Dynamic Topic-Noise Models for Social Media

Rob Churchill* and Lisa Singh

Georgetown University
Washington, DC, USA
*Corresponding Author. Email: rjc111@georgetown.edu

Abstract. Temporal topic models often cannot effectively approximate topics on social media data sets due to the noise levels inherent in these types of data. Topic-noise models are important for modeling the short, sparse, noisy posts that we see throughout social media platforms. We propose using topic-noise models for temporal topic modeling, specifically D-TND (dynamic topic-noise discriminator). It enables topic and noise distributions to be generated together, modeling both the relationships between words in documents and the evolution of words and noise. We also propose Dynamic Noiseless Latent Dirichlet Allocation (D-NLDA), which integrates D-TND's time-dependent noise distribution with the topic distributions of Dynamic LDA, and show its propensity for improving dynamic topic models by effectively separating noise and topics on two large Twitter data sets.

1 Introduction

Topic models are important unsupervised tools for quickly understanding large textual data sets. They can be particularly useful when attempting to understand the discussion surrounding a large number of social media posts [4, 22, 26]. A number of topic models have been designed specifically to more accurately model social media data [5, 27, 16, 21]. More recently, a class of topic models called topic-noise models was proposed to jointly model topic and noise distributions on social media data [8]. None of these models incorporate a temporal dimension. Temporal topic models enable researchers to not only identify the relevant underlying topics in a data set, but also to track the evolution of these topics through time. Recently, there has been a renewed interest in temporal topic models, with the publication of a graph-based dynamic topic model [12], and an embedding-based dynamic topic model [10]. Even though these and other dynamic topic models have been proposed, they do not explicitly model noise.

In this paper, we adapt a topic-noise model to a temporal social media setting, with the goal of improving topic coherence by successfully removing noise from evolving topics. We accomplish this by adapting a topic-noise model, Topic-Noise Discriminator (TND) [8] to a temporal setting, Dynamic Topic-Noise Discriminator (D-TND). D-TND takes advantage of the joint topic-noise distribution generation of TND, while at the same time enabling the tracking of topics and noise through time by passing topic and noise distributions from one time period

to the next. We then propose a new temporal model, Dynamic Noiseless Latent Dirichlet Allocation (D-NLDA), a temporal version of Noiseless Latent Dirichlet Allocation (NLDA) [8] that integrates the proposed D-TND with Dynamic LDA (D-LDA). The advantage of this approach is that the noise distribution of D-TND and topic distribution of D-LDA evolve together, allowing for more accurate filtering of noise, and better-trained topic-word distributions at each time period. As we will see, D-NLDA is much greater than the sum of its parts.

The contributions of this paper are as follows. 1) We propose a new *temporal topic-noise model* that models noise and topics over time. 2) We propose a new *temporal topic model* that accounts for noise and generates higher quality topic sets. 3) To improve scalability, we introduce a vocabulary limiting function that reduces the vocabulary size of temporal data sets while maintaining topic quality. 4) We conduct an empirical analysis, both quantitative and qualitative, using two large Twitter data sets, that demonstrates the abilities of D-TND and D-NLDA to scale to accommodate such data sets, and to successfully identify high-quality topics. 5) We publish our model and evaluation code for others to use to continue advancing research in the field of temporal topic modeling.[1]

The paper is organized as follows: Section 2 presents related literature. Section 3 defines the notation used throughout the paper, details the models that were used in creating our proposed models, and presents our proposed models. Section 4 presents our quantitative and qualitative empirical analyses of our models. Finally, Section 5 presents our conclusions.

2 Related Literature

2.1 Static and Social Media Topic Models

The most well-known topic model is Latent Dirichlet Allocation (LDA) [3]. The basis upon which many topic models are built today, LDA is a bag-of-words model that approximates topics by maximizing the likelihood of documents in a k-dimensional Dirichlet distribution, where k is the number of topics. As documents are observed, words are probabilistically placed into topics and the probability distribution of each document over the topic set slowly changes to reflect the co-occurrence of words within the data set. After the model is trained, words that occur together in the same documents are more likely to be in the same topic. The result is topics containing words that are related according to the observed documents.

It has been apparent for some time that social media data sets require specially-constructed topic models to deal with the noise levels, short length, and sparsity of the data at hand. Biterm Topic Model detects topics, not from unigrams, but from bigrams generated from text [27], decreasing the vocabulary to improve quality. Self-Aggregating Topic Model (SATM) follows a similar vein, aggregating related short posts into longer pseudo-documents and generating topics from the pseudo-documents. GPUDMM [16] improves the coherence

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

of topics by sampling related words from an embedding space. Percolation-based Topic Models (PTM) [5] detects topics in social media data using a graph structure. A word co-occurrence graph is broken down into small communities and then built back up into small but coherent topics. For a more complete survey of unsupervised topic models, including ones designed for social media, see [6].

Topic-Noise Models [8] jointly model topics and noise distributions in order to more effectively remove noise from topics. Churchill and Singh propose a topic-noise model called Topic-Noise Discriminator (TND), which adds a noise distribution to LDA [3]. They use TND in an ensemble with LDA, called Noiseless Latent Dirichlet Allocation (NLDA), to create low-noise topics in domain-specific social media data sets. In this paper, we use the static TND and NLDA as a starting point to build two dynamic models, D-TND and D-NLDA.

2.2 Temporal Topic Models

Topics over Time (TOT) [25] jointly models time and topics, allowing for a continuous timeline of topics as opposed to discretized time periods like in D-LDA. Other early temporal topic models include MTTM [18], continuous DTM [24], Topic Tracking Model [13], and MDTM [14]. Dynamic Topic Models (D-LDA) [2] is a direct temporal adaptation of LDA [3]. Approximated probability distributions from a given time period are passed into the subsequent time period, in order to track the evolution of topics over time. We will draw on this temporal structure to create our dynamic topic-noise model, which incorporates a noise distribution into the model.

Bhadury et al. optimize D-LDA using multithreading and an optimized inference algorithm [1]. Topic Flow Model (TFM) [9] models temporal social media data using a graph structure. It runs a directed depth-first search from selected seed words to connected words and back to confirm mutual association. Dynamic Embedded Topic Model (D-ETM) [10] takes the Embedded Topic Model (ETM) [11], and adds a time-varying aspect. D-ETM runs ETM for each time period in the data set, passing parameters into the next time period like in D-LDA. The graph-based Dynamic Topic Model (GDTM) [12] is a scalable dynamic topic model for social media. The model assigns documents to topics based on the overlap of documents' graph representations, and partitions the documents based on graph density. One issue with GDTM is that it does not output the most probable words per topic, instead opting to output partitioned documents. Because of this, it is not directly comparable to models such as DTM, D-ETM, and our proposed models. In our experiments, we test our models against D-LDA, TFM, ToT, and D-ETM. The largest difference between our models and this previous work is that we explicitly model noise as a separate distribution. None of these other dynamic models do that.

3 Approach

In this section, we define our notation (Section 3.1) and review D-LDA, TND, and NLDA (Section 3.2). We then describe how we adapt the topic-noise models

TND and NLDA to a dynamic setting to produce D-TND (Section 3.3) and D-NLDA (Section 3.4). We then propose a method for improving the scalability of dynamic topic models, with the goal of producing dynamic models capable of handling large social media data sets (Section 3.5).

3.1 Notation

Let D represent a *dataset* consisting of M documents, where $D = \{d_0, d_1, \ldots, d_{M-1}\}$. A *document* d is a group of N words, where $d = \{w_0, w_1, \ldots, w_{N-1}\}$. A vocabulary V is the set of unique words in D. In our context, a word is a unigram. However, words can be replaced by phrases without loss of generality.

A topic z consists of ℓ related words, $z = \{w_0, w_1, \ldots, w_{\ell-1}\}$. The words in z should be coherent and interpretable by a human. A topic set Z contains k topics, $Z = \{z_0, z_1, \ldots, z_{k-1}\}$, that represent a summary of D. A noise distribution Ω is a probability distribution over V, where each word has a non-zero probability of being a noise word. In the case of temporal models, we use discretized time. We refer to a data set as consisting of T time periods, $\{t_0, t_1, \ldots, t_T\}$, where topics within a time period are constructed together.

3.2 Dynamic LDA, Topic-Noise Discriminator, & Noiseless LDA

Dynamic LDA (D-LDA) was designed to approximate topics over time, but does not take into account the noise inherent in social media data [2]. Topic-Noise Discriminator (TND) was designed to simultaneously approximate noise and topic distributions in social media data sets. While it can be used as a standalone topic model, it is best used in an ensemble, like Noiseless LDA (NLDA)[8]. NLDA leverages the noise distribution of TND and topic-word distribution of LDA to produce more coherent, high quality topics in social media data sets [8]. We briefly describe these core components of our dynamic topic-noise models here.

D-LDA. D-LDA defines a topic-word distribution $\beta_{t,k}$, where t is the time period, and k is the number of topics. For a document d, its document-topic distribution $\alpha_{t,d}$ is a probability distribution over β_t. When generating a word for document d on time slice t, a topic z is chosen from β_t conditioned on $\alpha_{t,d}$. The word $w_{t,d}$ is drawn from $\beta_{t,z}$. This results in topics that are generated relative to time, as well as the observed documents.

TND. Topic-Noise Discriminator is a generative model that assumes that documents are a mixture of topics and noise. Words are drawn from a mixture of the topic-word distribution and noise distribution to generate documents. Each word in an observed document is assigned to the noise or topic-word distribution, based on its prior probabilities of being in each. A Beta distribution (Equation 1) is used to determine whether a word belongs to the noise or topic distribution. β_z^i is the frequency of word i in topic z, and Ω_i is the frequency of word i in the noise distribution. The γ parameter can be increased to weight the Beta distribution toward assigning a word to the chosen topic over the noise distribution.

$$\lambda = Beta(\sqrt{\beta_z^i + \gamma}, \sqrt{\Omega_i}) \tag{1}$$

NLDA. While TND effectively models noise, it does not always independently find the strongest topics. Noiseless LDA [8] joins the noise distribution from TND with the topic distribution of LDA [3] to produce more coherent topics than those generated by TND or LDA. Assuming that we have a noise distribution Ω from TND and topic-word distribution β_z from LDA, NLDA integrates them based on each word's probability of being in the noise distribution and topic-word distribution for a given topic, in a process similar to Equation 1.

3.3 Constructing a dynamic topic-noise model

We now describe how D-TND is constructed. $\beta_{t,k}$ is the topic-word distribution for t over k topics. The document-topic distribution α_t is a probability distribution over β_t. α and β are initially group Dirichlet priors (document-topic and topic-word distributions, respectively) in the first time period, but once trained, are passed to future time periods as individual priors. α_t and β_t are initialized from their $t-1$ counterparts.

We define Ω_t to be the noise distribution at time t. Like α_t and β_t, Ω_t is conditioned on Ω_{t-1}. This inherently assumes that words that were noise in $t-1$ are still noise in t. While this will make it harder for noise words from $t-1$ to be included in topics, it does not make it impossible, merely less likely.

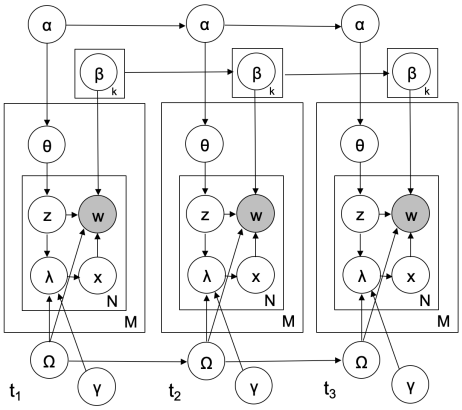


Fig. 1: Plate notation for D-TND, for three time periods.

Figure 1 shows plate notation for D-TND. Observed words are designated as noise or topic words within a time period based on a Beta distribution conditioned on the word's probability of being in the chosen topic or in noise. This process is represented by λ in Figure 1, and is tuned by γ. The designation is indicated by the switching variable x. For a given time period t, we generate a document d as follows:

1. Draw the number of words N for d.

2. Draw the topic distribution $\theta_{t,d}$ from the Dirichlet distribution, conditioned on α_t.
3. For each word w_i, $0 \leq i < N$:
   (a) Draw a topic z_i from the topic distribution $\theta_{t,d}$.
   (b) Draw a word from either z_i or the noise distribution Ω_t, according to λ_t, indicated by switching variable x.
   (c) If drawing from z_i, draw w_i from β_{t,z_i}.
   (d) If drawing from Ω_t, draw w_i from Ω_t.

3.4 Constructing dynamic NLDA

D-TND's most versatile feature is its noise distribution, which is trained alongside topics. Like TND for static models, D-TND can be easily integrated into generative temporal topic models. This makes D-TND particularly useful because researchers can use it in concert with whichever model they prefer.

Just as NLDA integrates TND's noise distribution and LDA's topic-word distribution, D-NLDA integrates D-TND's noise distribution and D-LDA's topic-word distribution. To create D-NLDA, we train a noise distribution Ω on D for each time period $t \in T$ using D-TND. Our assumption that noise and topics both evolve over time and in relation to each other allows us to track and integrate topics and noise in the style of NLDA, with a temporal aspect. We generate topics on D using D-LDA, and combine D-LDA's topic-word distribution β_t, k with D-TND's Ω_t to create topics for each time period. A word is removed or retained using the Beta distribution, conditioned on $\beta^i_{t,z}$ and $\Omega_{t,i}$ (Equation 2).

$$Beta\left(\sqrt{\beta^i_{t,z} + \gamma}, \sqrt{\Omega_{t,i}(\phi/k)}\right) \tag{2}$$

After the status of w_i has been determined, we follow the same guidelines as NLDA, incrementing $\Omega_{t,i}$ by one if w_i is noise, or by $\beta^i_{t,z}$ if w_i belongs to z. This ensures that, *for time period* t *only*, w_i has a high chance of not being put in another topic if it already belongs to one. As Ω has already been computed for all $t \in T$, this does not affect the status of w_i in future time periods.

3.5 Vocabulary Reduction to Improve Topic Model Performance

As we mentioned in Section 1, topic models are often too slow to infer topics on large data sets in a temporal setting. The original D-LDA [2], D-ETM [10], and ToT [25] only show results on data sets of tens of thousands of documents. In order to facilitate better scaling for topic models, we propose reducing the vocabulary size of data sets.

We define a *vocabulary limiting function* (VLF) to be a function that removes words from the vocabulary V, resulting in a smaller vocabulary V'. We define the frequency of a word $w_i \in V$ to be f_w. Given a threshold f_{min}, we compute the VLF as follows:

$$V' = V' \cup \{w_i\}\ \forall w_i \in V | f_{w_i} > f_{min} \tag{3}$$

In practice, we set f_{min} such that $|V'|$ is approximately equal to some target vocabulary size. It is also worth pointing out that this approach indirectly

reduces the size and possibly the number of documents. Instead of removing documents that may have important words, we remove words from documents that are less likely to be high probability words in a topic model. We evaluate the effects of VLF in Section 4. We note that the performance impact of this may be small in cases where the lowest frequency words are much less frequent than the average word.

4 Empirical Evaluation

In this section, we present our empirical evaluation of D-TND and D-NLDA using quantitative and qualitative approaches. We begin by describing our experimental setup, including data sets, preprocessing, and model parameters (Section 4.1). We then present a quantitative evaluation (Section 4.2), and a qualitative evaluation of our models' performance (Section 4.3).

4.1 Experiment Setup

Baseline Algorithms. In our experiments, we tested against four state-of-the-art temporal topic models: D-LDA [2], ToT [25], TFM [9], and D-ETM [10]. They are each described in Section 2.

Data Sets. In our analysis, we use two Twitter data sets. The first data set contains posts about the 2020 United States Presidential Election from August 1 to November 14, with weekly time periods. We refer to this data set as *Election 2020*. The second data set, *Covid-19*, contains posts about the Covid-19 pandemic, collected between March 2020 and February 2021, with monthly time periods. We collected these documents using hashtags related to the election and Covid-19, respectively, via the Twitter Streaming API, and randomly sampled 200,000 posts per time period.[2]

We use our vocabulary limiting function (VLF) to create different versions of each data set. The large version is the full vocabulary, ($f_{min} = 0$). We set f_{min} such that $|V'| \approx 10,000$ for each time period to get medium-size data sets.[3] f_{min} was set such that $|V'| \approx 5,000$ for each time period for small-size data sets.[4] Table 1 shows the exact effects of the VLF for each data set. While there is a significant reduction in vocabulary, the number of documents remains high. In Covid-19, just over 100,000 documents, or about 4%, are lost, while in Election 2020, about 200,000 documents, or about 6.67% are lost. As we will see, this loss in documents has very little effect on the quality of topics.

Text Preprocessing. Text processing can have a positive impact on topic model performance [7]. For our data sets, we tokenize on whitespace, remove

[2] Leaving data sets in their original form, with a large skew in data set size from time period to time period, reinforces the skews in more pronounced ways in the probability distributions, leaving effects on future time periods.

[3] $f_{min} = 15, 20$ in Election2020 and Covid-19 for the medium-size data sets.

[4] $f_{min} = 40, 50$ for Election2020 and Covid-19 for the small-size data sets.

Table 1: Data Set Qualities for different size variants of vocabulary. $|D|/t$ and $|V|/t$ are average data set size and vocabulary size within a time period.

		$\lvert D\rvert$	$\lvert D\rvert/t$	$\lvert V\rvert$	$\lvert V\rvert/t$
Covid-19	Large	2,400,103	200,008	1,041,552	172,116
	Medium	2,326,370	193,864	28,198	10,483
	Small	2,292,266	191,022	13,890	5,645
Election 2020	Large	3,000,042	200,002	648,193	96,671
	Medium	2,836,549	189,103	33,391	9,484
	Small	2,800,209	186,680	18,010	5,153

lowercase text, remove URLs, punctuation (including hashtags), and stopwords. We also remove deleted posts and user tags.

Model Parameters. We conduct a sensitivity analysis for D-TND and D-NLDA, testing each model with an array of different parameter settings. Due to space limitations, we present the results for the best-performing settings. For D-TND, we found the best parameter settings to be $\alpha = 1$, $\beta = 0.01$, $\gamma = 25$, and $k = 30$. The best settings for D-NLDA were the same settings as D-TND, with $\phi = 10$.[5] For D-LDA, the best parameter settings were $\alpha = 1$, $\beta = 0.01$, and $k = 30$. The chosen α and β parameters consistently resulted in better topic quality than other options. The γ parameter is less sensitive than α and β, but $\gamma = 16$ was also a reasonable choice. We found that $\gamma = 0$ or 36 were too extreme of settings for our data sets, designating too few and too many words as noise, respectively. Changing the ϕ parameter can lead to far more coherent topic sets. We found that $\phi = 5$ resulted in too few noise words being filtered from topics, but that $\phi > 15$ resulted in some quality words being removed from topics. We note that it is straightforward to quickly iterate through ϕ values, since the filtering of noise is the fastest part of the model. For D-ETM and ToT, the parameters suggested in the papers were used, with $k = 30$ to match the parameters of the other models. While a sensitivity analysis was conducted, It is possible that with more extensive hyperparameter tuning, performance could be improved.

4.2 Quantitative Analysis

Evaluation Metrics Similar to previous work, we assess a model's ability to detect coherent, interpretable topics using a normalized point-wise mutual information score (NPMI) [15]. NPMI attempts to quantify the relatedness of two words within a topic, given their cofrequency, and is a commonly used evaluation metric [8, 10, 16, 11, 21, 20]. For a pair of words (x, y), we define the probability of them appearing in the same document as $P(x, y)$. We define the probability

[5] Parameters for sensitivity analysis across our models: $k = \{10, 20, 30, 50, 100\}$, $\alpha, \beta = \{0.01, 0.1, 1.0\}$, $\gamma = \{0, 16, 25, 36\}$, $\phi = \{5, 10, 15, 20, 25, 30\}$

Table 2: Time per iteration on each data set (s=seconds, m=minutes).

Model	Covid-19			Election 2020		
	Large	Medium	Small	Large	Medium	Small
D-TND	19.0 s	18.2 s	14.6 s	21.9 s	15.2 s	13.8 s
D-LDA	**1.5 s**	**1.4 s**	**1.32 s**	**2.0 s**	**1.2 s**	**0.9 s**
D-NLDA	20.6 s	19.7 s	16.0 s	23.6 s	16.4 s	14.8 s
D-ETM	480 m	117 m	87 m	360 m	177 m	138 m

of any word w appearing in a document as $P(w)$. Using these probabilities, we compute the NPMI of a topic $z \in Z$:

$$NPMI(z) = \frac{\sum_{x,y \in z} \frac{\log(\frac{P(x,y)}{P(x)P(y)})}{-\log(P(x,y))}}{\binom{|z|}{2}}$$

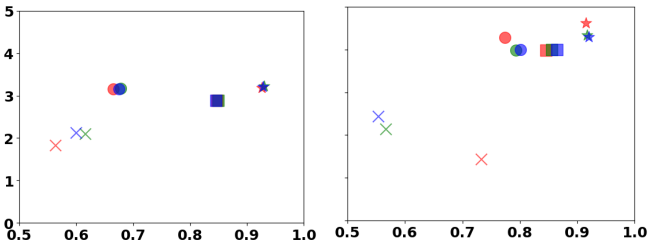
A higher NPMI indicates high topic coherence and lower noise penetration, or that a topic model is creating meaningful topics. We refer to the topic-wise NPMI score as *topic coherence*.

Unfortunately, a model can, in theory, find ten variants of the same meaningful topic. We care about the ability of a topic model to detect unique topics from the data,. *Topic diversity* is the fraction of unique words in the top 20 words of all topics in a topic set [11]. A model with high topic diversity is able to find almost entirely unique topics, while a model with low diversity is not able to successfully delineate between unique topics. *Topic quality*, proposed by Dieng et al. [10], is the product of the coherence and diversity scores. As we care about both metrics, a product of the two gives a good overall score for a topic set.

Given the size of our data sets, we are concerned about efficiency. For our experiments, models were run on a machine with twelve 2.2GHz virtual cores, with 50GB of memory. D-TND, D-NLDA, D-LDA, and D-ETM take advantage of parallelization or multi-threading (Mallet for D-LDA and D-TND [17], PyTorch for D-ETM [19]). ToT did not scale to the size of our data sets. It was allowed to run for three days, and did not complete an iteration for either data set. TFM ran for three days and did not finish constructing topics. The topics found contained only a single word, meaning its topic coherence would be zero. As a result, we do not include TFM and ToT in the analysis that follows.

Efficiency. To analyze efficiency, we compute time per iteration for the other methods (see Table 2). As we can see, D-LDA is the most efficient model. Because it is only computing a topic distribution and not a noise model as well, this result is not surprising. D-TND and D-NLDA are the next most efficient models and are comparable to each other. Our models are between 300 and 1500 times faster than D-ETM, the most recent temporal topic model in our study.

D-ETM is implemented using PyTorch, a highly optimized Python framework for neural networks [19]. It is run for ten iterations on the small and medium



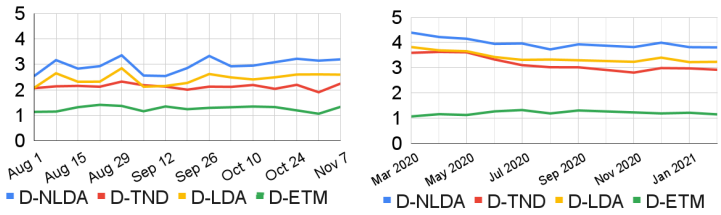
(a) Election 2020 results. (b) Covid-19 results.

Fig. 2: Coherence (y-axis) and Diversity (x-axis)

data sets and five iterations on the large data set given that each iteration took approximately 8 hours to run. D-LDA and our models were run for 500 iterations. Highlighting its ability to work on larger data sets. Part of the differential in computation time between D-ETM and the other models is likely due to the fact that D-ETM is implemented in Python, whereas the other models are implemented in Java. It is possible that a Java implementation of D-ETM would be faster than its Python implementation, but given the complexities of the underlying model, it is unlikely that a Java version of D-ETM would be faster than D-TND or D-NLDA.

Coherence and Diversity. This section focuses on the quality of the models. Figure 2 plots the mean coherence and diversity score for D-TND (circles) and D-NLDA (stars) alongside D-LDA (squares) and D-ETM (X) for each data set. Coherence is plotted on the Y-axis, and diversity is plotted on the X-axis. The results for the large-size data set are colored red, the medium-size blue, and the small-size green. The closer to the top-right corner of the plot a model is, the better. D-NLDA performs the best out of any model on both data sets. In the Election 2020 data set, there's little difference in D-NLDA's performance across the different-size data sets. In the Covid-19 data set, we see a slight deterioration in terms of coherence when we use VLF to remove words from the vocabulary. Table 1 shows the difference between the Election 2020 and Covid-19 data sets in terms of how many words are removed from each vocabulary. By aiming to retain approximately 10,000 and 5,000 words per time period in the medium and small data sets, far more words were removed from the Covid-19 vocabulary than the Election 2020 vocabulary. It seems that in the case of Covid-19, we removed too many words from the vocabulary. This adversely affected the topic quality. In the Election 2020 data set, we retained all or most of the topic words, which is reflected in the maintained high topic quality across data set sizes. While we can certainly improve the scalability of topic models by reducing vocabulary size, removing too many words can sacrifice topic quality.

D-NLDA has a slightly higher (1.5%) coherence than D-TND, but a 25% higher diversity score. Compared to D-LDA, its diversity is 7% higher, and its coherence is 8% higher. D-NLDA once again is the best model on all data set sizes, beating D-TND by 0.35 in coherence and 14% in diversity, and beating D-LDA by 0.64 in coherence and 7% in diversity. D-ETM's poor performance



(a) Election 2020 results over time. (b) Covid-19 results over time.

Fig. 3: Topic Quality Plot for Election2020 and Covid-19 medium-size data sets.

is likely due to its inability to finish enough iterations in a reasonable amount of time to detect high-quality topics. Finally, for all models except D-ETM, the vocabulary size had little effect on the overall topic quality. The low variance in performance on D-TND, D-LDA, and D-NLDA reflects our theoretical assertion that removing the lowest frequency words should have very little affect on topic model performance. For D-ETM, the coherence of topics increases with the use of VLF, indicating that D-ETM benefits from smaller vocabularies.

In order to understand how models perform over time in relation to one another, we plot topic quality, the product of the coherence and diversity scores, for each model in each time period in Figure 3. Plotting topic quality over time highlights the similarity of D-NLDA and D-LDA, but also highlights the clear improvement of D-NLDA with the addition of D-TND's noise distribution.

Table 3: Percent judge agreement on Covid-19 temporal topics.

Topic	Vaccines	Lockdowns	Cases	Testing	Schools	Masks	Global Impact	Economy	India	China
Agreed %	100	100	100	100	100	100	100	100	80	100

4.3 Qualitative Analysis

Our qualitative analysis shows D-NLDA's ability to track topics through time. For the Covid-19 data set, we asked five human judges to individually label ten topics generated by D-NLDA that persisted throughout every time period. In Table 3, we show agreement between the judges, with the agreed upon label for each topic. For all topics but one, every judge independently concluded the same topic label. These findings indicate that D-NLDA is able to generate high quality topics that humans can easily comprehend.

We highlight this ability with a deeper look at the evolution of the vaccine topic through time periods, seeing it evolve and grow (see Figure 4). Words highlighted green appeared in the topic in the previous time period, and words highlighted yellow appeared in the topic in any previous time period. The topic starts out with a wish for a vaccine and with concern for healthcare workers. It evolves into a reality in the middle of 2020 and goes through drug trials. Finally, the vaccine is approved in late 2020 and rolled out at the beginning of 2021.

March 2020	April	May	June	July	August	September	October	November
medical	vaccine	vaccine	vaccine	vaccine	vaccine	vaccine	vaccine	vaccine
workers	family	scientists	china	sarscov2	research	trials	emergency	vaccines
back	doctors	hydroxychloroquine	world	human	sarscov2	fauci	vaccines	effective
healthcare	nurses	research	virus	study	study	trial	public	trial
rate	treatment	treatment	lives	results	effective	experts	political	family
made	dying	study	black	research	scientists	find	health	results
night	heart	evidence	united	made	vaccines	clinical	data	pfizer
full	science	prevent	matter	early	immunity	trust	mental	missed
vaccine	policy	hands	human	shows	emergency	company	information	friends
action	scientists	spread	sarscov2	trials	governments	hold	control	spent
national	didnt	effective	chinese	complete	told	ready	billion	months

December	Dec (2)	January 2021	Jan. (2)	Jan. (3)	Jan. (4)	February	Feb. (2)	Feb. (3)
vaccine	vaccines	vaccine	vaccination	people	vaccines	vaccine	vaccines	vaccination
pfizer	effective	pfizer	india	video	doses	dose	countries	house
days	data	effective	minister	sick	healthcare	received	global	biden
vaccination	hospitals	rollout	working	vaccinated	workers	covidvaccine	india	president
heres	county	received	time	work	million	vaccinated	world	years
doses	california	covidvaccine	narendramodi	laurie_garrett	countries	pfizer	make	team
vaccinated	developed	started	prime	taking	county	rollout	access	past
years	christmas	receive	families	paid	iran	single	communities	economy
covidvaccine	shows	moderna	global	weeks	high	shot	safe	vaccinations
drleanawen	kids	mrna	response	america	frontline	distribution	country	mass
reach	tomorrow	dose	corona	bill	worker	leading	ensure	white

Fig. 4: Evolution of the Vaccine topic in the Covid-19 medium-size data set.

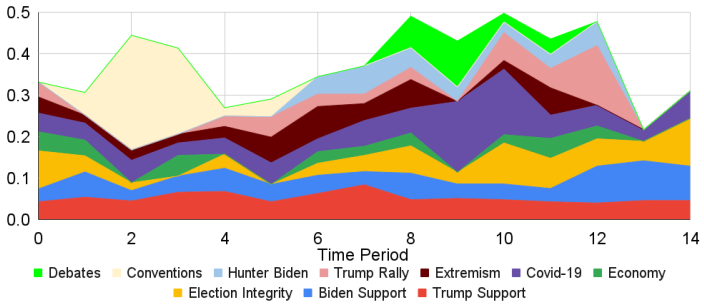


Fig. 5: Election 2020 Topic proportions (y-axis) over time periods (x-axis).

D-NLDA allows us to see a very detailed evolution of the Vaccine topic that contains limited noise throughout the entirety of the year, showing the promise of topic-noise models within a temporal setting.

In the Election 2020 medium data set, we show the ability of D-NLDA to accurately track multiple relevant topics through time. To produce topic labels for this data set, we relied on a manually-generated topic set, curated by political scientists who closely studied the 2020 Election on social media platforms [23]. Figure 5 shows how the topic proportions of selected topics change throughout the election. New topics emerge and disappear throughout the campaign. We can see the large impact of the party conventions in time periods two and three (late August 2020), and how quickly talk about conventions ceases after they are over. The same happens with topics about Presidential and Vice Presidential Debates in time periods eight to ten (early October).

The Conventions and Debates topics represent bursty topics, which appear out of nowhere and disappear quickly as attention turns away from them. These bursty topics could be missed or muddled with other topics by static mod-

els.In general, this topic flow visualization highlights the ability of a dynamic topic-noise model like D-TND or D-NLDA to produce highly relevant and easily understandable topics with a temporal aspect. The ability to understand how topics evolve over an election is important both for voters and candidates.

5 Conclusions

In this paper we create a dynamic temporal-noise model that incorporates a noise distribution into a temporal topic model for the first time (D-TND), and weave together D-TND with the well-known D-LDA model to create D-NLDA. These approaches bring to temporal topic models the noise-filtering benefits of topic-noise models that are so necessary for social media data sets.

We demonstrate the ability of our proposed methods to both scale to large temporal data sets, and produce high quality topics on the data sets through time periods spanning weeks and months. We show how using a vocabulary limiting function (VLF) can speed up topic models, and in some cases, produce better topics. Finally, we share our code on GitHub for others to use.[6]

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 also like to thank the Mosaic Project and SSRS for access to the Covid-19 Survey data set.

References

1. Bhadury, A., Chen, J., Zhu, J., Liu, S.: Scaling up dynamic topic models. In: The Web Conference (WWW) (2016)
2. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: International Conference on Machine Learning (ICML) (2006)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)
4. Bode, L., Budak, C., Ladd, J.M., Newport, F., Pasek, J., Singh, L.O., Soroka, S.N., Traugott, M.W.: Words that matter: How the news and social media shaped the 2016 Presidential campaign. Brookings Institution Press (2020)
5. Churchill, R., Singh, L.: Percolation-based topic modeling for tweets. In: KDD Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM) (2020)
6. Churchill, R., Singh, L.: The evolution of topic modeling. ACM Computing Surveys (CSUR) (2021)
7. Churchill, R., Singh, L.: textprep: A text preprocessing toolkit for topic modeling on social media data. In: The DATA Conference (2021)

[6] Our code can be found at https://github.com/GU-DataLab/topic-modeling

8. Churchill, R., Singh, L.: Topic-noise models: Modeling topic and noise distributions in social media post collections. In: International Conference on Data Mining (ICDM) (2021)
9. Churchill, R., Singh, L., Kirov, C.: A temporal topic model for noisy mediums. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) (2018)
10. Dieng, A.B., Ruiz, F.J.R., Blei, D.M.: The dynamic embedded topic model. CoRR (2019)
11. Dieng, A.B., Ruiz, F.J., Blei, D.M.: Topic modeling in embedding spaces. arXiv preprint arXiv:1907.04907 (2019)
12. Ghoorchian, K., Sahlgren, M.: Gdtm: Graph-based dynamic topic models. Progress in Artificial Intelligence **9**, 195–207 (2020)
13. Iwata, T., Watanabe, S., Yamada, T., Ueda, N.: Topic tracking model for analyzing consumer purchase behavior. In: International Joint Conference on Artificial Intelligence (2009)
14. Iwata, T., Yamada, T., Sakurai, Y., Ueda, N.: Online multiscale dynamic topic models. In: Conference on Knowledge Discovery & Data Mining (KDD) (2010)
15. Lau, J.H., Newman, D., Baldwin, T.: Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: Conference of the European Chapter of the Association for Computational Linguistics (EACL) (2014)
16. Li, C., Wang, H., Zhang, Z., Sun, A., Ma, Z.: Topic modeling for short texts with auxiliary word embeddings. In: SIGIR Conference on Research and Development in Information Retrieval (2016)
17. McCallum, A.K.: Mallet: A machine learning for language toolkit. (2002)
18. Nallapati, R.M., Ditmore, S., Lafferty, J.D., Ung, K.: Multiscale topic tomography. In: Conference on Knowledge Discovery & Data Mining (KDD) (2007)
19. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems (NIPS) (2019)
20. Qiang, J., Chen, P., Wang, T., Wu, X.: Topic modeling over short texts by incorporating word embeddings. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) (2017)
21. Quan, X., Kit, C., Ge, Y., Pan, S.J.: Short and sparse text topic modeling via self-aggregation. In: International Joint Conference on Artificial Intelligence (2015)
22. Singh, L., Bode, L., Budak, C., Kawintiranon, K., Padden, C., Vraga, E.: Understanding high-and low-quality url sharing on covid-19 twitter streams. Journal of Computational Social Science **3**(2), 343–366 (2020)
23. Singh, L., Ladd, J., Pasek, J., Traugott, M., Budak, C., Soroka, S., Agiesta, J., Sparks, G.: The breakthrough [polling project] (2020)
24. Wang, C., Blei, D., Heckerman, D.: Continuous time dynamic topic models. arXiv preprint arXiv:1206.3298 (2012)
25. Wang, X., McCallum, A.: Topics over time: A non-markov continuous-time model of topical trends. In: Conference on Knowledge Discovery & Data Mining (KDD) (2006)
26. Williams, J.B., Singh, L., Mezey, N.: # metoo as catalyst: A glimpse into 21st century activism. University of Chicago Legal Forum p. 371 (2019)
27. Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In: The Web Conference (WWW) (2013)