Tracking the Newsworthiness of Public Documents

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

Abstract

Journalists regularly make decisions on whether or not to report stories, based on "news values" (Gatlung and Ruge, 1965). In this work, we wish to explicitly model these decisions to explore *when* and *why* certain stories get press attention. This is challenging because very few labelled links between source documents and news articles exist and language use between corpora is very different. We address this problem by implementing a novel *probabilistic relational modeling* framework, which we show is a low-annotation linking methodology that outperforms other, more state-of-the-art retrieval-based baselines. Next, we define a new task: ***newsworthiness prediction***, to predict if a policy item will get covered. We focus on news coverage of local public policy in the San Francisco Bay Area by the *San Francisco Chronicle*. We gather 15k policies discussed across 10 years of public policy meetings, and transcribe over 3,200 hours of public discussion. In general, we find limited impact of public discussion on newsworthiness prediction accuracy, suggesting that some of the most important stories barely get discussed in public. Finally, we show that newsworthiness predictions can be a useful assistive tool for journalists seeking to keep abreast of local government. We perform human evaluation with expert journalists and show our systems identify policies they consider newsworthy with 68% F1 and our coverage recommendations are helpful with an 84% win-rate against baseline.[1]

1 Introduction

Despite much qualitative analysis of newsworthiness (Gatlung and Ruge, 1965; Kaniss, 1991) very little quantitative work has attempted to analyze: (1) *what* stories get covered, (2) *why* have they been covered, and (3) what *impact* does the coverage has? Not only could such work increase our understanding of coverage patterns and informational salience perceptions (Hamilton and Fallot, 1974), but it could lead to assistive tools to surface leads for a journalist to pursue (Cohen et al., 2011). To that end, we propose a new task, *newsworthiness prediction*, to predict whether a story should get covered, *according to previous coverage patterns*.

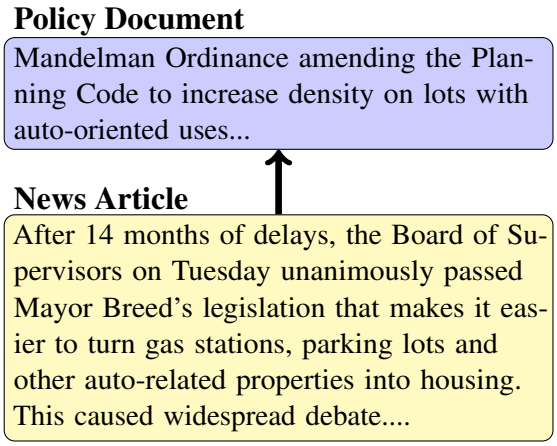

Figure 1: In this paper, we establish the *newsworthiness prediction* task. We (1) train models to infer when public policy items have been covered in the press and (2) predict if new items *will* be covered.

Determining that a story, in this work taken to be a local government policy item[2] was covered in media, as shown in Figure 1, is a challenging task. Unlike related tasks, like *citation prediction* (Shibata et al., 2012) or *cross document event-coreference* (Bagga and Baldwin, 1999), determining policy coverage requires us to establish links between documents in two different linguistic domains, with no pre-existing labels. Our first challenge, in this work, is to establish when a news article references a specific local policy document, i.e. to *link* them. We show that breaking this problem down into a chain of decisions, each conditional on the pre-

*Corresponding Author: spangher@usc.edu
[1]We release all code and data to our work here: https://github.com/alex2awesome/newsworthiness-public
[2]i.e. A motion of gov.: a proposal, bill, settlement, etc..

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14150–14168
August 11-16, 2024

vious[3], an application of probabilistic relational modeling (PRM) (Getoor et al., 2002), helps us outperform other retrieval-based baselines.

Next, having established links, we seek to predict if a *new* policy will get covered. We fine-tune language models based on large silver-linked corpora that we identify using the PRM. We find that although F1 is low for coverage prediction, our models are helpful to journalists, beating baseline 84% of the time and surfacing relevant items.

Finally, we ask whether policy text alone is enough to predict coverage. We study recordings of public meetings where policy proposals are addressed. 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 newsworthy characteristics might not get discussed.

In sum our contributions are:

1. We collect a large multimodal dataset of 13,000 SFBOS policy proposals spanning 10 years, 20,000 SFChron news articles and 3,200 hours of SFBOS meeting video (which we transcribe and diarize), in Sections 2.1, 3.1 in order to study newsworthiness in the local context of one city.

2. We link these corpora with a novel application of probabilistic relational modeling, outperforming modern baselines (Section 2.2). We find that between 2-6% of SFBOS policies get covered in SFChron (Section 2.4).

3. We establish a novel task, *newsworthiness prediction*, and use it to analyze what makes policy and public discussion newsworthy, finding that newsworthiness is predictable and has a strong non-temporal element (Section 3.3). We show journalists find our rankings helpful in surfacing newsworthy leads.

This work, to our knowledge, is the first to *operationalize* newsworthiness prediction by focusing on trying to predict historical coverage patterns. As such, we start small in order to prove that this direction is viable, focusing on one locality, San Francisco Board of Supervisors (SFBOS) and it's coverage in the the San Francisco Chronicle (SFChron). This means that our analysis is necessarily limited to "newsworthiness" as defined by the SFChron, on SFBOS policy text. However, we intend in future work to expand to different localities.

Press coverage of local policy can be crucial for the health of a community: it can increase civic engagement (Smith, 1987), reduce government malfeasance (Bruns and Himmler, 2011) and engender greater productivity in society (Snyder Jr and Strömberg, 2010). Yet, many newsrooms face severe economic challenges preventing them from providing robust coverage (Fisher and Park, 2022). We hope that our work can open the door to assitive tools to increase journalists' capabilities to cover their communities.

2 Policy Item ↔ Article Linking

Our goal in this section is to determine which articles cover which policies. We seek to model the likelihood a link l exists between an article, a, and a specific policy item, p, or $P(l|a,p)$.

We apply the *probabilistic relational model* (PRM) framework (Getoor et al., 2002) to solve this problem. In PRM, we learn conditional attributes $h_1, ...h_t$ of either the article, policy, or both and marginalize over them:

$$P(l|a,p) = \sum_{h_1} \dots \sum_{h_t} p(l|a,p,h_1,\dots,h_t) \dots p(h_1|a,p) \quad (1)$$

Where, as shown in Figure 2, h_2 might be "covers SFBOS", and h_3 might be "covers SFBOS votes/policy."[4] (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.) Not all politics articles are about SFBOS, and not all SFBOS articles cover policy. Such variety confounds unsupervised models, but is solvable when broken into easier-to-supervise subproblems. This is not dissimilar to Chain-of-Thought (CoT) (Wei et al., 2022), where language models decompose complex reasoning tasks.

2.1 Corpora: SFBOS Policy-Proposals and SFChron Articles

We focus on a specific local government, SFBOS, and a specific newspaper, SFChron, that has a ro-

[3]Shown in Figure 2, i.e. "article covers local politics" → "article covers city council meetings" → "covers past meeting" → "covers *this* past meeting"

[4]Because no natural linking information exists (i.e. hyperlinks in the article body), we typically model l_* on the text of the article and/or policy proposal.

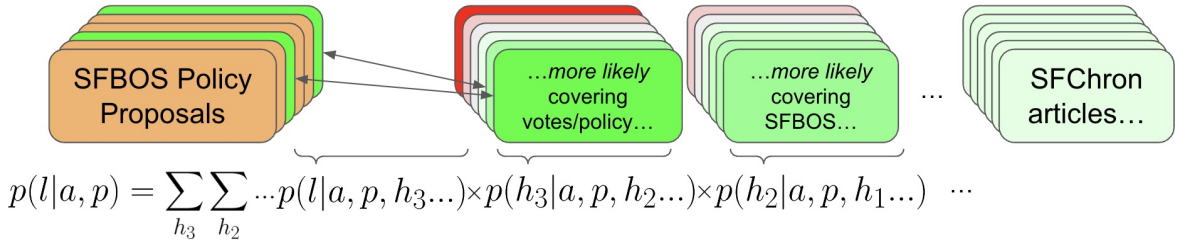

Figure 2: Our probabilistic relational modeling (PRM) process for whether an article a covers a city council proposal, p, i.e. are linked, l. PRM works by introducing auxiliary marginal variables $h_1, ...h_n$ that refine the link model, $p(l|a,p)$ through conditioning. In the diagram, moving from right-to-left, 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.

bust local news section. We start by gathering HTML of all SFChron articles published between 2013–2023 and via the Common Crawl[5]. We parse article text[6] and deduplicate based on text, and ultimately are left with a set of 202,644 SFChron articles[7]. We also scrape the public meeting calendar on the SFBOS website[8] to collect all SFBOS meetings between 2013-2023[9] and then collect the proposal text for 13,089 SFBOS policy proposals[10] 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 Devising Relational Chains

We manually identify a sequence of hidden attributes, h_i, to learn (shown in Figure 2) and handcraft models to learn each one. Each h_i is chosen after conducting error analyses to determine which areas the previous learned attributes, $h_{<i}$, fell short. We hire two journalists[11] to annotate data for each hidden attribute, h_i, and calculate that their inter-annotator agreement on these tasks is $\kappa > .8$.

1. h_1: "a covers SFBOS". We use the keyword t ="Board of Supervisors" to identify candidates. Then we delete t from these candidates, sample negatives and bootstrap a classifier to identify more candidates. Our annotators label 100 of these and we train a classifier $p(h_1|a)$.

2. h_2: "a covers votes/policy". From h_1 articles, our journalists 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 also ask GPT3.5 to confirm the government body is SFBOS (e.g. not "Oakland City Council"). 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 (Ramos et al., 2003) and Logistic Regression classifiers. Each hidden attribute, we find, can be learned with F1 > .8 and less than 100 annotations. For more details about learning h_i, see Appendix A.

To learn the final linking model, $p(l|a,p)$, we test different representations for articles and policies: TF-IDF, SBERT with the `all-MiniLM-L6-v2` model (Reimers and Gurevych, 2019) and OpenAI's `text-embedding-ada-002` embeddings. (We note that SBERT is run on the first 256 tokens of the article and policy text). Finally, we wish to choose a threshold, λ, for Equation 1, above which items will be considered a match. To help us choose λ and to evaluate our method, our annotators manually identify 100 true pairs, which we split 50/50 into $S_{gold,train}$ and $S_{gold,test}$.

[5]We search for all URLs matching wildcard pattern `https://www.sfchronicle.com/*`

[6]Using `https://github.com/codelucas/newspaper`.

[7]We release the full list of URLs in our experiment, as well as scripts to replicate our collection process.

[8]`https://sfgov.legistar.com/Calendar.aspx`

[9]Example meeting: `https://sfgov.legistar.com/MeetingDetail.aspx?ID=1108038&GUID=8B3A2668-90A9-43E9-A694-8747176617F4`

[10]Example of a policy proposal: `https://sfgov.legistar.com/LegislationDetail.aspx?ID=6251774&GUID=420031B2-94DE-440F-AB74-25FF091F2D61`

[11]Both journalists were U.S. citizens, but neither was a San Francisco resident. We adjust pay to be roughly $20 USD an hour.

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. l is defined as a mapping between News article $a \leftrightarrow$ Policy mapping p. We establish a score-threshold for $p(l|a,p)$ for each trial using our gold-labeled dataset, $S_{gold,train}$ and report f1-scores using $S_{gold,test}$. TF-IDF is defined Ramos et al. (2003). SBERT uses the `all-MiniLM-L6-v2` model (Reimers and Gurevych, 2019). OpenAI uses the `text-embedding-ada-002` model.

2.3 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.

Surprisingly, using PRM with TF-IDF outperforms different embedding methods like SBERT (Reimers and Gurevych, 2019) and OpenAI embeddings (Ryan Greene, 2022). We suspect that specific technical phrases are important for this task, which unsupervised embeddings might ignore; training a supervised retrieval architecture like Dense Passage Retrieval (DPR) might help represent these phrases in the embeddings, but as reported by Karpukhin et al. (2020) requires 100-1000 times more data than we have collected. Our PRM approach also outperforms retrieval-specific methods like BM25 (Robertson et al., 2009). Overall, these results indicate that attribute-specificity of PRM is crucial[12]. We note that our PRM approach can seen as a supervised variation of CoT reasoning (Wei et al., 2022) (albeit with a wide beam). As language models become cheaper and more scalable, more directly applying CoT-style approaches to either identify hidden attributes to train auxiliary classifiers, or directly link articles and policies, could be a viable approach.

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.

2.4 Linking Analysis

We scale our models across our entire corpus of SFChron and SFBOS articles from 2013-2023. Examining these links gives us insights into the amount of coverage devoted to public policy.

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. Although each policy is covered on average 1.8 times, the distribution is right-skewed and the median coverage is one time per policy. See Appendix B, Figure 6 for more details. 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".) Again, see Appendix B, Table 10.

Coverage of policies is constant across time. As shown in Figure 3, 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(Mathews, 2022) as a possible confounder to newsworthiness decisions[13]. We observe a brief

[12]To implement BM25, we index a and use p as a search query. We use the `retriv` Github package: https://github.com/AmenRa/retriv.

[13]Due to an ongoing economic crisis in journalism, many newspapers are shrinking, leading to less coverage and, possi-

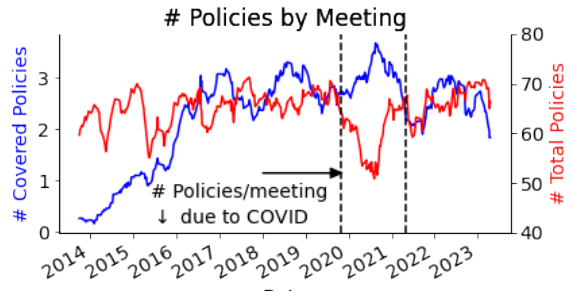

Figure 3: The number of policies introduced to SFBOS and those covered by SFChron, measured by the date the policy was introduced and whether $p(a, p) > \lambda$.

spike in % of p covered in 2020–2021. Closer examination reveals that the number of (a, p) pairs stay constant, however, the number of p *proposed* drops significantly, from an average of 64 policies per meeting prior to 2021, to 52 policies per meeting in the first 9 months of 2021. This is likely the result of COVID-related shutdowns. Conversations with SFChron journalists confirm this.

3 Newsworthiness Prediction

Next, we wish to ask *why* certain policy proposals are covered. To address this, we establish a new task, *newsworthiness prediction*: predict, given a policy item p, if an article will write about it. We use our linked dataset $\{(a, p)\}$, in Section 2, and treat this problem as a prediction problem where:

$$Y(p) = \begin{cases} 1, & \text{if } p \in \{(a, p)\} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

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.

Previously, *newsworthiness* has been addressed as a feature-detection problem, as in (Diakopoulos et al., 2010), where engineered-features measured specific criteria[14]. Journalists examined combinations of features to find newsworthy items but could miss items if their newsworthiness did not fit the measurements. Because we formulate our task as a prediction task, backed by a dataset, we can also expose new and possibly unexpected features. However, a prediction-based approach is limited in its own ways. We assume that past coverage patterns predict future patterns, and that journalists generally agree. We will explore these assumptions in Section 3.3, and we will also see notable cases where these assumptions *do* limit us.

Policy Features Analyzed
text of proposal
prior meetings proposal has been discussed
prior news articles linked to proposal
length of time proposal is discussed in meeting
transcribed text of city-council member's policy discussion
public commenters discussing the policy
summary of public commentary

Table 2: Summary of features for each policy item. Top section is generated via (a, p). Bottom section is generated via SFBOS video transcriptions.

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 meetings in our corpus[15] and we use the WhisperX package (Bain et al., 2023) to transcribe and perform speaker-diarization. See Appendix C for more about aligning transcripts. We associate each (a, p) with a specific meeting if: (1) p is discussed in the meeting and (2) a was published within a month of the meeting occurring.

Finally, in every SFBOS meeting, there is a special time for members of the public to speak, called "Public Comment". Since good newswriting is emotional (Uribe and Gunter, 2007), we hypothesize that "Public Comment" might offer an additional lens on a policy's newsworthiness. We determine which speakers are members of the public using diarization to identify speakers that *only* spoke during "Public Comment"[16]. Then, we calculate the lexical overlap between their speech and the policy

bly, changes in what is considered "newsworthy".

[14]E.g. "statistically anomalous" (Zhao et al., 2014), "sentiment=happy")

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

[16]We infer the sections of the transcript like "Public Comment" using time-stamped agendas, see Appendix C for more detail.

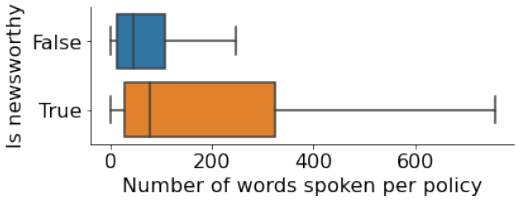

Figure 4: Number of words spoken per meeting for newsworthy policies versus non-newsworthy policies.

text. For more details about "Public Comment" and other meeting sections, please see Appendix E. We show all of the features that we use for newsworthiness prediction in Table 2.

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 3 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 8, 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 4), (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.) [17].

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 3 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, and it includes all features listed in Table 2. We limit the size of the prompt by providing only the first 50 words of the text fields (besides "proposal text"). We do not notice any impact of this truncation in early experimentation. We use this prompt to fine-tune the GPT3-Babbage model, shown to be a robust classifier (Spangher et al., 2023b), outperforming architectures designed for text classification (Spangher et al., 2021a). It could be that the length time spoken is a more important variable than the time spoken itself.

Policy text is the most predictive newsworthiness attribute, followed by meeting discussion and then public comment. 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

[17] 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
settlement	-0.22	department	0.23	like	0.58	caltrain	0.14
contract	-0.21	cannabis	0.22	need	0.43	rooms	0.13
expend	-0.19	election	0.21	president	0.37	amendments	0.12

Table 3: Most likely words associated with newsworthy policy proposals, meeting speech and public comment, measured by $p(w|Y(p) = 1) - p(w|Y(p) = 0)$, where $p(w|.)$ is based on observed word counts. Also shown in the left-most column is the *least* likely words (negative-valued). Colors shown are a heatmap for easy viewing.

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 the 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 at first 50 words. Further truncated in this example 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. All rows with (full) show models that were trained on full input prompt (Table 4). Rows with numbers, e.g. (1), etc. are ablation models trained with those parts of the prompt. 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.

split rather than a randomized split because our goal is ultimately to build a model that can predict future newsworthy items.

We find that the full prompt performs the best across all metrics we considered, but only marginally. As expected, ablating "Public Comment" from the prompt barely impacts performance, while ablating all "meeting info." impacts a little more. Removing "policy text" information, thus forcing the model to just rely on meeting text alone impacts performance dramatically. GPT3, unsurprisingly, outperforms a very simple classifier, TFIDF+Logistic Regression (LR in Table 5), but not by much, indicating that there might be simple textual cues that we are learning.

GPT4 might be capturing national newsworthiness trends. Vanilla GPT4 outperformed our expectation. We had hypothesized that many of SFChron's newsworthiness judgements on SFBOS were local. GPT4 underperforms most other classifiers, but not by much. Manual analysis we perform finds that many errors were GPT4 failing to identify *locally newsworthy* items (e.g. "local scooter ban", local street renaming) and that many correct

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 be prior to Jan 1st on each of those years to test whether GPT is learning to fit to specific newsworthy events (e.g. "COVID-19") too well, or whether it is picking up broader newsworthy trends. For definitions of metrics, see Table 5.

Task	Metric	Score
Identify Newsworthy Policies	Human F1	63.2
	(Model F1)	*(58.9)*
	Cohen's κ	36.3
Use top k=10 as recommendation system	Preference	84%
	ID Accuracy	74.2%
	Cohen's κ	60.0

Table 7: Results from human evaluation. Top row: journalists identify real newsworthy policies, by meeting, given a balanced dataset of 33% $Y(p) = 1$ and 66% $Y(p) = 0$ policies. Model f1-score is much higher than Table 5 because this is a balanced sample. Bottom row: preference test for lists of newsworthy minutes (generated viaour models vs. random) and identification (ID) accuracy for list-origin.

predictions were made on *nationally* newsworthy trends (i.e. "COVID-19 responses"). There are two likely conclusions: (1) SFChron has major overlaps for newsworthiness judgements 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 3 and Table 8 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 in our training and test data, or is it learning either (1) larger event-types (e.g. pandemics more generally, like "ebola", are recurrent and newsworthy) or (2) newsworthy language patterns and other non-semantic attributes (e.g. framing)?

To test this question, we retrain our model and increasingly restrict the date cutoffs of our training set to ask whether a model would correctly predict the newsworthiness of policies pertaining to specific events (e.g. "COVID-19") if the likelihood of them being in the dataset were to decrease. We show in Table 6 that, except for a dropoff after excluding data from 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[18] 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 $Y(p) = 1, 0$. As shown in Table 7, 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, compared to our models.

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

[18] Combined have $>$ 40 years of newsroom experience

to guess which list was generated via our method 74% of the time.

4 Related Works

Sociology of News Production Newsworthiness is a well-studied concept in communication and journalism studies, starting most famously with Gatlung and Ruge (1965)'s identification of "news values" like *timeliness*, *eliteness* and *proximity* in international reporting. Kaniss (1991) followed up with work focused on local news values, identifying *downtown proximity*, *economic boosterism* and *symbolification* as key local news values. Each of these works fit into a broader discipline of qualitatively studying newsroom practice, but are typically done via field studies and resulted in descriptive analyses, which could not be operationalized as predictive algorithms.

Local Policy and News Coverage Analysis Prediction We are not the first to gather and analyze local policy discussions at scale (Sorens et al., 2008; Brown et al., 2021; Maxfield Brown and Weber, 2022; Barari and Simko, 2023). Nor are we the first to study broad coverage patterns in local journalism (Hamilton, 2016; Baekgaard et al., 2014; Garz and Sörensen, 2017). Hamilton (2016), notably, studied *effects* of different coverage patterns in journalism on local government. Besides qualitative work (Felt, 2015), or work focused on social media (Graham et al., 2015), we believe we are one of the first to link news coverage to *specific* policy.

Link prediction is a well studied field (Kumar et al., 2020). PRMs were introduced (Getoor et al., 2002; Taskar et al., 2003) as a way of modeling attributes, but often suffered from high computational complexities. Our approach (a) uses a relatively small dataset and (b) uses entirely supervised models to ultimately make PRMs tractable here.

Newsworthiness prediction has been approached in different ways. (Spangher et al., 2021b) and (Nishal and Diakopoulos, 2022) sought to learn distant signals for document newsworthiness: either by classifying article layout in newspapers or by collecting attributes from crowd-workers, like "surprising", "impactful". Our work more directly addresses the question "will this be written about?" and allows us to study it in a data-driven manner.

This task is also called *lead generation* (Cohen et al., 2011). Of existing approaches to lead-generation, one is given by Diakopoulos et al. (2010), 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.

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 and risks, which we discuss in Sections 2.3 and 3. 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 (Oliver, 2003) or suicide coverage (Niederkrotenthaler et al., 2020). 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 (a point made by researchers studying other real-world predictive systems (Hong, 2023)).

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.

6 Ethics Statement

6.1 Limitations

We discuss a number of methodological limitations throughout our work, namely: (1) our assumptions as to linking articles give us an overreliance on lexical overlap, which is a bias our annotators might *also* share based on how they chose to retrieve (article, policy) pairs. (2) Past newsworthiness might not always generalize, and might degrade more over time. There are other limitations that exist, though. The datasets we used are all in English, and local to one geography, thus are possibly not representative.

We must view our work in newsworthiness prediction with the important caveat that non-Western news outlets may not follow the patterns. We might face fundamental differences in prediction ability or problem framing if we attempt to do such work in other languages.

6.2 Risks

Since we constructed our datasets using well-trusted news outlets and public meetings, we assumed that every informational sentence was factual, to the best of the journalist's ability, and honestly constructed. We have no guarantees that such a newsworthiness system would work in a setting where the journalist is acting adversarially.

There is a risk that, if such a work were used in a larger news domain, it could fall prey to learning newsworthiness of misinformation or disinformation. This risk is acute in the news domain, where fake news outlets peddle false stories that attempt to *look* true (Boyd et al., 2018; Spangher et al., 2020, 2018). We have not experimented how our classifiers would function in such a domain.

We used OpenAI Finetuning to train the GPT3 variants. We recognize that OpenAI is not transparent about its training process, and this might reduce the reproducibility of our process. We also recognize that it owns the models, and thus we cannot release them publicly. Both of these thrusts are anti-science and anti-openness and we disagree with them on principle. However, their models are still useful in a black-box sense for giving strong baselines for predictive problems and drawing scientific conclusions about hypotheses. By the camera ready, we will work to reduce these anti-science thrusts by experimenting with and releasing open sourced LMs. We experimented with them using DeepSpeed to run the GPT-Neo 6.7 model and the GPT Juno model on a V100 GPU. However, due to time constraints we were not able to get them working in time for submission. However, based on available evidence,[19] we expect them to work at a similar capacity and will report results on them separately when we do.

6.3 Licensing

The San Francisco Board of Supervisors dataset we used is released without any restrictions. We have had independent lawyers at a major media company ascertain that this dataset was low risk for copyright infringement. We do not release the San Francisco Chronicle dataset that we gathered, but we do release relevant URLs, which are public domain, and scripts for accessing the Common Crawl.

6.4 Computational Resources

The experiments in our paper required minimal computational resources. We used a laptop computer to run baseline logistic regression and TF-IDF matching experiments. We used OpenAI's fine-tuning and prompting architecture to train GPT3 models.

6.5 Annotators

We recruited annotators from two major newspapers that partnered with our institution during this work. They consented to the experiment and were paid at above $20 an hour. Both spent more than 5 years at their organizations. Neither organization is in the same locality as the San Francisco Chronicle. Both annotators are male. Both identify as cis-gender. Both are over 30 years old. This work passed a university Institutional Review Board.

References

Martin Baekgaard, Carsten Jensen, Peter B Mortensen, and Søren Serritzlew. 2014. Local news media and voter turnout. *Local Government Studies*, 40(4):518–532.

Amit Bagga and Breck Baldwin. 1999. Cross-document event coreference: Annotations, experiments, and observations. In *Coreference and Its Applications*.

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

[19] https://blog.eleuther.ai/gpt3-model-sizes/

Soubhik Barari and Tyler Simko. 2023. Localview, a database of public meetings for the study of local politics and policy-making in the united states. *Scientific Data*, 10(1):135.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Ryan L Boyd, Alexander Spangher, Adam Fourney, Besmira Nushi, Gireeja Ranade, James Pennebaker, and Eric Horvitz. 2018. Characterizing the internet research agency's social media operations during the 2016 us presidential election using linguistic analyses.

Eva Maxfield Brown, To Huynh, Isaac Na, Brian Ledbetter, Hawk Ticehurst, Sarah Liu, Emily Gilles, Sung Cho, Shak Ragoler, Nicholas Weber, et al. 2021. Council data project: software for municipal data collection, analysis, and publication. *Journal of Open Source Software*, 6(68):3904.

Christian Bruns and Oliver Himmler. 2011. Newspaper circulation and local government efficiency. *Scandinavian Journal of Economics*, 113(2):470–492.

Sarah Cohen, James T Hamilton, and Fred Turner. 2011. Computational journalism. *Communications of the ACM*, 54(10):66–71.

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.

Mylynn Felt. 2015. The incessant image: how dominant news coverage shaped canadian cyberbullying law. *UNBLJ*, 66:137.

Caroline Fisher and Sora Park. 2022. Economic and existential challenges facing journalism. *The SAGE Handbook of the Digital Media Economy*, page 280.

Marcel Garz and Jil Sörensen. 2017. Politicians under investigation: The news media's effect on the likelihood of resignation. *Journal of Public Economics*, 153:82–91.

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

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.

Melissa W Graham, Elizabeth J Avery, and Sejin Park. 2015. The role of social media in local government crisis communications. *Public Relations Review*, 41(3):386–394.

David L Hamilton and Roger D Fallot. 1974. Information salience as a weighting factor in impression formation. *Journal of Personality and Social Psychology*, 30(4):444.

James T Hamilton. 2016. *Democracy's detectives: The economics of investigative journalism*. Harvard University Press.

Jenny Hong. 2023. *Project Recon: A Computational Framework for and Analysis of the California Parole Hearing System*. Stanford University.

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

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

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.

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

Eva Maxfield Brown and Nicholas Weber. 2022. Councils in action: Automating the curation of municipal governance data for research. *Proceedings of the Association for Information Science and Technology*, 59(1):23–31.

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.

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.

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

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.

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

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Lilian Weng Arvind Neelakantan Ryan Greene, Ted Sanders. 2022. New and improved embedding model.

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.

Kim A Smith. 1987. Effects of newspaper coverage on community issue concerns and local government evaluations. *Communication Research*, 14(4):379–395.

James M Snyder Jr and David Strömberg. 2010. Press coverage and political accountability. *Journal of political Economy*, 118(2):355–408.

Jason Sorens, Fait Muedini, and William P Ruger. 2008. Us state and local public policies in 2006: A new database. *State Politics & Policy Quarterly*, 8(3):309–326.

Alexander Spangher, Xinyu Hua, Yao Ming, and Nanyun Peng. 2023a. Sequentially controlled text generation. *arXiv preprint arXiv:2301.02299*.

Alexander Spangher, Jonathan May, Sz-Rung Shiang, and Lingjia Deng. 2021a. 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.

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

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

Alexander Spangher, Gireeja Ranade, Besmira Nushi, Adam Fourney, and Eric Horvitz. 2018. Analysis of strategy and spread of russia-sponsored content in the us in 2017. *arXiv preprint arXiv:1810.10033*.

Alexander Spangher, Gireeja Ranade, Besmira Nushi, Adam Fourney, and Eric Horvitz. 2020. Characterizing search-engine traffic to internet research agency web properties. In *Proceedings of The Web Conference 2020*, pages 2253–2263.

Alexander Spangher, Xiang Ren, Jonathan May, and Nanyun Peng. 2022. Newsedits: A news article revision dataset and a novel document-level reasoning challenge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 127–157.

Ben Taskar, Ming-Fai Wong, Pieter Abbeel, and Daphne Koller. 2003. Link prediction in relational data. *Advances in neural information processing systems*, 16.

Rodrigo Uribe and Barrie Gunter. 2007. Aresensational'news stories more likely to trigger viewers' emotions than non-sensational news stories? a content analysis of british tv news. *European journal of communication*, 22(2):207–228.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Jian Zhao, Nan Cao, Zhen Wen, Yale Song, Yu-Ru Lin, and Christopher Collins. 2014. # fluxflow: Visual analysis of anomalous information spreading on social media. *IEEE transactions on visualization and computer graphics*, 20(12):1773–1782.

A Additional *Probabilistic Relational Modeling* Details

We show the F1 scores for each hidden attribute that we separately learn in our PRM chain in Table 9.

A.1 GPT3.5 Prompt

Hidden attribute h_3 relies on appropriate identification of referential timing from a news article. We craft a few-shot prompt as follows:

> You are a journalist in San Francisco Bay Area who covers local city council meetings and events.
>
> I'm going to show you some articles and you'll tell me the year, month and date that the meeting mentioned took place on. Look for clues relative to the article publish date I will provide. If you can only determine the year and month, that's OK. Ignore irrelevant dates.
>
> Here are examples:
>
> (Example 1) The article was published on: ⟨`article_publish date`⟩ Article Text: "During a school board committee meeting Tuesday night, district officials said they believed that the vast majority of the students lacking a class or two would still graduate on time..." Answer: Day-of-week: Tuesday Year: 2013 Month: 10 Day: 1 ...
>
> Ok, let's get started. Here is 1 article. What year, month and day did the meeting mentioned in the article occur? Look

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
settlement	bonds	services	section	covid-19	board
district	revenue	resolution	plan	business	federal
filed	services	california	act	election	supervisors

Table 8: 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.

	Description	F1
h_1	a covers SFBOS	.92
h_2	a covers votes/policy	.85
h_3	a covers recent policy from SFBOS	.9

Table 9: Accuracy for TF-IDF and Logistic regression classifiers at identifying h_i hidden attributes in the PRM model.

for clues relative to the article publish date I will provide. If you can only determine the year and month, that's OK. Ignore dates of events besides the meeting.

(Article 1) The article was published on: The article was published on: ⟨`article_publish date`⟩ Article Text: ⟨`article_text`⟩ Answer:

B Descriptive Statistics

In this appendix, we start by giving some more detailed statistics of our newsworthiness analysis. Then, we will discuss some data-processing challenges that we faced and overcame. We will discuss how we aligned transcripts with segments of the meetings, and we will discuss in more detail how we found and joined public commenters.

B.1 More Link Analysis

In Figure 5, we show the number of times a policy is presented at each meeting, and find a median of 3 times. This aligns with our understanding of how policy progresses from SFBOS; it is introduced, it must be discussed and then it might pass. In cases where a policy is only discussed 1-2 times, it's more likely that it did not pass.

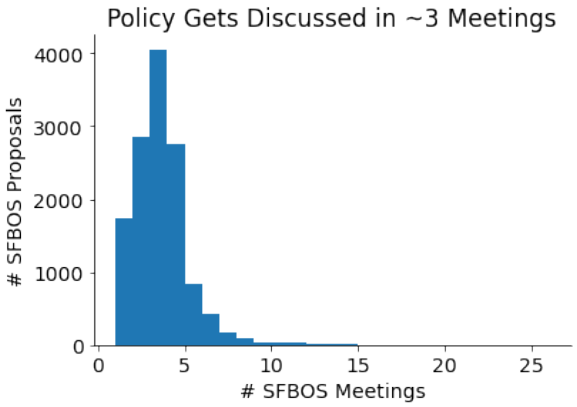

Figure 5: Amount of times policy-items get discussed in SFBOS meetings. Items go from proposals to bills and then get passed.

We also examine the amount of coverage given to policy items. As shown in Figure 6, most policy items are covered between 0-1 times. However, some policy items are covered many, many times. Table 10 shows the bills that have the most coverage. We see a combination of "COVID"-related bills, "nominations" and "transportation"-related bills. While these bills do not materially affect our newsworthiness considerations, since they are more anomalies, they do provide us an opportunity to observe how coverage unfolds over time. In the future, such work could be combined with (Spangher et al., 2022), (Spangher et al., 2023b) and (Spangher et al., 2023a) to provide more of a step-by-step analysis of how coverage of especially newsworthy policies unfolds and grows over time.

B.2 More newsworthiness Analysis

We show in this section additional results from our newsworthiness modeling. In Table 8, we

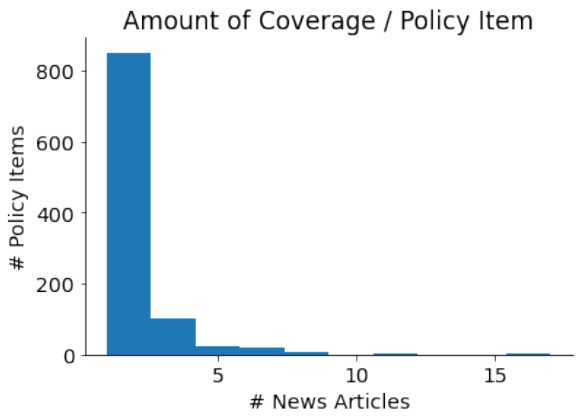

Figure 6: Number of news articles per policy item. Items get covered on average 1.8 times.

performed Latent Dirichlet Allocation (Blei et al., 2003) with the number of topics set to $k = 8$, as we saw a replication of topics after that. We then assigned each policy item to it's most highly-weighted topic, and counted the number of newsworthy and non-newsworthy policy items associated with each topic. We rank-order topics by the top most-newsworthy topics and the least most newsworthy topics, and shows the top 3 and bottom 3, color-coding them by their newsworthiness ranking.

As can be seen, topics like "COVID-19" and "Environment" are more present in newsworthy items, compared with "City Lawsuits". We assign titles to the topics based on a manual assessment.

C Aligning meeting transcripts with video

We collect $3,200$ hours of video data for SFBOS meetings from their hosted service[20] In this section, we describe how we parse the sections of the video that correspond to the policy-items.

Figure 7 shows an example landing page for one SFBOS meeting, held on March 21, 2023. As seen on the left-hand side, the video is shown. On the right-hand side, a nested, hyperlinked agenda document is shown. Capital-letter headers are canonical meeting sections, and are relatively constant across meetings.

Some of the lines are shaded in blue, meaning that they are hyperlinked to a timestamp in the video. In the agenda, any line that starts with a 6-digit code refers to discussions around a policy-item that the SFBOS wishes to pass. Some of the policy-item lines have hyperlinks pertaining to them while others do not.

We manually examined agendas. Many links were missing simply because several policies were discussed together in the agenda. However, others seemed to be missing randomly, leading us to believe that the agenda hyperlinks were incompletely linked.

We wish to reconstruct as completely as we can the time-stamped agenda so that we could get an accurate segmentation for the meeting, so we aim to fill in the missing agenda items. We explore a very simple hypothesis: we assume that meetings were highly organized, and there were consistent phrases used to transition to different agenda items.

So, we seek to classify transitionary phrases. We train a classifier that takes as input a list of diarized transcriptions, t, which each have their own start and end times (t_s, t_e) annotated from the transcription process, and predicts:

$$Y(t) = \begin{cases} 1, & \text{if } \exists \text{ hyperlink } \in (t_s, t_e) \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

We recognize this is noisy, as $Y(t)$ can $= 0$ if both: (A) a segment is not a transition *or* (B) it is simply missing a transition. However, we train a simple classifier using bag-of-words representations for diarized segments and logistic regression, and we achieve f1-score=.86 on held-out data. We analyze the outputs of the classifier and see that it discovers relevant transitional speech, see Table 11 for examples.

Having labeled our transcripts with each diarized segment's likelihood of being transitional speech, we then iterate through each transcript and peg each unlabeled block to either: (a) the next most likely transitional segment or (b) the previous item's segment, if no segment exists above .8 likelihood. In this way, we allow multiple agenda items to be discussed in the same short segment.

This is an exceedingly simple approach that does not consider semantic similarities between the meeting agenda as an approach like, say, dynamic time warping () might. We are confident that our approach could be improved, and maybe in future work improvements could result in additional signal being observed.

[20] Called `Granicus`, which is a service provider used for many local governments.

Top Policies by News Coverage	#
Commending Supervisor London Breed. Resolution Commending and Honoring Supervisor London Breed for her distinguished service as a Supervisor of the City and County of San Francisco.	17
Transportation, Public Works Codes - Unauthorized Powered Scooter Violations, Powered Scooter Share Program. Ordinance amending Division I of the Transportation Code to establish a violation for Powered Scooters that are a part of a Powered Scooter Share Program, to be parked, left standing, or left...	16
Emergency Ordinance - Limiting COVID-19 Impacts through Safe Shelter Options. Emergency ordinance to require the City to secure 8,250 private rooms by April 26, 2020, through service agreements with hotels and motels for use as temporary quarantine facilities for people currently experiencing homele...	16
Nomination Process and Appointment of a Successor Mayor. Motion to take nominations and appoint a successor Mayor to fill a vacancy in the Office of the Mayor, during a Committee of the Whole hearing of the Board of Supervisors of the City and County of San Francisco on January 23, 2018.	14
Mario Woods Remembrance Day - July 22. Resolution declaring July 22 as Mario Woods Remembrance Day in the City and County of San Francisco.	12
Approving Submission of Sales Tax to Support Caltrain Service - November 3, 2020, Election. Resolution approving the Peninsula Corridor Joint Powers Board's placement of a three-county measure to impose a one-eighth of one percent (0.125%) retail transactions and use tax to be used for operating and...	12
Park Code - Golden Gate Park Access and Safety Program - Slow Street Road Closures - Modified Configuration. Ordinance amending the Park Code to adopt the Golden Gate Park Access and Safety Program, which includes restricting private vehicles on certain slow street segments in Golden Gate Park inclu...	11

Table 10: Top SFBOS policies, by the number of times they were covered in the SFChron (#). Includes a mix of office-related, transportation bills and COVID bills.

D Additional Joining Information

When we extract agendas in the SFBOS video viewer, as shown on the righthand side of Figure 7, we find that we are able to retrieve a total of only 10,877 out of 13,089 policies which were listed on the SFBOS legislative calendar website as being discussed during meetings. This is strictly a subset. All policies gathered from video viewer agendas are listed in the SFBOS legislative calendar website.

It's likely that this discrepancy results from policies that were introduced but did not make it past preliminary stages of investigation. For instance, here is an example of a proposal that was listed in the legislative calendar website: https://sfgov.legistar.com/LegislationDetail.aspx?ID=2070276&GUID=D31163A0-D5F8-41E7-AB90-4DECAF9E6693 as having been presented during a SFBOS meeting on 11/18/2014. However, the actions associated with that item, as told in the website, are: "RECEIVED AND ASSIGNED", "REFERRED TO DEPARTMENT", "TRANSFERRED", and "FILED PURSUANT TO RULE 3.41". Here is the video page of the 11/18/2014 meeting: https://sanfrancisco.granicus.com/player/clip/21460. As can be seen, policy number 141197 is not listed in the agenda. That is different from, say, this proposal: https://sfgov.legistar.com/LegislationDetail.aspx?ID=6122328&GUID=B5231DEE-0596-463F-8934-A84468D131ED, which was "CONTINUED" and "HEARD AND FILED", with meeting details associated with each one.

So, a logical explanation is these policies that were never brought to discussion during meetings, thus they do not appear in meeting agendas. However, we cannot discount the possibility that errors were made in creating the agendas. In this case,

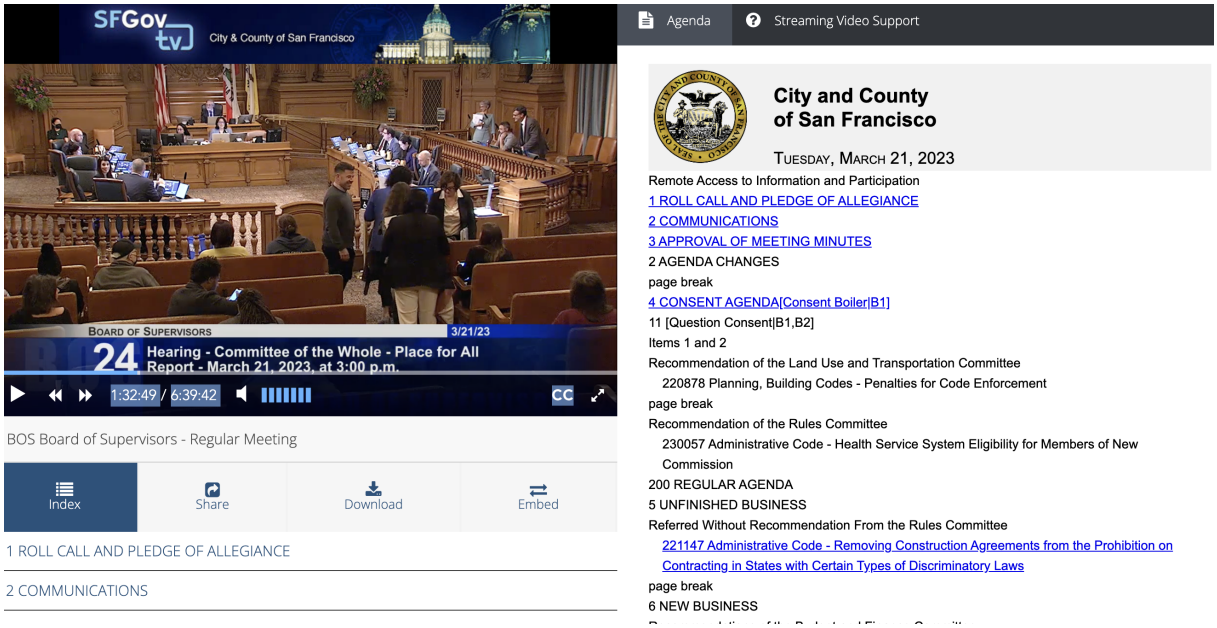

Figure 7: A screen shot of the SFBOS video-hosting website for the URL: `https://sanfrancisco.granicus.com/player/clip/43243`. Seen on the left is the video, from which we are able to download an audio `.mp3` file. On the right is an agenda items from which we can parse timestamps for policy discussions.

we were not able to track policies that were genuinely discussed. Nevertheless, we will refer to these policies as "unpassed-policies".

This affects 164 unpassed-policies that we have identified as being covered by SFChron, out of a total of 1,015 policies, or 16% of newsworthy policies. These unpassed-policies were covered 298 times, out of a total of 2001 articles. We give a sampling of these missing newsworthy unpassed-policies in Table 14. While it's entirely likely that the *fact* that these policies were not discussed *lead* to them being newsworthy, we do not consider such a distinction in our modeling. We leave this to future work.

E Additional Meeting Exploration

Having parsed each agenda and time-pegged each line-item in the agenda, we are able to roll up the time spent in each section-header. Table 13 shows the length of time spent in each section.

As can be seen, the "PUBLIC COMMENT" section occupies a major part of meetings, in terms of the amount of time spent in each meeting, and yet very few policy-proposals are explicitly discussed during this period. We hypothesize that public comment is a potentially newsworthy period in the meeting, where members of the public are able to raise the emotional tenor of a piece of policy (which makes for good news-writing (Uribe and Gunter, 2007)). So, we attempt to join publicly-made comments to entire text of the policy discussion.

As discussed in the main body, we defined "public commenters" as members of the public who only speak during the "PUBLIC COMMENT" section of the meeting. Given timestamps for this section, and speaker diarization, we are able to filter out all speakers besides those that speak during public comment. Next, we use word-overlap between the speaker's speech and the policy text to determine whether the speaker is addressing a particular topic. For the sake of brevity, we assume that each speaker only addresses one comment.

We show in Table **??** some examples of public commenters. As can be seen, they address policy with a personal tenor. However, there are also comments that are rather noisy (e.g. meandering, not on topic, not very focused.) We feel that more work is needed to make the public comment section a usable part of this analysis.

Transitional Phrase
Madam clerk, please call item next item item 33.
Without objection, this resolution is adopted unanimously. Item number nine. Item nine.
Those items are adopted unanimously. Next item, please.
Item number 61. Item 61.
Without objection the resolution is adopted unanimously. Item 44. Item 44
Next item. Item 24.
Without objection, the resolution is adopted unanimously. Next item. Item 52.
Madam clerk, please call item the following items together item 44 45 and 50 and item 54.
Madam Clerk, would you call item 5 please?
Without objection, this resolution is adopted unanimously. Madam Clerk, item number 18.
Madam Clerk, please call the next item.

Table 11: Examples of agenda-item transitions that we identified and then used to parse the agenda..

Policy	Public Comment Speaker (transcribed text)
130049 Resolution supporting Senator Dianne Feinstein's Assault Weapons Regulatory Act of 2013.	Good day, Supervisors. My name is [redacted]. I'd like to start by saying that I do not own any firearms, and I do not oppose sensible gun control legislation. Yet I rise today in opposition to Item...
130151 Resolution opposing the indefinite detention provisions of the National Defense Authorization Act, instructing public agencies to decline requests by Federal agencies acting under detention pow...	Hi, I'm [redacted] from the Libertarian Party of San Francisco and we fully support David Chu's resolution against the detention provisions of the NDAA. Under the guise of the War on Terror, The...
130257 Resolution standing with Muslim and Arab communities in the face of anti-Arab and anti-Muslim bus advertisements.	Hello, my name is [redacted] and I'm a staff attorney at the Asian Law Caucus. And I'd like to speak today about the racist advertisements, the anti-Muslim and anti-Arab advertisements that have ...
130425 Resolution authorizing the Department of Public Library to retroactively accept and expend a grant in the amount of up to $750,000 of in-kind gifts, services, and cash monies from the Friends o...	Good afternoon, Supervisors. Stop the corporate rape of the public library. Don't give money to the Friends of the Library. Don't accept money from the Friends of the Library. Before we begin, we shou...
131071 Accept and Expend Grant - Library Programs - Friends of Public Library - Up to $720,000 - FY2013-2014	Good afternoon. I'm [redacted], Executive Director of Library Users Association. I would like to ask the supervisors to have a hearing on library plans and priorities and performance. The library,...

Table 12: Sampling of public comments, mapped to the policies we infer that they are supporting.

Meeting Section	Time (Min)	# Policies	# Speakers
COMMITTEE REPORT	38.1	3 (+/- 5)	5 (+/- 9)
SPECIAL ORDER	29.3	5 (+/- 8)	15 (+/- 22)
PUBLIC COMMENT	23.9	0 (+/- 1)	16 (+/- 13)
CONSENT AGENDA	4.6	3 (+/- 4)	3 (+/- 5)
NEW BUSINESS	3.5	11 (+/- 9)	10 (+/- 10)
REGULAR AGENDA	2.5	5 (+/- 7)	4 (+/- 5)
IMPERATIVE AGENDA	1.8	2 (+/- 3)	3 (+/- 7)
FOR ADOPTION WITHOUT COMMITTEE REFERENCE	1.4	5 (+/- 4)	7 (+/- 9)
UNFINISHED BUSINESS	1.3	3 (+/- 6)	3 (+/- 4)
ROLL CALL	1.0	2 (+/- 3)	13 (+/- 14)
PROPOSED RESOLUTION	0.4	2 (+/- 3)	1 (+/- 3)
COMMUNICATION	0.3	0 (+/- 1)	3 (+/- 4)
APPROVAL OF MEETING MINUTE	0.3	0 (+/- 1)	1 (+/- 1)
AGENDA CHANGE	0.2	0 (+/- 1)	2 (+/- 2)
PROPOSED ORDINANCE	0.2	1 (+/- 2)	2 (+/- 12)
ADJOURNMENT	0.0	1 (+/- 2)	2 (+/- 2)

Table 13: Top-level parts of SFBOS meetings and the average amount of time spent on each one, according to our inferred timestamps. Also shown are the mean # of policy items discussed in each part, on average, as indicated by the agenda, and the mean # of speakers per section, as per diarization.

Sample of Newsworthy Policies that were not found in SFBOS Video Agendas
Committee of the Whole - Standing Briefings Related to the COVID-19 Health Emergency Response on Board Tuesdays at 3:00 p.m.. Motion directing the Clerk of the Board of Supervisors to schedule standing Committee of the Whole hearings every Tuesday that the Board of Supervisors has a regular meeting ...
Hearing - Federal Budget Cuts to Health Care, Immigration Services, Homeless Services, and Services for the LGBTQ Community. Hearing on the federal budget cuts to health care, immigration services for undocumented San Franciscans, services for the LGBTQ community, homeless services, and cuts to serv...
Concurring in the Continuation of the Declaration of a Local Health Emergency - Monkeypox Virus Outbreak. Motion concurring in the continuance of the San Francisco Health Officer's August 1, 2022, Declaration of Local Health Emergency regarding the outbreak of the Monkeypox virus.
Appropriation - Department of Building Inspection Fund to Department of Emergency Management for Tall Building Seismic Safety Project - $250,000. Ordinance appropriating $250,000 of fund balance in the Department of Building Inspection fund to Department of Emergency Management for Tall Building Sei...
Hearing - 2022 Aging and Disability Affordable Housing Needs Assessment Report. Hearing requesting the key findings and recommendations made in the 2022 Aging and Disability Affordable Housing Needs Assessment Report; and requesting the Department of Disability and Aging Service, Mayor's Office on H...

Table 14: Sample of newsworthy policies that were not found in agendas listed on SFBOS video page viewers. We believe that the most were never discussed, but there could be errors in creating the agenda.