A Guided Topic-Noise Model for Short Texts

Rob Churchill
Lisa Singh
Rebecca Ryan
Georgetown University
Washington, DC, United States

Pamela Davis-Kean
University of Michigan
Ann Arbor, MI, United States

ABSTRACT

Researchers using social media data want to understand the discussions occurring in and about their respective fields. These domain experts often turn to topic models to help them see the entire landscape of the conversation, but unsupervised topic models often produce topic sets that miss topics experts expect or want to see. To solve this problem, we propose Guided Topic-Noise Model (GTM), a semi-supervised topic model designed with large domain-specific social media data sets in mind. The input to GTM is a set of topics that are of interest to the user and a small number of words or phrases that belong to those topics. These seed topics are used to guide the topic generation process, and can be augmented interactively, expanding the seed word list as the model provides new relevant words for different topics. GTM uses a novel initialization and a new sampling algorithm called Generalized Polya Urn (GPU) seed word sampling to produce a topic set that includes expanded seed topics, as well as new unsupervised topics. We demonstrate the robustness of GTM on open-ended responses from a public opinion survey and four domain-specific Twitter data sets .

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; **Natural language processing**; **Machine learning**.

KEYWORDS

topic modeling, social media, semi-supervised topic model, guided topic model, topic-noise model, seed topics

ACM Reference Format:
Rob Churchill, Lisa Singh, Rebecca Ryan, and Pamela Davis-Kean. 2022. A Guided Topic-Noise Model for Short Texts. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3485447.3512007

1 INTRODUCTION

Researchers across disciplines use publicly available social media data to try to understand conversations taking place about the most important topics of the day. These researchers are usually experts in their field, and fall into one of two groups: 1) they have no knowledge of the types of conversations taking place, or 2) they have a sense of the topics relevant to their research that are most likely to appear in public conversations or that are most important to investigate based on theories developed within their disciplines. Unfortunately, no one topic model is well-designed for both of these tasks. For the first task, it is typical to use an unsupervised topic model. For the second task, unsupervised models may not be as fruitful since they do not produce a set of topics that includes the topics researchers expect, or want to see. This typically occurs when a topic is of interest to a researcher, but it is less prevalent or less cohesive than other topics in the data set. In the case of the 2020 United States Presidential Election, topics specific to presidential policies such as immigration, gun control, and the economy were outshone by other salient topics such as Covid-19, racial tensions, and foreign interference in the democratic process. Domain experts know that topics such as economic issues exist, but unsupervised topic models are incapable of discovering them due to noise and low topic frequency relative to more prominent topics. In these scenarios, researchers may consider creating completely manual topics or supervised topic models. While both of these options can produce more targeted topics, one of our goals is to reduce the amount of work a research team must do in order to produce coherent, research-driven topics. Therefore, for this scenario, we consider semi-supervised topic modeling.

Some semi-supervised topic models that allow users to provide 'seed words' as guidance for topic models have been proposed [1, 14, 17, 23]. However, they have not been designed for social media, where noise, including domain-specific flood words, are prevalent and short documents increase the sparsity of the data.

To help researchers (or other users) attain high quality topics from social media data (or short posts/open-ended responses in general) for specific research domains, we propose Guided Topic-Noise Model (GTM). The input to GTM is a set of topics that are of interest to the researcher and a small number of words or phrases that belong to those topics. These seed topics are used to guide the topic generation process, and can be used interactively, expanding the seed word list as the model provides new relevant words for different topics (see Figure 1). This model is intended for the expansion and interactive verification of known or suspected topics. If the seed topics do not exist at all in the data set, it is likely that they will (and should) disappear from the final topic set. However, if they do exist, we want to include them and to augment them with other relevant topic words. To achieve this, we use a combination of informed model initialization, selective oversampling, and noise filtering. Finally, our approach generates additional topics that were not seeded, perhaps missed by researchers. Researchers can then evaluate the full list of topics, and iteratively adjust them as needed.

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France.

ACM ISBN 978-1-4503-9096-5/22/04…$15.00
https://doi.org/10.1145/3485447.3512007

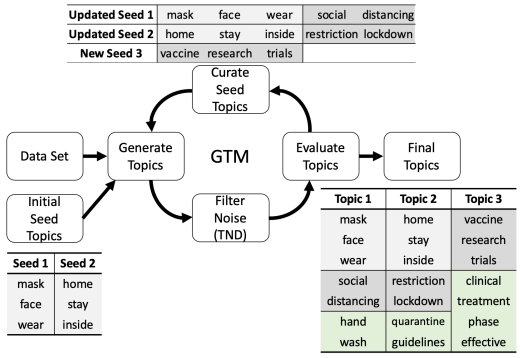


Figure 1: GTM Flow Chart. Words are shaded differently by each round they are added to the topic.

To summarize, the main contributions of this paper are as follows. 1) We propose a new semi-supervised topic model, the Guided Topic-Noise Model (GTM), that allows users to provide guidance to the topic model in the form of seed topics, and through human-model interaction, produce topics related to the seeds, as well as other coherent topics. 2) GTM incorporates a novel sampling strategy that oversamples seed words within a topic, thereby allowing both lower frequency seeded topics and higher frequency seeded topics to coexist. 3) We conduct extensive empirical analysis (both quantitative and qualitative) using four Twitter data sets and an open-ended survey data set, and show the ability of Guided Topic-Noise Model to produce more complete, accurate topic sets. 4) We make our GTM code and evaluation code used in our experiments available to the research community.[1]

2 RELATED LITERATURE

Unsupervised and Supervised Topic Models. While many unsupervised topic models exist (see [7] for a survey), the most widely used unsupervised topic model is Latent Dirichlet Allocation (LDA) [3]. LDA uses a bag-of-words model to find the parameters of the topic/term distribution that maximize the likelihood of documents in the data set over k topics. More recently, topic models have been adapted to incorporate word embedding. For example, GPUDMM draws a set of related words from the embedding space to sample alongside each observed word [20].

Many topic models have been proposed specifically to handle social media data, graph-based models [5, 6, 10], generative models [26, 27, 31], and dimensionality reduction models [32]. Churchill and Singh recently proposed a Topic Noise Discriminator (TND) for generating a new class of topic-noise models for social media data sets [9]. TND works by modeling two distributions, a topic-word and noise distribution, simultaneously. The authors use TND in an ensemble with LDA, and refer to this as Noiseless LDA (NLDA). They show that topic quality across different data sets is more consistent with NLDA. In our experiments, we test against three unsupervised models, LDA, GPUDMM, and NLDA,

Supervised models use manually labeled documents to learn topics [2, 28]. Our proposed model works in the other direction. Instead of labeling documents, GTM uses a small number of seed words assigned to topics to guide the topic generation process by encouraging those seed words to remain close to each other.

Semi-Supervised Topic Models. There have been a handful of semi-supervised topic models proposed in recent years, some with specific applications (learning from product reviews or images) [21][33][30].

Interactive Topic Model (ITM) allows users to mold topics in an iterative manner, interpreting topics and altering the model as it runs [13, 15, 16, 19, 29]. Andrzejewski et al. proposed a model that allows users to encode pairs of words that should (Must-links) and should not (Cannot-links) be placed together in a topic [1]. To accomplish this, the authors encoded the topic-word distribution as a set of Dirichlet tree distributions, called a Dirichlet Forest (DF), to produce the model DF-LDA. DF-LDA was extended by Kobayashi et al. to allow for more complex logical expressions than Must-link and Cannot-link [18]. Expressions such as *and, or, and negation* were added to allow for easier linking of groups of words. One weakness of DF-LDA is that in order to model increasingly large numbers of constraints (Must-links, Cannot-links, and other expressions), the number of Dirichlet Forest models must be increased substantially. In order to translate a set of seed topics into constraints, dozens of Must-links must be added, limiting the ability of DF-LDA to scale.

Similar to GTM, Guided LDA (GLDA) was proposed to allow users to provide guidance to LDA in the form of 'seed words' [17]. While the problem formulation is similar, GLDA models two separate topic-word distributions: one unsupervised and one supervised by the seed topics and then associates the seeded and unseeded topics. Documents are generated as a mix of two distributions, the latter distribution generating only seed words. Our approach models seed topics and other words in the same distribution.

More recently, Gallagher et al. proposed Correlation Explanation (CorEx) Topic Model [14]. Similar to our proposed method, the authors used 'anchor words' for topics. Unlike GTM, topic sets are created by grouping words based on correlation to the anchor words. Category-Name Guided Text Embedding Topic Model (CatE) [23] is a non-generative model that learns the embedding vector of each document in the data set, as well as the vectors of each word using a word embedding space to select similar words to seed words for each topic. The main difference between our proposed method and these approaches is that our model is generative, while these are not. Also, CorEx does not allow for words to be placed in multiple topics with different probabilities. This can result in inflexible topics. We will show that GTM is able to produce higher quality topics that are more in line with what experts expect as a result. We compare our model to semi-supervised methods CorEx, CatE, and GLDA.

3 APPROACH

In this section, we describe the Guided Topic-Noise Model (GTM) in detail, beginning with notation, followed by the model itself.

3.1 Notation

Let D represent a *data set* consisting of M documents or posts, where $D = \{d_0, d_1, ..., d_{M-1}\}$. A *document* d is a collection of N words, where $d = \{w_0, w_1, ..., w_{N-1}\}$. While we focus on words, one could easily generalize this to bigrams and/or phrases.

[1] GTM code can be found here: https://github.com/GU-DataLab/topic-modeling

Algorithm 1 Guided Topic-Noise Model (GTM)

1: **INPUT:** D, S, k, α, β
2: **OUTPUT:** Topic Set T
3: **repeat**
4: T = generate_topics(k, S, D, α, β)
5: T = filter_noise(T, D)
6: *done* = evaluate_topics(T) [Human]
7: **if** !*done* **then**
8: S = curate_seed_topics(T, S, D) [Human]
9: **end if**
10: **until** *done*
11: **return** T

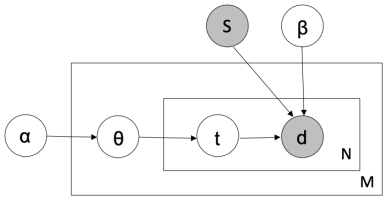


Figure 2: Plate Notation for GTM Generative Phase

A topic t consists of a set of ℓ words, $t = \{w_0, w_1, ..w_{\ell-1}\}$, where the words in t are coherent and interpretable. A topic set T consists of k topics, where $T = \{t_1, t_2, ..t_k\}$. A noise set Ω consists of a set of H words, $\Omega = \{w_1, w_2, ..w_H\}$, where the words in Ω represent noise. A topic set can be initialized using a set of seed topics S. A seed topic s is identical in nature to a topic t, except that it is predefined by the user as opposed to generated by a topic model. A seed topic can also include phrases in the form of ngrams if the user believes they will be more informative.

3.2 Guided Topic-Noise Model (GTM)

Algorithm 1 shows GTM at a high-level. The inputs to GTM are the data set D, the initial seed topics S, and the total number of desired topics k, where $k \geq |S|$. If $k > |S|$, $k - |S|$ unsupervised topics will be approximated alongside the $|S|$ guided topics. GTM begins by generating topics using the seed words S (Section 3.2.1). Noise is then filtered from the topics (Section 3.2.2). The topics are then evaluated by the user (Section 3.2.3), and if necessary, the seed topics are further curated and the process is rerun. The human-driven feedback that we incorporate iteratively is an important facet of GTM. While GTM can produce good topics without human feedback, since we already incorporate human guidance at the instantiation of the model, we give users the chance to improve their topic seeds after topic generation.

3.2.1 The Generative Phase. **Generating Topics with Guidance.** While we could use the seed topics to label documents as belonging to one topic or another from the start, we do not view this as advantageous. Given that the size of the seed topics are expected to be very small compared to the vocabulary, we would be attaching labels to documents with incomplete knowledge, and risk mislabeling documents from the outset, leading to poor topics. Instead, in GTM seed topics are injected into the generative process at the word level, as opposed to the topic or document level (see Figure 2 plate notation). More specifically, the high level generative process of GTM (line 4 of Algorithm 1) is as follows.

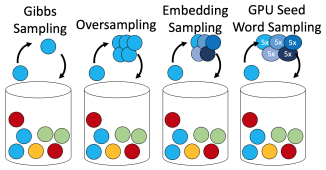


Figure 3: Sampling Schemes for Gibbs Sampling, Oversampling, Embedding Sampling, and GPU Seed Word Sampling.

For $d \in D$:

(1) Draw the number of words N for d.
(2) Draw the topic distribution θ from the Dirichlet distribution, conditioned on the parameter α.
(3) For each word w_i, $0 \leq i < N$:
  (a) Draw a topic t_i from θ.
  (b) If seed topic s_i exists:
    (i) Draw a word w_i based on the probability of w_i given t_i and S and conditioned on the parameter β.
  (c) Otherwise:
    (i) Draw a word w_i based on the probability of w_i given the topic t_i and conditioned on the parameter β.

We note that if the topic being drawn is a seed topic, seed words are given preference.

Generative Model Initialization. When initializing the model, the generative process of many unsupervised models randomly assigns a topic to each document, and then randomly assigns a topic to each word in the document. We do the same for each document and for each word that is not a seed word. For seed words, we assign them to the correct topic based on their seed topic. When $k \geq |S|$, additional unseeded topics can also be generated.

Sampling with Guidance. Figure 3 shows the differences between traditional Gibbs sampling, oversampling [24], embedding sampling (GPUDMM [20]), and our proposed method, GPU seed word sampling. Traditionally, LDA-inspired models use Gibbs sampling, which in each iteration draws a word (a ball in Figure 3), and with respect to the topic distribution of the observed document, reassigns the word to a topic (replaces the ball in Figure 3). Gibbs sampling works well in many regards, especially in unsupervised models where there is no prior knowledge about specific words.

The Generalized Polya Urn topic model [24] alters the Gibbs sampling scheme of LDA to use Generalized Polya Urn (GPU) sampling. The difference between GPU sampling and Gibbs sampling is that instead of observing a word and replacing it in the drawn topic, the observed word is replaced in the drawn topic with multiple copies of itself, oversampling words with low frequency in the data set to increase their probability in their corresponding seed topics. Figure 3 shows the oversampling approach on the middle-left, where one ball is drawn and five are replaced in its stead.

In GPUDMM [20], the authors, instead of oversampling the observed word, used the observed word to find related words in an embedding space, and sampled each of the related words as well as the observed word. Figure 3 shows the embedded sampling scheme where one ball is drawn and it is replaced alongside related words from the embedding space.

Instead of an embedding space containing related words, we are given as guidance something much more valuable. Seed topics, assembled by someone with prior knowledge of the domain, are direct confirmation of word relatedness, as opposed to the implicit relatedness given by embedding spaces. We do not need to use embedding spaces here to improve topic coherence because we already have highly related sets of seed words. Also, embedding spaces are made to be more generic, providing word relations that are most common for the word as opposed to relationships that are domain specific. For example, a general embedding space would not put the words *email* and *scandal* close to each other. However, a researcher studying the 2016 US Presidential election would. We call our sampling algorithm *GPU seed word sampling*. In GPU seed word sampling, when a non-seed word is observed, Gibbs sampling is performed, replacing it with a single copy of itself. However, when a seed word is observed, its whole seed topic is sampled *and placed in the correct topic.* Each seed word is oversampled by the product of a global factor γ and an individual factor δ_i, the inverse document frequency for word w_i. Figure 3 shows GPU seed word sampling on the right, where the seed word is sampled alongside multiple copies of its fellow words from the same seed topic.

By performing GPU seed word sampling on the seed topic words, we are able to maximize the probability of each seed topic regardless of the frequency of the topic in the data. This allows us to generate more complete topics for all seed topics. We are also pushing documents containing a seed word toward the seed topic. While the document will still have a probability of being composed of every topic, we are guaranteeing that it always has a significant probability for the topic of the seed word it contains.

3.2.2 The Noise Filtering Phase. GTM is designed with social media in mind. A common source of frustration for those using topic models on social media is noise. To filter noise from GTM, we propose using it in an ensemble with the Topic Noise Discriminator (TND), as detailed by Churchill and Singh [9]. TND works by modeling two distributions, a topic-word and noise distribution, simultaneously. When generating a document, a word can be drawn either from the topic distribution of the document, or from noise. This approach more accurately captures the random nature of noise in social media data, and produces a robust noise distribution that is implicitly conditioned on the topic-word distribution. To incorporate our generative model with TND, we follow the framework of NLDA [9]. We combine our topic-word distribution θ with the noise distribution Ω of an instance of TND modeled on our data set.[2] For each topic, we decide whether each word should be in the topic or in noise for that specific topic by drawing from a Beta distribution conditioned on the word's probability of being in the topic and the noise distribution. Equation 1 shows how the Beta distribution is calculated for word i on topic t, with respect to Ω. ϕ is a variable that can be used to tune the weight of the noise distribution, and the $skew$ variable is used to tune the weight of the topic-word distribution.

$$Beta\left(\sqrt{\theta_t^i + skew}, \sqrt{\Omega_i(\phi/k)}\right) \quad (1)$$

[2]In Section 2 we noted that DF-LDA was extended to allow for certain words to be excluded from topics altogether [18]. The advantage of using TND is that noise removal is determined probabilistically, with no user input required.

Data Set	# Docs	Vocab
Survey School	2,697	2,781
Gun Violence	145,602	86,913
Covid-19	620,297	432,555
Election 2020	1,226,369	345,603
BLM	1,300,340	397,728

Table 1: Data Set Sizes

3.2.3 The Human Phases: Topic Evaluation and Seed Topic Curation. While not strictly necessary (a user could provide no feedback and output the first round of topics generated and filtered), we encourage users to evaluate and adjust the approximated topics.

Topic Evaluation. Since we assume that the users of GTM will have some knowledge of the domain, users should find it easy to evaluate the quality of topics. Evaluation of topics should include identifying seed words in each topic, identifying potential new topic words, and identifying noise words. The goal of iterative topic evaluation is to see more new topic words and less noise.

Seed Topic Curation. If users decide that they want to further hone the semi-supervised topics, they can alter their seed words for the next iteration of topic generation. New seed topics can be added to the list if users see potential new topics of interest in the $k - |S|$ unsupervised topics. For existing seed topics, users can add new seed words, likely that have appeared in the topic set. In this phase, it is a good idea to remove words that seem too general. Often, these general words are relatively harmless to topic coherence. However, removing them can sometimes reveal a part of a seed topic that was hidden by the inclusion of such a general word.

At the end of this process, we should be left with topics that greatly improve when compared to the seed topics.

4 EMPIRICAL EVALUATION

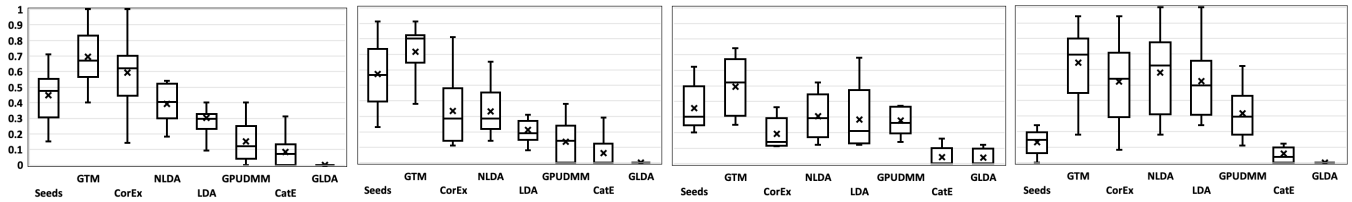
In this section, we present our empirical evaluation of Guided Topic-Noise Model (GTM). We begin by describing our experimental setup. We then present our quantitative and qualitative evaluations.

4.1 Experiment Setup

Baseline Models. We compare our model to GLDA [17], CorEx [14], and CatE [23] since they are semi-supervised at the word-level, and begin with seed words, similar to GTM.[3] We also compare to unsupervised models to understand the ability of GTM to capture topics that might otherwise be overlooked or incoherent. With this in mind, we compare GTM to Latent Dirichlet Allocation (LDA) [3][4], Generalized Polya Urn Dirichlet Multinomial Mixture (GPUDMM) [20], and Noiseless Latent Dirichlet Allocation (NLDA) [9]. We include LDA because of its ubiquity and consistent performance across many types of data, GPUDMM because of its theoretical contributions to GTM and its uses in social media data, and NLDA because it filters noise and is the state of the art for social media.

Data Sets. We evaluate model performance on four unique Twitter data sets in English collected using the Twitter API, and an open-ended response survey data set. For each data set, domain experts were asked to identify seed topics and a full set of topics.

[3]We do not compare to DF-LDA [18], because it does not scale to larger data sets.
[4]Specifically the MALLET implementation of LDA [22]



(a) Election 2020 (b) BLM (c) Gun Violence (d) Covid-19

Figure 4: Topic Recall Box Plots on Twitter Data Sets. x denotes mean recall.

The first Twitter data set contains tweets about the 2020 US Presidential Election (Election 2020)between August 15 and November 15, 2020. It was collected using hashtags about the election, e.g. #biden and #trump. For the Election domain, two male professors and one female Ph.D. in social science served as our experts. The second data set contains tweets about the Covid-19 Pandemic (Covid-19), collected between April 1 and December 31, 2020. This data set was collected using hashtags related to Covid-19, e.g. #coronavirus and #covid. For the Covid-19 domain, our experts were two female public policy professors. The third contains tweets about the BlackLivesMatter (BLM) movement, collected between May 15 and July 15, 2020 using the hashtag #BlackLivesMatter. We focus on these two months because they include when George Floyd was killed and the protests that followed. For the BLM domain, two female law professors were as our experts. The last contains tweets about gun violence, collected in 2017, a period with multiple mass shootings. Again we use keywords and hashtags related to conversation about gun violence and gun rights to collect the data. One female professor and one male researcher served as our experts. In total, we worked with nine domain experts.

We also have survey data that included multiple open-ended response and was given to a probability-based web panel containing 9,544 U.S. adults [11]. Responses were collected between April and June 2021. The question that we focus on here asked about challenges related to children learning from home during the pandemic. For this question, there were approximately 2,700 responses. Table 1 contains statistics about each data set after preprocessing.

Data Preprocessing. Churchill and Singh [8] show that text preprocessing can improve topic model performance. We use the following elementary pattern-based preprocessing for all the data sets: lowercase, remove stopwords, remove punctuation. For the Twitter data sets, we remove URLs, deleted posts and user mentions.

Model Parameters. We conduct a sensitivity analysis for GTM, testing a range of parameter settings to determine which parameters were the most stable. Because of space limitations, we only present the results for the best performing settings.[5] We determined the best parameter values by comparing topic diversity, recall, and entropy on the smallest data set, the survey data. For GTM, the best parameters were $k = 5|S|, \alpha = 0.1, \beta = 0.01, \gamma = 1$. We experimented with $k = x|S|$, where $x \in (1, 10)$, and $\gamma = \{1, 2, 5, 10, 20, 50, 100\}$. For TND, we used the suggested parameters provided by the authors [9], with $k = 5|S|$ and $\mu = 0$. For ensembling GTM and TND to filter noise, we set $\phi = 10$. For LDA, $\alpha = 0.1$, $\beta = 0.01$, and for GPUDMM, $\alpha = 0.1$, $\beta = 0.1$, were the best parameter settings. For GPUDMM, we used GloVe Twitter word embeddings [25] with 50 dimensions. For both models, we used the same k value as was used for GTM to provide the best comparability. For GLDA, CatE, and CorEx, we used our seed topics, with $k = 50$.[6]

Topic model implementations. GTM is implemented using MALLET [22], a parallelized Java implementation of generative models, including LDA. All of the models we run are for the same number of iterations on the same number of threads on the same machine (24 2.2GHz vCPUs, 40 GB memory). [7]

4.2 Quantitative Evaluation

Evaluation Metrics. We are interested in models that can find a diverse set of useful topics. Topic diversity is the fraction of unique words in the top 20 words of all topics in a topic set [12]. A model with high diversity partitions words into topics with little overlap.

In order to test GTM, we asked our experts for seeds topics to guide the model. We used the seed topics as the seeds for GTM, CorEx, CatE, and GLDA. We also asked for extended topics, curated by the experts, for evaluation. We use these expert curated topics as our set of ground truth topics. To measure the ability of models to produce topics similar to the ground truth, we employ *topic recall*. Topic recall is the fraction of words from the full topic that an approximated topic recovers in the top x words. For our experiments, we used the top 50 words per topic. In order to ensure fairness and not punish models for having topics that do not fall into the final topic set, for each full topic, we count only the approximated topic with the highest recall.

We also seek to measure the improvement offered by using GTM to augment seed topics. We had experts choose the words generated by GTM under the guidance of the seed topics that they believe truly belong to the underlying topic. Topic improvement is the percent increase in topic size after augmenting the seed topics using GTM.

Finally, we are interested in understanding the amount of information present in our topic set. There are a number of different ways to measure this. A simple way is to look at the probability of every ground truth word $p(w)$ in every topic and sum the probabilities. In our case, ranking is more important than probability, where the rank of a word w in topic t is the position of w in t based on $p(w)$ given t. So, the most probable word to appear in t is rank

[5] We found that k was a more sensitive parameter than in normal topic models. Having enough extra unsupervised topics allows for cleaner seed topics. γ was less sensitive.

[6] CatE was designed to be given one word per topic. We tested using different 'best' words per seed topic and found that the results were better using the entire seed topic.

[7] Our runtime is slower than LDA and NLDA due to the more complicated sampling scheme required, but is still faster than CorEx, CatE, GLDA, and GPUDMM.

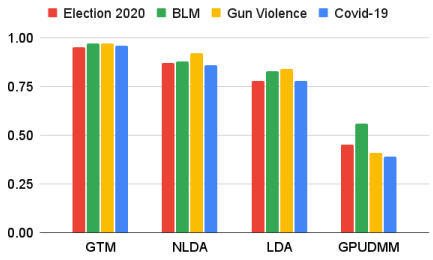


Figure 5: Topic diversity comparison.

$R = 1$, and the least probable is rank $R = |V|$. The problem with this simple ranking approach is that the penalty for having a poorly ranked word is much higher than the reward for having a highly ranked word. In reality, if a topic word is ranked as the 1000th or the 3000th word, both are bad and overwhelm the information calculation. Therefore, to adjust for that, we propose using a variant of Shannon's entropy, *topic entropy*, that uses the number of digits in the ranking R of each word as a proxy for the number of bits needed to represent the topic word: $E = \sum_1^{|t|} c \times log_{10} R[(p(w), t])$, where $R[p(w), t]$ is the rank of each ground truth topic word for each topic and c is a constant. We suggest using higher c-values to further distinguish between rankings when $|V|$ is large. We set $c = 1$ for our experiments. A low topic entropy score indicates a set of topics that are more compressed since they contain more information in fewer bits.

Comparing GTM to Unsupervised Models. An important distinction between GTM and unsupervised models is its adherence to the seed topics that guide it. But how well does it adhere, and how coherent and unique are the generated topics?To address these questions, we compute topic recall, topic entropy, and diversity.

Figure 4 displays a box plot of topic recall for each model on each Twitter data set. In this section, we focus on the relationship between GTM and NLDA, LDA, and GPUDMM (the unsupervised models). In the Election 2020, BLM, and Gun Violence data sets, there is a stark contrast between GTM and the unsupervised models. The mean recall for GTM is nearly twice as high as any unsupervised model. In the Covid-19 data set, the unsupervised models perform better, with NLDA's average recall at 58%, still lower than GTM (64%). The unsupervised models are not able to provide the same level of recall that GTM is capable of since they are not given any guidance. While not surprising, it is important to verify. We also pause to note that in all cases, the seeds alone do not sufficiently describe any topic. Additional meaningful words are added to every topic by GTM across all the data sets, highlighting the difficulty for researchers to identify all the meaningful words apriori.

Figure 6 displays the topic entropy of GTM and all of the baseline models on the Twitter data sets (a lower score is better). As we can see, GTM is consistently the best across each data set, from 7 to 22% better than the next best model. While LDA, NLDA, and GPUDMM each perform well on one or two data sets, none perform as well or consistently as GTM.This result shows the positive impact of using semi-supervision to detect topics that experts care about.

	Survey	Election	Covid-19	GV	BLM
GTM	**34% (3.2)**	**176% (8.8)**	**402% (25.5)**	**114% (10.2)**	**108% (9.4)**
CorEx	8% (0.67)	120% (6.0)	220% (20.9)	69% (6.6)	10% (0.8)
CatE	15% (0.77)	82% (4.1)	47% (3.2)	38% (4.0)	29% (2.5)
GLDA	6% (0.67)	0% (0)	0% (0)	0% (0)	0% (0)

Table 2: Topic Improvement. Average % increase (average total increase) in size of topics after augmentation.

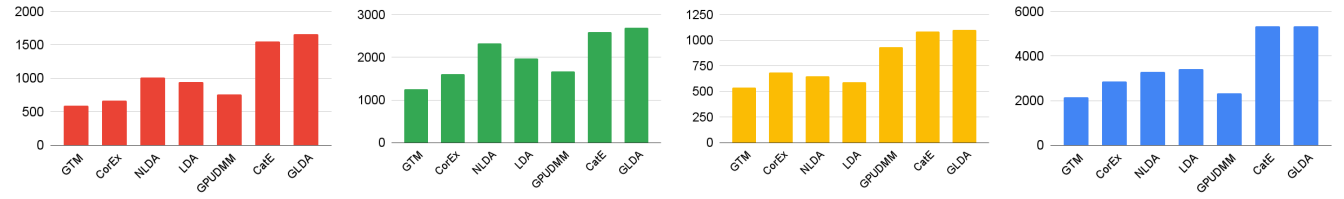
Figure 5 shows a histogram of the topic diversity scores for GTM and each unsupervised model on each Twitter data set.[8] What we see here is that GTM scores higher in diversity than NLDA, LDA, and GPUDMM. This is a consequence of using TND [9] for noise filtering, and also a mark of the helpfulness of seed topics. NLDA also uses TND for noise filtering, and improves on the diversity of LDA by 5-10% depending on the data set, but GTM improves on NLDA's diversity by 5-10% because of the guidance of the seeds. This high diversity means that GTM is both accurate and precise, recovering more topic words than other models, but also without repeating topics. Surprisingly, considering its relative success on topic ranking, GPUDMM has very low diversity across the board.

Comparing GTM to Semi-supervised Models. We now focus on comparing GTM to similar semi-supervised models. We compare to GLDA, CatE, and CorEx given that their approaches are most similar to ours. Beginning with recall in Figure 4, CorEx came close to GTM in the Election 2020 and Covid-19 data sets, but was significantly lower in the BLM and Gun Violence data sets. GTM had better recall for all of its topics and had less spread across all of them. GTM's improvement over CorEx ranged from about 10% on the Election 2020 data set to nearly 40% on the BLM data set. CatE and GLDA performed surprisingly poorly. GLDA seemed to suffer heavily from noise inundation, with all of its topics containing words that had little to do with the seed topics, but which had high frequencies. CatE also suffers from noise penetration.

Returning to Figure 6, we now evaluate the topic entropy of the semi-supervised models. CorEx produces topics with entropy close to that of GTM, but is beaten by GPUDMM in the Covid-19 data set and by NLDA and LDA in the Gun Violence data set. CatE and GLDA both have trouble dealing with noise words in the Twitter data sets, and for that reason have very high entropy, meaning that they are not able to accurately rank words in their optimal topics.

To further compare the semi-supervised models, we use topic improvement as shown in Table 2. We compare the average topic improvement of GTM, CorEx, CatE, and GLDA on each data set. The percent improvement is followed by the average number of words added per topic in parentheses. The improvement numbers do not account for new topics that were added. In the case of the survey data set, the improvements are much more modest because experts curated longer seed word lists than with the Twitter data sets. Even with the larger seed word list, GTM still resulted in an improvement of 34%, with an average of 3.2 words added per topic. Using CorEx and GLDA, only 0.67 words were added per topic, for an increase of 8% and 6%, respectively. The percent increase varies because the size of seed topics varies. So, if one word was added to a topic of size 4, that would relate to an increase of 25%, while adding one word to a topic of size 10 would only be a 10% increase.

[8] CorEx and CatE are not pictured because by their models' definitions, words are forced into only one topic. GLDA's diversity was poorer than that of GPUDMM due to noise. So we focus on the comparison to unsupervised models.



(a) Election 2020 (b) BLM (c) Gun Violence (d) Covid-19

Figure 6: Topic Entropy Histograms on Twitter Data Sets.

Masks	Origins	Lockdown	Cases	School	Economy	Vaccines & Treatments	Hospitals	Testing	General US Government	General Pandemic	Community Support	Personal Stories	Information Ecosystem
mask	china	lockdown	cases	kids	business	vaccine	care	positive	government	pandemic	support	wife	hoax
masks	chinese	home	deaths	online	economy	research	hospital	testing	govt	spread	community	lost	msnbc
face	wuhan	stay	total	education	impact	trials	workers	tested	relief	outbreak	provide	remember	twitter
wear	chinas	inside	death	students	economic	treatment	medical	infected	fauci	fear	response	christmas	potus
social	communist	safe	confirmed	school	jobs	hydroxychloroquine	ventilators	tests	congress	insurance	local	loved	president
protect	hong	stayhome	number	teachers	market	effective	healthcare	test	governor	stop	needed	passed	lies
wearing	kong	quarantine	reported	children	restaurants	drug	hospitals	negative	federal	global	families	sharing	truth
distancing	wuhancoronavirus	measures	toll	schools	small	trial	beds	samples	democrats	disaster	sign	worth	foxnews
hands	party	restrictions	reports	open	businesses	developed	patient	results	bill	fight	proud	heres	american
hand	realjameswoods	staysafe	active	parents	industry	clinical	frontline	rapid	house	slow	emergency	weve	whitehouse
wash	body	rules	highest	learning	companies	researchers	staff	kits	court	close	efforts	alexberenson	called
covering	political	guidelines	rises	reopen	technology	potential	blooddonorsin	district	senate	gavinnewsom	deliver	busy	tombx7m
hygiene	evidence	stayathome	update	reopening	digital	study	recovered	antibody	situation	prevent	communities	happen	trump2020
physical	found	socialdistancing	recovered	child	future	sarscov2	component	persons	give	americas	vulnerable	happened	maga
distance	virus	staying	recoveries	student	sector	phase	type	breaking	current	continues	team	imagine	trumpvirus
practice	forced	healthy	spike	university	tech	remdesivir	delhi	reveals	billion	amid	helping	majority	blame
prevent	created	shutdown	tally	exams	recovery	pfizer	plasma	telangana	gates	continue	continue	reading	biden
cdcgov	policy	place	bringing	closed	crisis	antibodies	blood	odisha	money	follow	officials	lots	realdonaldtrump
mandatory	gordongchang	order	lakh	decision	employees	develop	meet	pradesh	request	insurance	access	sadly	america
public	clear	indiafightscorona	24hrs	return	innovation	early	personal	tamil	jim_jordan	operations	initiative	heard	joebiden

Figure 7: Covid-19 Twitter Topics. Top 20 words: seed words (gray), new topic words (green), non-topic words (orange).

A lower average percentage increase but the same average number of words added per topic means that more words were added to large topics than to small ones, which we see as having less impact on the overall topic set. In the Twitter data sets, the seed topics ranged from five to fifteen seed words per topic. In this case, the average improvement was over 100% for each data set, peaking at over 400% on the Covid-19 Twitter data set, which had an average of seven seed words per topic. CorEx and CatE did not add as many words per topic for any Twitter data set. CorEx was generally better than CatE and GLDA because of their inability to deal with noise.

Using GTM to augment and identify new topics can help save valuable time for domain experts who want to understand their data quickly, but who already have partial knowledge of what exists.

4.3 Qualitative Evaluation

Improving Expert Understanding of Data Sets. One important aspect of GTM is its ability to create highly relative topics around the seeds while at the same time unveiling topics that experts may have missed. We demonstrate this using the Twitter Covid-19 data set. Domain experts identified 11 seed topics in the data. After reviewing the GTM topics, they discovered that they had actually missed three topics – community support, personal stories, and information ecosystem. Figure 7 shows the top 20 words for each topic guided by a seed, as well as the three new topics. Seed words are highlighted in gray on top of each column, new topic words are highlighted in green in the middle, and non-topic words are highlighted in red at the bottom. The three new topics are on the right side, separated from the seeds by a vertical line. As we can see, nearly all the words in the first seven topics (except for the Origins topic) are seed or topic words. The large green band emphasizes the value of a guided model for improving topics of interest to researchers and also for identifying new topics.

Covid-19 Remote Schooling Challenges Survey. Figure 8 shows the final topics decided on by the experts after using GTM to augment their seed topics. Like Figure 7, the seed words are at the top of each column in gray, and the new topic words are in green. First, we can see on the right side of the bottom row that the researchers found a new topic, about working parents, in the topic set. These new words did not count toward the 34% improvement since the topic was not a seed topic, but together they are another important topic underlying the survey responses. In the other topics, we can see many highly informative words and phrases, such as 'one-on-one', 'weight', 'gain', and 'poor grades'. These added words all add to the context of their respective topics, making for higher quality, more interpretable topics.

The domain experts who curated these topics had two goals in doing so. The first was to get a more complete, descriptive set of topics to convey the main concerns and opinions of respondents to the survey. The second was to quantify those concerns and opinions by classifying responses using the topics. Because manually-curated topics consist of only a small portion of the vocabulary, we do not expect every single response to be classified under the topic set. However, we want to maximize the number that are classified, and by adding more words to each topic we hope to improve that number. The manually curated seed topics alone were able to classify 70% of the responses. Using GTM to augment the researchers' seed topics, we were able to increase that figure to 77.3%, an improvement of over 7%.

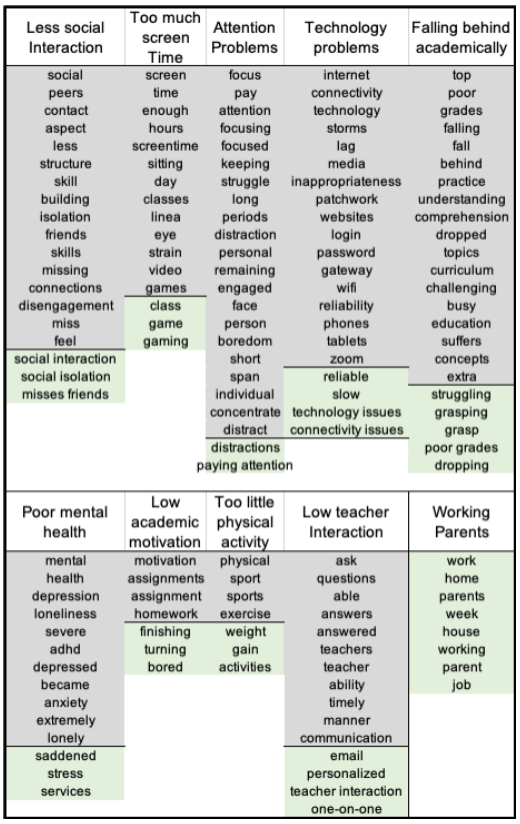


Figure 8: Home-schooling Survey Topics. Seed words (gray), new topic words(green).

GTM and CorEx Qualitative Comparison. In our final qualitative analysis, we chose a seed topic from the Election 2020, BLM, and Gun Violence data sets and show the resulting topics produced by GTM and CorEx. Figure 9 shows the top 20 words for three topics as they were found by GTM and CorEx. The three topics are Mail-In Voting (Election 2020), Victims (BLM), and Gun Ownership (Gun Violence). The words are ordered by topic and non-topic words.

The Election 2020 Mail-In Voting topic was found with high accuracy by both models. Each model produces topics with 15 words chosen from the top 20. The other five words in CorEx are clearly noise words, including four user handles and the generic word 'says'. On the other hand, the five words not chosen by the experts from GTM are more likely less relevant topic words that do not add enough context to be added to the final topic. Words such as 'joebidens', 'changed', and 'stake' may be related to voting.

The differences in the BLM Victims topic are more stark. In GTM, nearly all words refer to victims of police violence. However, in CorEx, the topic seems to have been mixed together with a different topic related to signing petitions (as well as and noise words like user handles). Bowman-Williams et al. showed how, in the case of the BlackLivesMatter domain, millions of tweets were posted focusing on the victims [4], and as such a topic about victims should be easily detectable. In this case, the overlap of the underlying petition topic led to the failure of CorEx to isolate topics related to

Election 2020 - Mail-In Voting		BLM - Victims		GV - Gun Ownership	
GTM	CorEx	GTM	CorEx	GTM	CorEx
ballot	ballots	george	elijahmcclain	amendment	stopthenra
early	court	georgefloyd	sign	fight	concealed
voter	voter	floyd	thread	rights	carry
fraud	ballot	breonna	petitions	owner	reciprocity
ballots	fraud	taylor	petition	regulated	hr38
state	early	tamir	byersfilms	constitution	opposeccr
mail	mail	rice	help	arms	stopccr
county	state	rayshardbrooks	tpwkhollands	2ndamendment	nras
illegal	supreme	eric	nonblack	militia	antinra
absentee	votes	ahmaud	ways	bear	arming
send	voting	arbery	signatures	carrying	credomobile
mailin	absentee	brooks	shawnwasabi	2017	bill
register	mailin	remember	kiesdaya	hate	sign
polling	electoral	rayshard	ardentlyswift	wearorange	dream
investigation	courts	garner	support	pledge	must
changed	tomfitton	martin	seconds	written	priority
entire	loudobbs	elijah	sha_elise24	society	fetus
find	sidneypowell1	class	danitycafe	awareness	senate
joebidens	jsolomonreports	mefeater	educational	racism	block
stake	says	wouldve	featuring	biases	abortion

Figure 9: Topic Comparison. New topic words (green), and non-topic words (orange).

the seed words. Only one victim name is in the top 20 words. We note that LDA, NLDA, and GTM all detected a petition topic. In the case of GTM, it was a topic that was not a seeded topics. This example topic is representative of the performance of CorEx on the BLM data set as a whole. Because it relies on word correlations, correlations that cross topics can lead to muddled topics.

In the final topic here, the Gun Ownership topic, we see another interesting development. GTM and CorEx both find words relevant to the seed topics, but the topics are distinctly different. CorEx's topic refers specifically to the National Rifle Association (NRA), while the GTM topic focuses more on Second amendment rights. Both add in more general terms, in the case of CorEx, some that are not relevant to gun violence, e.g. abortion and fetus.

5 CONCLUSIONS AND FUTURE WORK

In this paper we proposed a new semi-supervised topic model that allows for supervision based on identifying seed topics, Guided Topic-Noise Model (GTM). We showed how users can provide seed topics with a small set of seed words to guide GTM toward a more complete set of topics. GTM does this through interactivity, clever model initialization, and a new sampling algorithm that takes full advantage of semi-supervision. We combine this with noise filtering to further improve the topic diversity of the final topic set.

We demonstrated the effectiveness of Guided Topic-Noise Model through extensive experiments using four novel domain-specific Twitter data sets and a data set containing survey responses about the Covid pandemic. We used quantitative and qualitative analysis to show that GTM is a novel, effective semi-supervised approach capable of producing rich topics that align with the seed topics more than other unsupervised and semi-supervised approaches. Finally, we share our models and evaluation code on GitHub to further topic modeling research, as well as any other research that would benefit from guided topics.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation grant numbers #1934925 and #1934494 and by the Massive Data Institute (MDI) and McCourt Impacts at Georgetown University. We would like to thank our funders. We would also like to thank the Mosaic Project and SSRS for access to the Covid-19 Survey data set.

REFERENCES

[1] David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *International Conference on Machine Learning*. 25–32.

[2] David M Blei and Jon D McAuliffe. 2010. Supervised topic models. *arXiv preprint arXiv:1003.0783* (2010).

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[4] Jamillah Bowman Williams, Naomi Mezey, and Lisa Singh. 2021. #BlackLivesMatter: Getting from Contemporary Social Movements to Structural Change. *California Law Review Online* 12 (2021).

[5] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. 2010. Emerging Topic Detection on Twitter Based on Temporal and Social Terms Evaluation. In *ACM KDD Workshop on Multimedia Data Mining*. 1–10.

[6] Rob Churchill and Lisa Singh. 2020. Percolation-based topic modeling for tweets. In *KDD Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)*.

[7] Rob Churchill and Lisa Singh. 2021. The Evolution of Topic Modeling. *ACM Computing Surveys (CSUR)* (2021).

[8] Rob Churchill and Lisa Singh. 2021. textPrep: A Text Preprocessing Toolkit for Topic Modeling on Social Media Data. In *International Conference on Data Science, Technology, and Applications (DATA)*.

[9] Rob Churchill and Lisa Singh. 2021. Topic-Noise Models: Modeling Topic and Noise Distributions in Social Media Post Collections. In *International Conference on Data Mining (ICDM)*. 71–80.

[10] Rob Churchill, Lisa Singh, and Christo Kirov. 2018. A Temporal Topic Model for Noisy Mediums. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. 42–53.

[11] P. Davis-Kean, R. Ryan, L. Singh, and N. Waters. 2021. Groundhog day: Homeschooling in the time of Covid-19. *MOSAIC Data Brief: Measuring Online Social Attitudes and Information Collaborative* (10 2021).

[12] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2019. The Dynamic Embedded Topic Model. *CoRR* abs/1907.05545 (2019). arXiv:1907.05545 http://arxiv.org/abs/1907.05545

[13] Jacob Eisenstein, Duen Horng Chau, Aniket Kittur, and Eric Xing. 2012. TopicViz: Interactive topic exploration in document collections. In *Extended Abstracts on Human Factors in Computing Systems*. 2177–2182.

[14] Ryan J Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. 2017. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics* 5 (2017), 529–542.

[15] Enamul Hoque and Giuseppe Carenini. 2015. Convisit: Interactive topic modeling for exploring asynchronous online conversations. In *International Conference on Intelligent User Interfaces*. 169–180.

[16] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive topic modeling. *Machine learning* 95, 3 (2014), 423–469.

[17] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. 204–213.

[18] Hayato Kobayashi, Hiromi Wakaki, Tomohiro Yamasaki, and Masaru Suzuki. 2011. Topic Models with Logical Constraints on Words. In *Workshop on Robust Unsupervised and Semisupervised Methods in Natural Language Processing*. 33–40.

[19] Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies* 105 (2017).

[20] Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016. Topic Modeling for Short Texts with Auxiliary Word Embeddings. In *Conference on Research and Development in Information Retrieval (SIGIR)*. 165–174.

[21] Fangtao Li, Sheng Wang, Shenghua Liu, and Ming Zhang. 2014. Suit: A supervised user-item based topic model for sentiment analysis. In *AAAI Conference on Artificial Intelligence*.

[22] Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. (2002).

[23] Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang, and Jiawei Han. 2020. Discriminative topic mining via category-name guided text embedding. In *The Web Conference (WWW)*. 2121–2132.

[24] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Empirical Methods in Natural Language Processing (EMNLP)*. 262–272.

[25] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.

[26] Jipeng Qiang, Ping Chen, Tong Wang, and Xindong Wu. 2017. Topic modeling over short texts by incorporating word embeddings. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 363–374.

[27] Xiaojun Quan, Chunyu Kit, Yong Ge, and Sinno Jialin Pan. 2015. Short and sparse text topic modeling via self-aggregation. In *International Joint Conference on Artificial Intelligence*.

[28] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Empirical Methods in Natural Language Processing (EMNLP)*. 248–256.

[29] Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the loop: User-centered design and evaluation of a human-in-the-loop topic modeling system. In *International Conference on Intelligent User Interfaces*. 293–304.

[30] Yang Wang and Greg Mori. 2009. Human action recognition by semilatent topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 10 (2009), 1762–1774.

[31] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A Biterm Topic Model for Short Texts. In *The Web Conference (WWW)*. 1445–1456.

[32] Xiaohui Yan, Jiafeng Guo, Shenghua Liu, Xueqi Cheng, and Yanfeng Wang. 2013. Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In *SIAM International Conference on Data Mining (SDM)*. 749–757.

[33] Liansheng Zhuang, Haoyuan Gao, Jiebo Luo, and Zhouchen Lin. 2013. Regularized semi-supervised latent dirichlet allocation for visual concept learning. *Neurocomputing* 119 (2013), 26–32.