We show via an ablation experiment that each attribute h_i is important for our final prediction: Table 1 shows how F1 drops from 68% to 16% when we remove h_i-conditioning steps. Despite our positive results, we acknowledge that our approach is limited in several ways. First, as mentioned above, our identification of hidden attributes was based on manual error analysis and, ultimately may not a scale to new domains. Secondly, another limitation we face is that if there is no lexical overlap between a and p, we would not discover a link even if there were one. Also, we might be more exposed to this risk than the results show: in constructing S_{gold}, our annotators might have also faced a similar bias depending on the retrieval mechanisms (e.g. search) they used. A more comprehensive evaluation set would be generated by journalists *as they are working* on stories. We discuss further limitations in Section 5.

2.3 Linking Analysis

We scale our models across our entire corpus of SFChron and SFBOS articles from 2013-2023.

Roughly 7.8% of SFBOS policy proposals get covered, or 1,105 out of 13,089 policies. These policies are covered by a total of 1,828 news articles. The policies that are covered many times are a mixture of staffing (e.g. "Nomination of a Successor Mayor"), transportation bills (e.g. "Unauthorized scooter violations") and emergency ordinances (e.g. "COVID-19 Safe Shelter Operations".)

Coverage of policies is constant across time. As shown in Figure **??**, between 1–3 policies are covered per meeting, out of between 50–60 presented. This equates to between 2%–6% of proposals being covered consistently throughout our 10 years window. Coverage is relatively constant throughout the observation period, removing newspaper decline[7] as a possible confounder to newsworthiness decisions.

3 NEWSWORTHINESS PREDICTION

Next, we wish to ask *why* certain policy proposals are covered. To address this, we train *newsworthiness predictor models*. Our goal is twofold: (1) Learning a good model can show us which features of policy-items lead to coverage. (2) Performing this task well at inference time takes us steps closer to building tools that will be useful for surfacing potential stories.

3.1 Newsworthiness Training Corpus

We extract features from the linked (a, p) pairs derived in the first section to construct our training corpus. As shown in Figure 1, in the news article, there are remarks: "After 14 months of delay", "widespread debate" that seem to indicate that there aspects of this policy that are *not* solely related to its topic that made it newsworthy.

To capture some of these features, we include SFBOS meetings where these policies are discussed. We download audio for all

[2] `all-MiniLM-L6-v2` model
[3] We release the full list of URLs in our experiment, as well as scripts to replicate our collection process.
[4] a

PRM-Chain	TF-IDF	SBERT	OpenAI Embeddings
$p(l\|a,p)$, base	16.0	32.1	30.3
$\sum_{h_1} p(l\|a,p,h_1)p(h_1\|a,p)$	28.5	33.9	37.5
$\sum_{h_1,h_2} p(l\|a,p,h_1,h_2)p(h_2\|h_1,a,p)...$	55.3	48.2	53.5
$\sum_{h_1,h_2,h_3} p(l\|a,p,h_1,h_2,h_3)p(h_3\|h_1,h_2,a,p)...$	**68.2**	55.6	62.6

Table 1: Results from training PRM chains, using different sentence embeddings to calculate l.

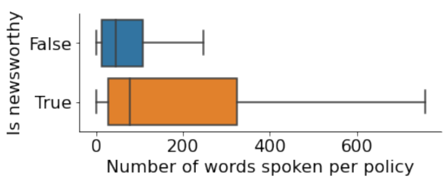

Figure 3: Number of words spoken per meeting for newsworthy policies versus non-newsworthy policies.

meetings in our corpus[5] and we use the WhisperX package [1] to transcribe and perform speaker-diarization. Finally, in every SFBOS meeting, there is a special time for members of the public to speak, called "Public Comment". We hypothesize that "Public Comment" might offer an additional lens on a policy's newsworthiness. We use diarization to identify speakers that *only* spoke during "Public Comment".Then, we calculate the lexical overlap between their speech and the policy text.

3.2 Newsworthiness Descriptive Analysis

Before showing results from the predictive modeling, we show descriptive results. Our main takeaway from this section is that policy text, meeting text and public speakers each are conveying *different* newsworthiness information. We point these out because we will show in the next section, despite clear differences observed in the features that we gathered, not all are semantically useful.

Policy Text, Meeting Speech and Public Comment all cover different newsworthy topics. We see a clear pattern in the kinds of words and topics used in newsworthy policies, meeting speech and public commenters. Table 2 shows the top most likely words in each aforementioned text category, calculated as $\Delta p(w) = p(w|Y(p) = 1) - p(w|Y(p) = 0)$. In the written policy text, we observe topic-specific words like "housing", "covid" and "cannabis" more in newsworthy policies. Topics that were more likely to receive coverage, shown in Table 3, include "Hearings" and "Environment". However, meeting speech for newsworthy policies (which is primarily speech of the SFBOS Supervisors and staff) is directed at deliberation, like "think" and "know". Finally, during public comment, we see topic-specific speech, but related to a different set of concerns, like "solar", "caltrain", "hotels". We hypothesize that these are each different aspects of newsworthiness that are being conveyed.

Newsworthy Policies are addressed for longer at meetings, by more people. Policies that end up getting covered in SFChron are also discussed at greater length than policies that are not: this includes (1) more words spoken (Figure 3), (2) more minutes spent discussing (7.7 minutes vs. 2.1), and (3) more speakers spent addressing it (4 speakers vs. 2.2. This number includes members of the public and council members.) [6].

The number of public commenters we are able to associate with specific policies, on the other hand, is a relatively small number. We are only able to establish an expected $n = .06$ speaker per newsworthy policy and $n = .04$ speaker per non-newsworthy policy. This amounts to 768 speakers associated, overall, with 13,089 policies. Thus, we hypothesize that public comment will not impact our modeling performance, despite observations in Figure 2 that public commenters tend to speak to different topics. *We acknowledge this as yet another limitation of our work and dataset.* We hope that future work can either (1) establish better methodologies to associate more public commenters with policies (2) collect larger public meeting datasets or (3) incorporate other channels (e.g. social media).

3.3 Results and Insights

In order to jointly model numerical and textual features, we choose to format our features jointly as a prompt. The structure of our full prompt is shown in Table 4. We use this prompt to fine-tune the GPT3-Babbage model, shown to be a robust classifier [16] outperforming architectures designed for text classification [14].

Policy text is the most predictive newsworthiness attribute. In our first set of experiments, we ablate the prompt to explore which components of the policy are the most important for assessing newsworthiness. We perform a temporally-based train/test split hinging on 2021/1/1. We balance our training set, with $n_{train} = 641/627$ ($Y(p) = 1/0$), and leave our test set unbalanced, with $n_{test} = 180/2310$. We perform a time-based split rather than a randomized to test how well we can predict future newsworthy items.

The full prompt performs the best across all metrics. Ablating "Public Comment" from the prompt barely impacts performance, while ablating all "meeting info." impacts a little more. Removing "policy text" information impacts performance dramatically. GPT3 outperforms TFIDF+LogReg (LR in Table 5), but not by much, indicating the power of simple textual cues.

GPT4 might be capturing national newsworthiness trends. Vanilla GPT4 outperformed our expectation. Manual analysis we perform finds that many errors were GPT4 failing to identify *locally newsworthy* items (e.g. "local scooter ban") and that many correct predictions were made on *nationally* newsworthy trends (i.e. "COVID-19 responses"). There are two likely conclusions: (1) SFChron has major

[5]Example: https://sanfrancisco.granicus.com/player/clip/43908.

[6]Journalists gave us initial feedback, saying that city councils sometimes shove important policies into sections of the meeting like "Consent Calendar" and "Roll Call", which are typically *not* addressed for a long period of time. This implies either that these cases are truly a minority, or that not enough attention is being paid to these sections of the meeting.

Δ Word Distributions for Newsworthy vs. Non-Newsworthy Text							
Policy Text				Meeting Speech		Public Comment	
authorizing	-0.41	housing	0.35	supervisor	1.98	budget	0.40
county	-0.30	health	0.31	think	0.89	philippines	0.16
grant	-0.26	board	0.30	know	0.82	solar	0.15
lawsuit	-0.25	ordinance	0.29	want	0.78	medical	0.15
bonds	-0.23	covid	0.28	people	0.76	covid	0.14

Table 2: Most likely words associated with newsworthy policy proposals, meeting speech and public comment. Also shown in the left-most column is the *least* likely words (negative-valued).

City Lawsuits	Tax/Revenue	Basic Services	Environment	COVID-19	Hearings
francisco	<number>	department	planning	ordinance	health
san	exceed	grant	code	tax	hearing
city	city	housing	findings	tent	case
county	contract	program	environmental	hotel	commission
lawsuit	authorizing	health	street	emergency	filed

Table 3: Selection of top topics obtained by running LDA with $k = 10$. Color-coding shows the likelihood of a newsworthy city council meeting minute containing a topic, with green being more likely and purple being less likely. Titles are inferred topics.

Full Prompt Example
(1) Policy description: "Priority for Veterans with an Affordable Housing Preference under Administrative..."
Presented in 2 prior meetings, 0 news articles
(2) Introduced by 4 speakers in the meeting for 0.7 minutes:
"...Without objection, this ordinance is finally passed unanimously. Madam Clerk..."
(3) 1 members of the public spoke for 1 minutes.
"<SPEAKER 1> spoke for 1 minutes and said: "Hello, this is [redacted]. I would like to oppose motions affirming..."
Is this newsworthy? Answer "yes" or "no".

Table 4: Example prompt that shows 3 primary components: (1) Policy text, (2) Meeting text and (3) Public commentary text (name censored). Text is truncated for brevity. Section lines/numbers shown for clarity.

Model	F1	ROC	R@10	MRR
Fine-tuned GPT3-Babbage				
full	**25.1**	**75.9**	**64.1**	**29.2**
(1), (2)	24.2	71.2	63.1	27.2
(1)	16.2	64.5	52.2	23.1
(2), (3)	14.4	57.6	37.2	15.9
LR, full	19.7	67.3	51.1	22.8
GPT4, full	18.4	62.6	40.6	16.2
GPT3.5, full	13.4	63.2	46.7	21.3

Table 5: Results from fine-tuning GPT3 on full and ablated versions of the prompt. Bottom sections show our baselines, Logistic Regression (LR) and vanilla GPT4/GPT3.5. Metrics are: F1, ROC-score over logits for "yes" tokens, Recall@10 (R@10) of each meeting (i.e. we surface the 10 most likely newsworthy items, count recall) and Mean Reciprocal Rank (MRR) of newsworthy policies, per meeting.

Train	F1	ROC	MRR	R@10	n
'13-'21	25.4	75.9	.26	64.4	1,595
'13-'20	18.9	68.8	.22	52.8	1,289
'13-'19	21.8	69.9	.22	53.9	1,084
'13-'18	19.5	67.8	.23	55.0	867
'13-'17	17.9	66.1	.22	52.2	693

Table 6: We alter the training split date cutoffs to test whether GPT is learning specific newsworthy events (e.g. "COVID-19") too well, or whether it is picking up broader news values [3].

overlaps with national newspapers, and (2) general newsworthy language and framing is *also* used for local newsworthiness.

Newsworthiness judgements are surprisingly consistent across time, with one major exception. Table 2 and Table 3 show that words related to specific events (e.g. those related to "COVID-19") are reflected in the perceived newsworthiness of policy: is the model fitting to a specific event (e.g. "COVID-19") that happens to be newsworthy, or is it learning either (1) news values [3] (2) newsworthy language patterns/framing? To test this question, we retrain our model and restrict the training date cutoffs. We show in Table 6 that, except for a dropoff after 2021, our performance does not significantly change.

To test whether this is the result of GPT3's pretraining, we test and are able to replicate these findings with baseline Logistic Regression models. An error analysis shows that "COVID-19"-related news was the least likely to be predicted correctly, and is the main contributor to this performance decrease, whereas there are numerous other specific events that emerge (e.g. environmental, transportation-related, fire-arms related events.) that our models predict correctly. We take this as evidence that *major* anomalous events, like COVID-19 specifically, do become newsworthy and are unpredictable given our current approach. This highlights an important limitation of our approach, as mentioned in Section 3. These need to be taken into account if these tools are deployed: they must be used along with other models better tuned to these blind spots.

Human journalists find our newsworthiness judgements predictable and helpful. Finally, we recruit two expert journalists[7] and conduct human experiments with two aims: (1) is our "newsworthiness" definition repeatable and (2) are our models helpful? For the first, we test how well *humans* able to identify newsworthy SFBOS policies. We construct a dataset by taking newsworthy policies from SFBOS meetings in our test set and a sampling nonnewsworthy policies in a 1-to-2 ratio of $Y(p) = 1, 0$. Our best models achieve 58.9 F1-score on this dataset, and humans score almost equivalently. It's tempting to think our models have reached a ceiling; however, the journalists are not San Francisco-based, and are thus untrained. To test how useful these models can be, we surface 10 policies from each meeting and ask journalists to (a) indicate which policies they might write about and (b) guess whether the list was a newsworthiness list or a random sample (they were told that it was a secondary method, not random). We found, for (a), that journalists preferred our lists to random 84% of the time, and for (b) were able to guess which list was generated via our method 74% of the time.

4 RELATED WORKS

Newsworthiness is a well-studied concept sociological concept, starting with [3]'s identification of "news values" like *timeliness*, *eliteness* and *proximity*. [5] followed up with work focused on local news values (e.g. *downtown proximity*, *economic boosterism*). Because these features are complex, neither work can easily be operationalized as predictive algorithms. However, newsworthiness prediction has been approached in different ways. [15] and [9] sought to learn distant signals for document newsworthiness: either by classifying article layout in newspapers or by collecting attributes from crowd-workers. Our work more directly addresses the question "will this be written about?" and allows us to study it in a data-driven manner. Another approach is given by Diakopoulos et al. [2], where a piece of content's *relevance* to a given topic, its *uniqueness*, and its *sentiment* is quantified. Then, these metrics surface tweets related to presidential speeches. Such metrics-based systems can be interpretable, but can also miss newsworthy items that are not ranked highly by such metrics. Our work might benefit from including these metrics, and our dataset might learn to rank them well among our other features. Finally, link prediction is a well studied field [6]. PRMs were introduced [4] as a way of modeling links, but suffer from high computational cost. Our approach (a) uses a small dataset and (b) uses entirely supervised models to make PRMs tractable.

5 DISCUSSION AND CONCLUSION

In summary, we established links between a large corpus of news articles and local policy proposals we did so using a classical method, probabilistic relational modeling, that outperformed retrieval-based methods and embedding-based methods with only a small amount of annotated data. We used the assessed newsworthiness of prior articles to build models to predict the newsworthiness of articles. We found that the performance of our models did not degrade over time, and we found that expert journalists agreed with our newsworthiness assessments and found our tools helpful.

Our work faces many limitations. Notably, we assumed that historical coverage patterns are a reasonable starting point for modeling future newsworthiness predictions. While we found that this yielded useful models, there might be cases where news values evolve and prior decision-making is morally and ethically unacceptable, for example with crime [10] or suicide coverage [8]. Our work would serve enforce such historical patterns. Also, it might miss major, atemporal results, like COVID-19. Both of these represent considerable risks, and indicate that human involvement remains crucial in any kind of newsworthiness prediction system.

Despite these risks and limitations, we see this work as presenting a crucial starting point for a larger research direction in newsworthiness prediction. By establishing "newsworthiness" as a well-defined predictive task, we hope to have opened the door to future work applying these concepts. We intend in upcoming work to explore ways to introduce control and explainability into the newsworthiness prediction pipeline that we have outlined here.

REFERENCES

[1] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

[2] Nicholas Diakopoulos, Mor Naaman, and Funda Kivran-Swaine. 2010. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pages 115–122. IEEE.

[3] J Gatlung and Mari Holmboe Ruge. 1965. The structure of foreign news. *Journal of Peace Research*, 2(1):64–91.

[4] Lisa Getoor, Nir Friedman, Daphne Koller, and Benjamin Taskar. 2002. Learning probabilistic models of link structure. *Journal of Machine Learning Research*, 3(Dec):679–707.

[5] Phyllis Kaniss. 1991. *Making local news*. University of Chicago Press.

[6] Ajay Kumar, Shashank Sheshar Singh, Kuldeep Singh, and Bhaskar Biswas. 2020. Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and its Applications*, 553:124289.

[7] Nick Mathews. 2022. Life in a news desert: The perceived impact of a newspaper closure on community members. *Journalism*, 23(6):1250–1265.

[8] Thomas Niederkrotenthaler, Marlies Braun, Jane Pirkis, Benedikt Till, Steven Stack, Mark Sinyor, Ulrich S Tran, Martin Voracek, Qijin Cheng, Florian Arendt, et al. 2020. Association between suicide reporting in the media and suicide: systematic review and meta-analysis. *Bmj*, 368.

[9] Sachita Nishal and Nicholas Diakopoulos. 2022. From crowd ratings to predictive models of newsworthiness to support science journalism. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28.

[10] Mary Beth Oliver. 2003. Race and crime in the media: Research from a media effects perspective. *A companion to media studies*, pages 421–436.

[11] Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

[12] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

[13] Naoki Shibata, Yuya Kajikawa, and Ichiro Sakata. 2012. Link prediction in citation networks. *Journal of the American society for information science and technology*, 63(1):78–85.

[14] Alexander Spangher, Jonathan May, Sz-Rung Shiang, and Lingjia Deng. 2021. Multitask semi-supervised learning for class-imbalanced discourse classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 498–517.

[15] Alexander Spangher, Nanyun Peng, Jonathan May, and Emilio Ferrara. 2021. Modeling" newsworthiness" for lead-generation across corpora. *arXiv preprint arXiv:2104.09653*.

[16] Alexander Spangher, Nanyun Peng, Jonathan May, and Emilio Ferrara. 2023. Identifying informational sources in news articles. *arXiv preprint arXiv:2305.14904*.

[7] Combined have > 40 years of newsroom experience