Online Confirmation-Augmented Probabilistic Topic Modeling in Cyber-Physical Social Infrastructure Systems

Jiajia Xie
Georgia Institute of Technology
Atlanta, United States
jxie@gatech.edu

Christin J. Salley
Georgia Institute of Technology
Atlanta, United States
csalley3@gatech.edu

Neda Mohammadi
Georgia Institute of Technology
Atlanta, United States
nedam@gatech.edu

John E. Taylor
Georgia Institute of Technology
Atlanta, United States
jet@gatech.edu

ABSTRACT

Online probabilistic topic models serve as essential analytical tools within Cyber-Physical Social Infrastructure Systems (CPSIS), enabling the analysis of real-time data streams. These models empower operators and decision-makers with actionable insights, anomaly detection, predictions, optimized resource allocation, user engagement, and social feedback, all critical for responding to evolving CPSIS conditions. While these models use inferred topic-assignment distributions to create lower-dimensional representations, applying them to online user-generated streams, like social media and community apps, has historically posed challenges due to sparse relevant content, leading to suboptimal performance. Our study proposes a novel and expanded version of topic models that integrates the variational lower bound with a linear reward function, supervised by a label associated with the confidence of relevant content presence. We introduce a learning algorithm designed for these augmented topic models. Our empirical experiments, conducted on real-world datasets, provide compelling evidence that our approach uniquely enhances the potential of any topic model in CPSIS for downstream tasks in information management. These enhancements encompass improved topic interpretability, enhanced data labeling precision, and the refinement of similarity metrics, reinforcing the effectiveness of our online confirmation-augmented probabilistic topic modeling approach in processing and analyzing CPSIS real-time data streams.

CCS CONCEPTS

• **Information systems → Data stream mining**; *Data streaming*;
• **Computing methodologies → Online learning settings**.

KEYWORDS

information augmentation, online machine learning, topic models, variational bayes

ACM Reference Format:
Jiajia Xie, Christin J. Salley, Neda Mohammadi, and John E. Taylor. 2023. Online Confirmation-Augmented Probabilistic Topic Modeling in Cyber-Physical Social Infrastructure Systems. In *The 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '23), November 15–16, 2023, Istanbul, Turkey.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3600100.3626341

1 INTRODUCTION

Consider a scenario in which online data streams are leveraged by Cyber-Physical Social Infrastructure Systems (CPSIS) such as emergency response, disaster management, or public health monitoring and disease control systems. Due to the real-time coordination in place in these systems, they rely on user-generated data (e.g., social media content, data generated through traffic-information applications, health-related data, and emergency hotlines) to monitor, detect, and respond to events and trends in a timely and effective manner. These systems, depending on the specific monitoring requirements, operate within a predefined spatiotemporal window, which is typically a short timeframe that encompasses real-time data collection and response. In this context, additional data becomes available online, helping to identify when and where specific relevant information may emerge after a certain delay. The objective is to develop a topic model based on the online data stream, with a focus on the topics of interest. The model continuously learns from historical and real-time data to enhance its detection algorithms and improve response strategies.

Streaming data analysis in this way is crucial, as it enables the discovery of relevant topics within the selected content, which in turn plays a pivotal role in tasks like information augmentation [13, 19, 22], detecting traffic or crisis events[4, 15, 18, 20, 24], and more. The key component of the aforementioned studies is the topic representations inferred from short text-based data. Conventional topic models, such as Latent Dirichlet Allocations (LDA) [3] and Mixture of Unigrams (MUG) [10], when operating in fully unsupervised settings, often exhibit suboptimal performance resulting from the sparse presence of relevant contents. The reasons for this can be attributed to two factors: (1) user-generated data being exceptionally brief [7, 21], and (2) online data inherently containing noise [9]. While various unsupervised topic models have been developed to address data sparsity [11, 16], less emphasis has been placed on leveraging the data itself for model enhancement. Explicit

BuildSys '23, November 15–16, 2023, Istanbul, Turkey

ACM ISBN 979-8-4007-0230-3/23/11.
https://doi.org/10.1145/3600100.3626341

methods for improving data quality through post-processing and annotations are often time-consuming and costly, especially given the online nature of the data [2].

In this study, as we introduce this challenge, our approach aims to integrate online data with topic models. We do so by introducing a confidence score associated with the periodic arrival of content of interest, which is derived from the combination of various sources of information. To demonstrate the aforementioned scenario involving online data, we provide a real-world example. Consider the data source as a stream of tweets within a specific geographic area, recalling that the Twitter API V2 has enabled streaming via bounding boxes [6]. Suppose we are interested in posts related to emergency events within this area. Waze, one of the largest GPS navigation apps, offers interactive features that allow users to share real-time traffic information and report crisis events. For the purposes of this study, Waze can serve as a valuable information resource, as its data can be correlated with the presence of relevant tweets within a common spatiotemporal window. One valuable aspect is *nThumbsUp*, which reacts and directly reflects the event's significance within the online community. The primary concept here is to gather data that quantifies confidence within a smaller online community. If this score reaches a significant level, it may indicate that a larger online community, such as Twitter, is also discussing events of interest. See Figure 1 for the schematic demonstration.

The main problem addressed in this study is the design of models capable of interactively confirming the presence of relevant information within the topic representations of interest, particularly in an online context. To the best of our knowledge, no prior studies have specifically addressed this setting leveraging a confidence score as a form of weak supervision to enhance an otherwise purely unsupervised model. We propose a novel online machine learning framework that integrates a linear reward function linked to the confirmation confidence (e.g., *nThumbsup*) with the variational-Bayesian lower bounds of probabilistic topic models. The only modification applied to the topic model involves the incorporation of a variational distribution for document-topic assignments through a bilinear function that connects variational posterior parameters and confirmation parameters. Our experimental results, obtained using real-world data, highlight the following advantages of the entire framework:

(1) The linear reward function for confirmation will eventually reveal topics linked to the events of focus.
(2) Empirically, simple baseline models, LDA and MUG, when augmented with the confirmation model, yield improved semantic interpretations. The results imply that the framework can be extended to other topic models.
(3) Our method can improve downstream tasks for event detection and data augmentations. Additional experiments demonstrate improvements in data labeling for classification and in measuring similarity/dissimilarity.
(4) A real-world case study demonstrates a potential implementation of our model for augmenting Waze alerts using the *nThumbUps* feature.

The remainder of this paper is organized as follows: In Section 2, we review literature on topic models and their applications. Section 3 introduces essential preliminary notations and problem statements. Section 4 delves into the primary methodology. Section 5 outlines the experiment design and results, while, Section 6 provides a case study.

2 LITERATURE REVIEW

To the best of our knowledge, this presents the first attempt at the interactive learning of a confirmation model and a topic model within an online setting, aiming to address the problem we have presented. Our work is related to probabilistic topic models with variational inferences and their applications in event detection and information augmentation.

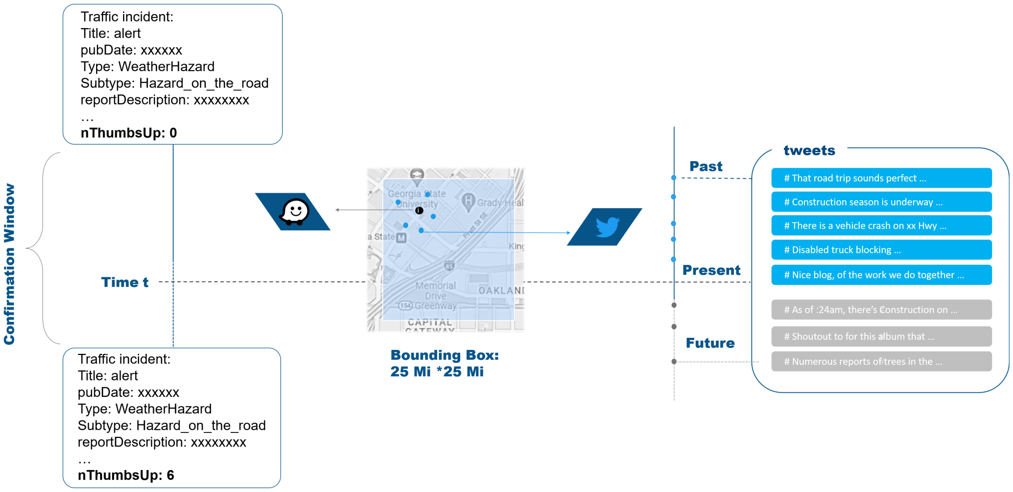
One of the earliest probabilistic topic models is the well-known Latent Dirichlet Allocation introduced by Blei, Ng, and Jordan [3]. The parameter estimation of LDA is challenging as the posterior distribution is computationally intractable [14]. Variational inference, where the posteriors are assumed to be multinomial and Dirichlet [3], has been one approach to address this issue. Hoffman et al. present the online versional variational inference of LDA [5]. However, LDA and many of its extended versions struggle with learning topics from documents in the format of short texts, such as social media data from Facebook or Twitter, which typically contain only one or two topics, rather than a mixture of all topics [16]. One preliminary model for addressing the sparsity of topics in short text is the mixture of unigrams model [10]. This idea has been further extended by Lin et al. to a dual-sparse topic model [7]. Other works address the sparsity issue by expanding the dimensionality, such as a bi-term model [21]. We refer to two surveys on probabilistic topics addressing the issue of sparsity in short texts [11, 16]. While various modeling methods exist, limited attention has been paid to incorporating data and weak supervision into variational inference in topic models, which is the focus of what we propose.

CPSISs rely on robust information management practices including the application of topic models to collect, process, and analyze real-time data streams from diverse sources. These analyses are vital for detecting events, anomalies, emerging patterns, and ensuring the seamless operation of these systems. Topic models are particularly effective in event detection and information augmentation. In event detection, the objective is to measure the uniqueness of identified patterns in the data. In information augmentation, the goal is to match similar content from different information sources to provide comprehensive context and background for users. Technically, both tasks often employ topic models for either (1) data labeling [4, 13, 15, 19] or (2) similarity measurement between topic distributions of two documents, [22, 22, 24]. The former serve as sources of data annotations for supervised/semi-supervised [13, 15, 19] classifiers, while the latter aims to retrieve lower dimensional representations for clustering analysis. For example, by computing the cosine similarity between the target and a potential candidate, we can assess the relevancy [20, 22] or uniqueness [24] of the candidate compared to the central topic.

3 DEFINITIONS AND PRELIMINARIES

3.1 Problem Statements

Let $t \in [T]$ represent discrete timestamps where $[T] = \{1, 2, \dots T\}$. We assume that at each $t \in [T]$, we are provided with a set of



Traffic incident:
Title: alert
pubDate: xxxxxx
Type: WeatherHazard
Subtype: Hazard_on_the_road
reportDescription: xxxxxxxx
…
nThumbsUp: 0

Traffic incident:
Title: alert
pubDate: xxxxxx
Type: WeatherHazard
Subtype: Hazard_on_the_road
reportDescription: xxxxxxxx
…
nThumbsUp: 6

That road trip sounds perfect …
Construction season is underway …
There is a vehicle crash on xx Hwy …
Disabled truck blocking …
Nice blog, of the work we do together …
As of :24am, there's Construction on …
Shoutout to for this album that …
Numerous reports of trees in the …

Figure 1: Illustration of a real-world scenario. The left-hand side is a Waze alert, which has *nThumbsUp* being constantly collected from online users to support the reliability of events. The right-hand side is the pool of tweets being posted within the same spatial-temporal window. In the real world, if an event has influenced a relatively small online community, the same event may have already influenced larger online communities, such as Twitter.

tweets $D^t \subset \mathcal{D}$, where each $d \in D^t$ is a vector in the bag-of-words format. In addition, we are given a binary label $y^t \in \{0, 1\}$, but there is no guarantee that y^t will arrive at timestamp t. It is common for y^t to have a certain delay. A value of $y^t = 1$ indicates the presence of some d of interest, while $y^t = 0$ indicates the absence of such d of interest. A real-world example of y^t is the *nThumbsUp* illustrated in 1. We should manually tune a threshold τ such that if *nThumbsUp* $> \tau$ we set $y^t = 1$ and 0 otherwise.

We present the formulation of topic modeling in terms of dimensional reduction. That is:

Problem 1. *Given D^t, learn/update the parameters of a topic model $\mathcal{H} : \mathcal{D} \longrightarrow \mathbb{R}^K$, where K is the number of topics and $\sum_k \mathcal{H}(D^t)_k =$* 1.

Therefore, the output is a K dimensional multinomial distribution over K topics of D^t. Meanwhile, the confirmation model can be defined as follows:

Problem 2. *Given D^t, learn/update the parameters of a reward function $f : \mathcal{D} \times \mathcal{Y} \longrightarrow \mathbb{R}$ so that it will gain more reward if $y^t = 1$ prior to the reveal of the ground truth y^t.*

It is important to note that Sub-problem 1 is an unsupervised learning problem and Sub-problem 2 is supervised. Although, these two problems may initially appear independent of each other, the primary contribution of this study is to propose a model capable of effectively addressing both problems interactively in an online machine-learning setting. We will demonstrate that by simultaneously solving Sub-problem 2, which involves confirmation, alongside Sub-problem 1, the model can yield a more focused (or skewed) distribution of topics of interest.

Table 1 provides a summary of all the notations used in this study and their corresponding descriptions.

Table 1: Notations and Descriptions

Notations	Descriptions
D^t	The set of documents (e.g. tweets) collected at time t
y^t	The binary label indicating the existence of targeted tweets at t
T	The maximum amount of timestamps
K	The Number of the topics.
W	The set of words.
β_k	The random latent variables of word-topic distribution.
γ_d	The random latent variables of document-topics distribution.
π	The parameters of the linear reward function for confirmation.
ϕ_{dwk}	The parameter of the variational posterior $q(Z_{dw} = k)$.
γ_d	The parameters of the variational posterior $q(\xi_d)$
λ_k	The parameters of the variational posterior $q(\beta_k)$
$\hat{\theta}^t$	All parameters of the topic model depends on t
$\bar{\theta}$	All parameters of the topic model independent of t

3.2 Variational Lower Bounds

Let Z^t represent all the latent variables of the topic models parameterized by $\theta^t = (\bar{\theta}, \hat{\theta}^t) \in \Theta$, the variational lower bound of a topic model with given D^t is:

$$\begin{aligned} \log p(D^t|\theta^t) &\geq \mathbb{E}_{q(Z^t)}\{\log p(D^t, Z^t|\theta^t)\} - \mathbb{E}_{q(Z^t)}\{\log p(Z^t)\} \\ &:= \ell(\theta^t|D^t) \end{aligned} \tag{1}$$

where $p(Z^t)$ is the prior distribution and $q(Z^t)$ is a variational posterior distribution we select. For generality, θ^t has parts $\bar{\theta}$ independent of t.

4 PROPOSED METHOD

4.1 Online Machine Learning

Our method further requires that the topic models must have a K-multinomial variational posterior. For instance, in LDA, there is a variational posterior of the per-word $w \in W$ topic assignment $q(Z_{d,w} = k) = \phi_{dwk}$ [5]. In MUG, there is a variational posterior of per-document topic assignment $q(Z_d = k) = \phi_{dk}$ [10]. The key idea of the interactive model is to introduce a linear reward function parameterized by $\pi \in \mathbb{R}^K$ such that $\sum_k \pi_k = 1$. We assume LDA is the topic model for the rest of the section. The linear reward function is defined as:

$$f(\pi|\phi^t, D^t, y^t) = \sum_{d \in D^t} \sum_w \sum_k \pi_k \phi^t_{dwk} y^t \tag{2}$$

The linear rewards' parameters π interact with the LDA model via the posterior distribution such that;

$$q(Z^t_{dw} = k) = \phi^t_{dwk} \pi_k / (\sum_{k'} q(Z^t_{dw} = k')) \tag{3}$$

Let us take a closer look at the linear function. When $y^t = 1$, which happens when we have enough confidence to confirm the presence of relevant $d \in D^t$, the multinomial distribution ϕ^t_{dwk} temporarily assumed given to us must be distinct from $y^t = 0$. For maximizing $\sum_{t=1}^{T'} f(\pi^t|\phi^t, D^t, y^t)$ for every $T' \in [T]$, the problem is equivalent to the well-known online learning problem: learn from K experts' advice [1]. π is encouraged to assign more weights to the topics that frequently appear when $y^t = 1$.

The main idea is to maximize both the linear reward function and the variational lower bound of the topic models simultaneously. In the context of online convex optimization, we set $g(\theta^t, \pi) = \ell(\theta^t|D^t) + f(\pi|\phi^t, D^t, y^t)$, and the benchmark of success is:

$$\mathcal{R}(\pi^t, \theta^t) := \max_{\pi, \theta^t} \sum_t g(\theta^t, \pi) - g(\theta^t, \pi^t) \tag{4}$$

subject to constraints $\sum_k \phi^t_{dwk} = 1$ and $\sum_k \pi^t_k = 1$, and recall that $\phi^t \in \Theta$ and $\mathcal{R}$ is the regret function.

4.2 An Online Algorithm

To solve the online problem, we derive an online algorithm with a theoretically guaranteed bound on the regret function $\mathcal{R}$. Due to page limit, we only present the iterative updates on π^t and ϕ^t and refer to [5] for the rest of the other parameters' updates.

For LDA, there are two additional prior distributions:

$$\beta_k \sim \text{Dirichlet}(\eta) \tag{5}$$

where $\beta_k \in \mathbb{R}^{|W|}$ is a distribution over words for each topic. Besides, for each document d

$$\xi_d \sim \text{Dirichlet}(\alpha) \tag{6}$$

where $\xi_d \in \mathbb{R}^K$ is a distribution over topics. $\eta, \alpha \in \mathbb{R}$ are two scalar hyperparameters of the model, which define the Dirichlet priors to be symmetric.

The variational inference of LDA also requires the variational posteriors of $q(\beta_k)$ and $q(\xi_d)$. They are;

$$q(\beta_k) = \text{Dirichlet}(\lambda_k) \tag{7}$$

and

$$q(\xi_d) = \text{Dirichlet}(\gamma_d) \tag{8}$$

where $\lambda_k \in \mathbb{R}^{|W|}$ and $\gamma_d \in \mathbb{R}^K$ are vector parameters. In terms of θ^t, $\lambda = \bar{\theta}$ since it does not depend on time t, and $(\gamma^t, \phi^t) = \hat{\theta}^t$ since the document $d \in D^t$ depends on time t.

Algorithm 1

```
procedure OnlineModel($K, \alpha, \eta, \rho^0, \bar{\theta}^0, \pi^0$)
    for $t \in T$ do
        $y^t$ has been revealed at time $t$.
        while Change of $|\frac{1}{K}\sum_k \gamma^t_{dk}| \le 0.00001$ do
            $\phi^t_{dwk} \propto \exp(\mathbb{E}_q\{\beta_{kw}\} + \mathbb{E}_q\{\xi_{dk}\} - \frac{y^t}{\pi^t_k})$
            Update $\hat{\theta}^t$ based on [5].
        end while
        Solve $\bar{\theta}^t_\Delta, \pi^t_\Delta = \arg\max \quad g(\theta^t, \pi) - \|\pi\|^2$
        s.t. $\sum_k \pi_k = 1$
        $\bar{\theta}^t = (1-\rho^t)\bar{\theta}^{t-1} + \rho^t\bar{\theta}^t_\Delta$
        $\bar{\pi}^t = (1-\rho^t)\bar{\pi}^{t-1} + \rho^t\bar{\pi}^t_\Delta$
        Update $\rho^{t+1}$
    end for
end procedure
```

Assuming a learning rate ρ^t is given to us, our online algorithm is based on the incremental updates from each subproblem's optimal at t. For each $t \in [T]$, the subproblem is:

$$\begin{aligned} \bar{\theta}^t_\Delta, \pi^t_\Delta = \arg\max \quad & g(\theta^t, \pi) - \|\pi\|^2 \\ \text{s.t.} \ & \sum_k \pi_k = 1 \end{aligned} \tag{9}$$

where a ℓ-2 regularizer, $\|\pi\|^2$, is added to the objective. The incremental updates for these two time-independent variables are:

$$\bar{\theta}^t = (1-\rho^t)\bar{\theta}^{t-1} + \rho^t\bar{\theta}^t_\Delta \tag{10}$$

$$\bar{\pi}^t = (1-\rho^t)\bar{\pi}^{t-1} + \rho^t\bar{\pi}^t_\Delta \tag{11}$$

In addition, solving ϕ^t_d is different from the above online problem as the solution in nature depends on time t. The subproblem for ϕ^t given π^t and θ^t attained from the updates is:

$$\begin{aligned} \max \quad & g(\theta^t, \pi^t) \\ \text{s.t.} \ & \sum_k \phi^t_{dwk} = 1, \ \forall d \in D^i, \ \forall w \in W \end{aligned} \tag{12}$$

The above subproblem has a closed-form solution as well.

$$\phi^t_{dwk} \propto \exp(\mathbb{E}_q\{\beta_{kw}\} + \mathbb{E}_q\{\xi_{dk}\} - \frac{y^t}{\pi^t_k}) \tag{13}$$

The above equation implies that if $y^t = 1$, and the topic weight for confirmation π^t_k is small, ϕ^t_{dwk} tends to be zero. If $y^t = 0$, we recover the same update as in [3, 5]. Algorithm 1 describes all computations for each $t \in [T]$. Overall, we repeat the computations of the two time-dependent parameters ϕ^t_d and γ^t until the convergence of γ^t is satisfied.

5 EXPERIMENTS

To verify the effectiveness of our augmented model, we experimented with real-world data. We considered two standard probabilistic models, LDA and MUG, due to their efficient variational inference [3, 5]. Importantly, our method is compatible with any probabilistic topic model featuring a multinomial per document-topic variational posterior distribution, a structure found in many existing models [7, 21, 23]. Future research may explore integrating our approach with other topic models.

We employed our model for two key downstream tasks in information science: data labeling and similarity measurement. The first task involves obtaining interpretable topic representations and assessing clustering correlation with human-generated labels. The second task leverages the representational space to reveal semantic content similarities and dissimilarities between documents.

5.1 Data and Ground Truth Labels

Hurricane-related Tweets. We will conduct experiments on a real-world Twitter dataset during Southern US hurricanes to evaluate disaster information augmentation in real-world applications [13, 15]. The dataset includes manually generated labels for five different classes. The data set consists of geotagged tweets collected from three hurricanes that occurred in Florida in 2020: Hurricane Eta (31/10/2020-14/11/2020), Hurricane Isaias (31/7/2020-4/8/2020), and Hurricane Sally (14/9/2020-28/9/2020). The entire data set consists of 10,210 tweets, which are evenly distributed over the first four categories of events as below:

(0) Broadcast/News - Includes tweets related to news, government updates, alerts, and official sources information.
(1) Power - Includes tweets related to power outages, power lines/systems, lights, Wi-Fi, Internet connectivity, etc.
(2) Traffic Incident - Includes tweets related to car crashes, road congestion, evacuations, traffic updates and incidents.
(3) Forecast/Weather - Includes tweets related to weather conditions, forecasts, rainfall, flooding, etc.
(4) Miscellaneous - Includes tweets that do not fit into other categories or are unrelated to the disaster.

Waze. A Waze alert is in a standardized schematic of *Type* plus *Subtype* and *Description*. An example of a traffic alert is provided:

alert: Traffic Accident, Minor Accident $onI - 75, Rear - end$

The content of these alerts tends to be similar due to the limited format and the specific categories used to describe the incidents. This categorization enables rapid identification of the alert's nature, including incident type (e.g., accident) and severity (e.g., minor).

5.2 Metrics

Perplexity. We use perplexity on out-of-sample data as as a model fit measure [5]. Perplexity is defined as the geometric mean of the inverse marginal likelihood of each word in the tweet set.

Topic Coherence. The two downstream tasks necessitate that topic representations are interpretable for readers. Topic Coherence (TC) measures the degree of semantic similarity among high-scoring words (top 15 in our case) [12]. We employ the "Umass" version of TC [8], which calculates the word-wise score function based on the document co-occurrence of the two words. The overall score is obtained by summing the score of every word-word pair and taking the average among all topics.

Adjusted Mutual Information. Normalized Mutual Information measures the agreement between two clusters and quantifies the similarity between two cluster assignments [17]. In our experiment, we have labeled tweets in 5 categories, denoted as $C \in [K]$. To match the number of classes, we set $K = 5$ resulting in an assignment score into 5 clusters. For each tweet $d \in D$, we assign it to the cluster with the highest score denoted as $C^{'} \in [K]$. The mutual information is measured as:

$$MI(C, C^{'}) = \sum_{i=1}^{K} \sum_{i=1}^{K} \frac{|C_i \cap C_j^{'}|}{|D|} \log_2 \frac{|D||C_i \cap C_j^{'}|}{|C_i||C_j^{'}|}$$

Normalized Mutual Information (NMI) is a normalization of mutual information, which scales the score between 0 and 1. A higher NMI score indicates a higher level of agreement between the two clusters. The Adjusted Mutual Information (AMI) is an extension of NMI that takes into account the size of clusters, making the score independent of size K.

Recall. The recall score of a binary classification model is computed as follows:

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

5.3 Experiment Designs

We conducted our experiments in an online machine-learning setting using real-world data but with simulated streams for event types of *Traffic Incident* and *Forecast Weather*. In the simulation, at each time step t, a batch of tweets and a binary label y^t are sampled. The batch size is uniformly distributed between 10 and 15. If $y^t = 1$, relevant tweets related to the targeted event type were included in the batch, ranging from 5 to 1, while the remaining tweets were randomly selected from the *Miscellaneous* class. The label y^t was only revealed to the model after t. Additionally, to test the robustness of the model, the accuracy of y^t could be compromised to some extent. The overall expectation of the model augmented with the confirmation we proposed is that it will gradually outperform the baseline model in all evaluation metrics regardless of the perturbation in y^t and D^t.

Hyper-parameters. In our online machine learning framework we consider several hyper-parameters that impact the performance of the model: (1) K the number of topics, (2) ρ^t, the step size of updating parameters at each step, (3) N, the number of runs of each experiment, (4) eta, the hyperparameter of the prior β, and (5) α, hyperparameter of the prior ξ. Throughout the experiments, we take $\rho^t = (t + 2)^{-0.7}$ for $t \in [T]$ and $\alpha, \eta = \frac{1}{K}$. We enumerate K from the set of $\{5, 12, 15, 21\}$. Let $N = 30$. All experimental results are aggregated from the N samples. For y^t, we test a label accuracy of $\{70\%, 80\%, 90\%\}$. The experiments are performed on a computer with AMD Ryzen 7 5700G 3.80GHz CPU, 16GB memory, and Nvidia Ge-Force RTX 3060 graphics.

Labeling Data. A core task task in employing topic modeling in information studies is to label data based on the clusters inferred by the models [13, 15]. we adhered to the standard practice of randomly reserving 10% of the data as a test set for each run. We then evaluated the AMI between the ground truth labels and the cluster assignments obtained from the topic model, considering a total of $K = 5$.

Matching Similar Tweets. For each Waze alert, one of the experiments in this study is to match its content with tweets found in a nearby spatial-temporal window. We, therefore, treat a Waze alert as another tweet and compute the cosine similarity scores of every other tweet within the specified window.

For each $y^t = 1$ revealed afterward, we retrieve the top 5 tweets with the highest similarity scores, considering them as the predicted relevant tweets. We assess the success of this matching process by reporting the recall. Note that we do not compare the models based on precision. This is because the baseline methods employed in this study are inherently unsupervised, meaning they do not utilize y^t. Consequently, the unsupervised topic model consistently output matched tweets, regardless of y^t, which results in low precision.

Baseline. To demonstrate the effectiveness of the augmented models, we also include a baseline model with:

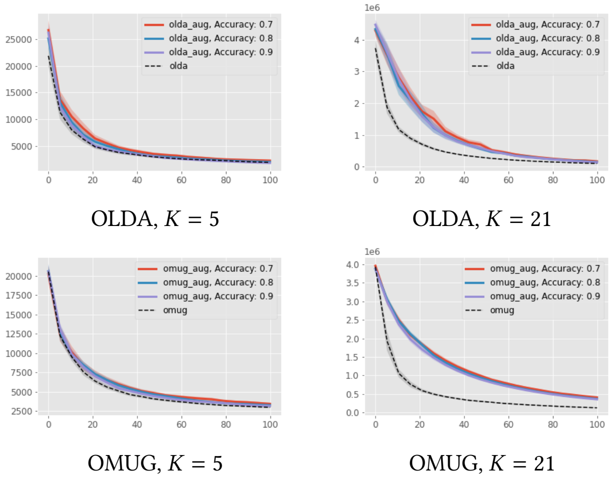
$$\pi_k = \frac{1}{K}$$

the parameters remain constant throughout the online experiment. We use Online Mixture of Unigrams (OMUG) and Online Latent Dirichlet Allocation (OLDA) to denote the baseline models.

5.4 Results

Convergence and Fit. The perplexity of all models is presented in Figure 2.Each sub-figure represents instances of LDA and MUG, along with their respective 95% confidence intervals. The results indicate that the convergence criteria are satisfied. The inclusion of the augmented linear reward function and the new posterior does not impede the training of the topic models. All models exhibit a good fit when measured on the held-out evaluation dataset. In general, when the value of K is large, we can anticipate that the model will require more computations to converge.

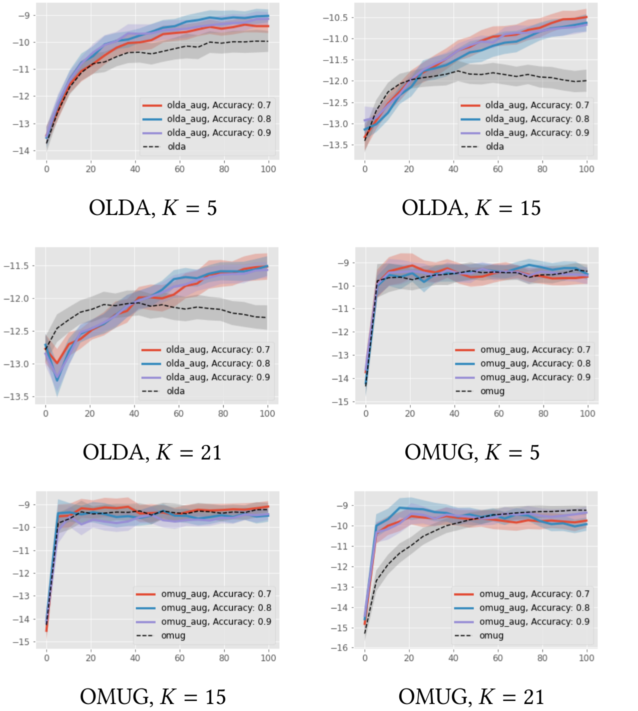
For $K = 21$, the LDA model augmented with the component we proposed takes 60 timestamps, denoted as t, to reach a similar perplexity level as a baseline LDA model using variational inference with 40 timestamps. However, for MUG model, when $K = 21$, the augmented version never reaches the same level of perplexity as the baseline within 100 iterations. In addition, LDA exhibits slower convergence when the accuracy of y^t goes down. In contrast, MUG demonstrates robustness under the uncertainty of y^t.



OLDA, $K = 5$ OLDA, $K = 21$

OMUG, $K = 5$ OMUG, $K = 21$

Figure 2: The Perplexity of Models in t



OLDA, $K = 5$ OLDA, $K = 15$

OLDA, $K = 21$ OMUG, $K = 5$

OMUG, $K = 15$ OMUG, $K = 21$

Figure 3: The Topic Coherence of Models in t. The Y-axis is the Perplexity. X-axis is the number of timestamps t.

It's important to note that the convergence on the held-out dataset does not necessarily prevent overfitting of the models. This consideration holds significance when analyzing the other experimental results presented below.

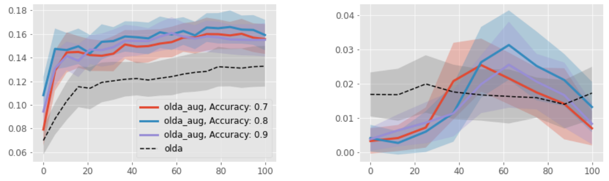
Interpretability of Topics. Our second remark is that our method significantly improves the semantic interpretability of topics in terms of topic coherence. As shown in Figure 3, the augmented LDA significantly outperforms the baseline LDA after 50-60 iterations. Moreover, the advantage of our augmentation method becomes more prominent as K increases. When $K = 15$ and $K = 21$, the baseline LDA tends to be overfitted after 50 iterations as evidenced by a continuous decrease in TC. The augmented LDA does not have this issue under all settings.

With regard to MUG, it's worth noting that a baseline MUG already outperforms LDA, with the optimal value ranging from -9 to -9.5. This observation aligns with previous studies on topic models for short texts[11]. The one-text-one-topic assumption of short text is more reasonable than a mixture of topics, especially for tweets [21]. Nevertheless, our method enjoys faster convergence in terms of TC, usually within 10 iterations when K is large. For small K, the increase is marginal. The overfitting problem arises again as there is a drop of TC in augmented MUG $K = 21$ after 80 iterations.

In summary, our method transforms LDA, which was not initially considered suitable for short texts, into a competitive model compared to MUG.

Labeling Data for Classifications. To evaluate the effectiveness of the proposed method on labeling data, we computed the AMI between ground truth labels and cluster assignments of the held-out data set for $K = 5$. Figure 4 presents the results of the AMI for all models. It is evident that augmented LDA consistently outperforms the baselines. Augmented MUG, however, peaks at 60-80 iterations, then dramatically decreases and becomes worse than the baseline model. We identify the drop as another evidence of the overfitting of MUG. The results suggest to use of LDA for data labeling, supported by empirical evidence of superiority, and is consistent with [13, 15].

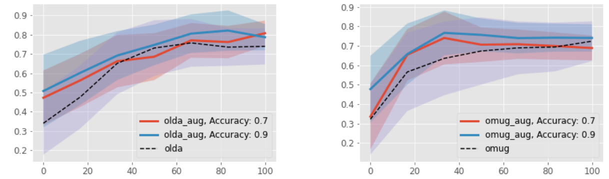
Interestingly, our method under the worst y^t, a 70% of accuracy, has the optimal AMI in both LDA and MUG. We investigated this observation in-depth and provided explanations for it. First, note that the augmented LDA with 80% accuracy is the worst case, which confirms that a higher accuracy does improve the model in terms of AMI. Both 80% and 90% suffer from overfitting as indicated by the nearly identical drop in MUG. The presence of 30% of noisy tweets accidentally expands the training data set, which mitigates the over-fitting issue and contributes to an LDA and a delay of drop in MUG. The analysis suggests that our results can be sensitive to data and hyper-parameters, such as the incremental update rate ρ^t. We will discuss it in the limitations section.



OLDA, $K = 5$ OMUG, $K = 5$

Figure 4: The Adjusted Mutual Information of Models in t

Similarity Measure. Figure 5 shows the recall of matching similar tweets with a single Waze alert. For better visualization, we only present two cases of accuracy, 90% and 70%. The case of 80% is omitted since it closely resembles the other cases in LDA and it overlaps with the baseline in MUG. The experiment of matching similar tweets is greatly dependent on chosen hyperparameters, particularly on K and label accuracy. As seen, the augmented LDA outperforms the baseline LDA across all uncertainties when $K = 12$. The recall is compromised by the lower accuracy of y^t and it is below the baseline for a 70% of accuracy. Other K either have slightly inferior performance or marginal merit compared to the baseline. The augmented MUG dominates the baseline under 90% accuracy of y^t and it is marginally enhanced compared to the baseline. Overall, the experimental findings highlight the influence of hyperparameters on the performance of matching similar tweets and reinforce the advantages of our augmented models over baselines.



OLDA, $K = 12$ OMUG, $K = 21$

Figure 5: Recall of Matching Similar Tweets in Models at t

6 CASE STUDY

To demonstrate the effectiveness and potential implementation of the proposed method in the real world, we also perform a case study using real-world data streams.

We utilize an unlabeled data set consisting of Waze data and tweets from 10/18/2021 to 10/31/2021 within 115 bounding boxes covering Georgia, USA. The entire dataset consists of about 12,000 events and over 600,000 geo-tagged tweets with lengths greater than 5. For each Waze incident, we recorded the debut time, the current time of API calls, and the number of thumbs-ups from other users. We consider tweets as potential matches if their debut time is within a 30-minute time window and in the same bounding box. To label the matches, we set $y^t = 1$ if the number of thumbs-ups is above 3, and $s_t = 0$ otherwise. For each Waze alert with $y^t = 1$, we apply an LDA model, $K = 21$, augmented with the confirmation part to output the top four similar tweets. We randomly sampled five instances from the results, and they are presented in Table 2.

Table 2 shows successful cases of implementing the augmented LDA with real-world data. Nevertheless, it is important to note that the bounding boxes used in the filtering process do not completely eliminate other incidents that may be spatially and temporally close to the target incident.

Waze Events	Relevant Tweets
alert ACCIDENT ACCIDENT MINOR I-575 S	Accident. left two lanes blocked in Cherokee on I 575 SB after Sixes Rd/Exit 11 ATLTraffic @MARTAservice there are multiple lights out on the walkway from Buckhead station towards Tower Place. He shoots under par and places at the national and regional tournaments ...
alert ACCIDENT ACCIDENT MAJOR I-85 S	Accident. left two lanes blocked in Hapeville on I-85 SB near Sylvan Rd/Central Ave/Exit 75, stop and go traffic b… Disabled vehicle, shoulder blocked in CollegePark on I-285 WB near I-85 (SW ATL)/Exit 61 (WB), stop and go traffic… Join the Lane Construction team! See our latest job opening here: https://xxxxxx Construction Craft-Workers ...
alert ACCIDENT ACCIDENT MAJOR I-285 E	Accident. right three lanes blocked in SandySprings on I-285 EB at Roswell Rd (GA-9)/Exit 25 (EB), stopped traffic… Accident. right shoulder blocked in Dekalb on I-285 SB at Lawrenceville Hwy (US-29)/Exit 38, stop and go traffic b… Cozy Cabin Overlooks the Suwannee River: A North Florida couple builds their family-friendly forever home along the… ...
alert traffic JAM JAM HEAVY TRAFFIC I-85 N	McDonough traffic might be the reason I actually atl cntrl delt Accident, right lane blocked in Norcross on I-85 SB at Jimmy Carter Blvd (GA-140)/Exit 99, stop and go traffic bac… Accident in Brevard on US 1 Both NB/SB between CO Hwy 502/Coquina Rd/Barnes Blvd and Eyster Blvd traffic ...
alert traffic JAM JAM HEAVY TRAFFIC I-75 S	Accident, left lane blocked in Snellville on Stone Mtn Fwy (Hwy 78) WB at Scenic Hwy (GA-124) ATLTraffic https://t.co/bABElTW6T2 Accident. two left lanes blocked. in Polk on I-4 EB before US 27 (MM 55) traffic Closed due to accident in Osceola on Poinciana Blvd SB south of US 17-92/Orange Blossom Trail and before Reaves Rd… Accident. left three lanes blocked in Jonesboro on I-75 SB after Tara Blvd/Old Dixie Hwy/Exit 235 ATLTraffic ...

Table 2: Relevant Tweets of Bounding Box 1, *Accident* and *Jam*. Blue texts indicate a truth positive, Red texts describe incidents that occurred in nearby bounding boxes.

7 LIMITATIONS AND FUTURE WORKS

One of the limitations of this study is the relatively limited exploration of hyperparameters. We acknowledge that the choice of hyperparameters, including the update weight $\rho^t = (t+2)^{-0.7}$, is somewhat arbitrary and not optimized for all models. This update rate is not suitable for MUG as it resulted in overfitting within just 50 iterations. The update rate should decay at a much faster rate compared to the one used for LDA. Additionally, other initialization hyper-parameters, such as ρ^0, η, and α, can all impact the performance of the two downstream tasks. In general, these hyperparameters should be fine-tuned for each specific model augmented with our proposed component. However, we adopted the default settings from[3, 5]. In addition, our exploration was limited to a small parameter set, with K values chosen from 5, 12, 15, 21, due to the considerable time required to complete each experiment (around 1.5 hours in average). We recognize the need for a more thorough investigation of hyperparameters as a future endeavor. This would entail refining, parallelizing, and addressing numerical issues in the current code implementation.

Another perturbation that our method is sensitive to is the accuracy of y^t. Initially, we anticipated that our method could still perform accurately with a 55% accuracy. However, the results indicate that an accuracy less than 70% will significantly compromise the performance of our method, see Figure 5 for the case of OMUG. We suspect that the bi-linear function may not be robust enough to handle noise in y^t. Exploring alternative stable functions as potential augmentation components is an avenue for future research.

8 CONCLUSION

Our research highlights the critical role of online probabilistic topic models in enabling the real-time analysis of complex data streams in Cyber-Physical Social Infrastructure Systems (CPSIS). These models empower infrastructure operators and decision-makers to extract actionable insights, detect anomalies, make precise predictions, optimize resource allocation, engage users, and leverage social feedback. However, traditional probabilistic topic models face challenges when applied to user-generated content, which is often sparse and dynamic. We propose a novel framework that integrates a linear reward function, guided by the confidence levels associated with relevant content, into the variational lower bound of the likelihood of Bayesian topic models. This innovative approach enhances topic retrieval, improving interpretability and generalizability across various topic models. Our empirical experiments and case study, conducted on real-world datasets, showcase the effectiveness of our learning algorithm in enhancing topic models through two important downstream tasks: information augmentation and event detection. It significantly improves topic interpretability, data labeling precision, and similarity metric refinement, making it a valuable tool for processing and analyzing real-time data streams in CPSIS. The effectiveness of our online confirmation-augmented probabilistic topic modeling approach for processing CPSIS real-time data streams contributes to informed decision-making, efficient infrastructure management, and proactive engagement with evolving CPSIS conditions. In the evolving field of CPSIS, our approach shows potential for unlocking new insights and addressing integration challenges.

ACKNOWLEDGMENTS

This material is based upon work supported by the Georgia Department of Transportation (GDOT) under Grant No. RP 20-13. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Georgia Department of Transportation.

REFERENCES

[1] Jacob Abernethy, Chansoo Lee, Abhinav Sinha, and Ambuj Tewari. 2014. Online linear optimization via smoothing. In *Learning Theory*. PMLR, 807–823.
[2] Charu Aggarwal. 2011. *An introduction to social network data analytics*. Springer.
[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[4] Chao Fan, Fangsheng Wu, and Ali Mostafavi. 2020. A hybrid machine learning pipeline for automated mapping of events and locations from social media in disasters. *IEEE Access* 8 (2020), 10478–10490.
[5] Matthew Hoffman, Francis Bach, and David Blei. 2010. Online learning for latent dirichlet allocation. *advances in neural information processing systems* 23 (2010).
[6] A. Khalid. 2019. Twitter removes precise geo-tagging option from tweets. https://www.engadget.com/2019-06-19-twitter-removes-precise-geo-tagging.html.
[7] Tianyi Lin, Wentao Tian, Qiaozhu Mei, and Hong Cheng. 2014. The dual-sparse topic model: mining focused topics and focused terms in short text. In *Proceedings of the 23rd international conference on World wide web*. 539–550.
[8] 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*. 262–272.
[9] Fred Morstatter, Shamanth Kumar, Huan Liu, and Ross Maciejewski. 2013. Understanding twitter data with tweetxplorer. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1482–1485.
[10] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine learning* 39 (2000), 103–134.
[11] Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. 2020. Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering* 34, 3 (2020), 1427–1445.
[12] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*. 399–408.
[13] Christin Salley, Neda Mohammadi, and John E Taylor. 2022. Semi-Supervised Machine Learning Framework for Fusing Georeferenced Data from Social Media and Community-Driven Applications. In *Comput. in Civil Engineering 2021*. 114–122.
[14] David Sontag and Dan Roy. 2011. Complexity of inference in latent dirichlet allocation. *Advances in neural information processing systems* 24 (2011).
[15] Iris Tien, Aibek Musaev, David Benas, Ameya Ghadi, Seymour E Goodman, and Calton Pu. 2016. Detection of Damage and Failure Events of Critical Public Infrastructure using Social Sensor Big Data.. In *IoTBD*. 435–440.
[16] Ike Vayansky and Sathish AP Kumar. 2020. A review of topic modeling methods. *Information Systems* 94 (2020), 101582.
[17] Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary?. In *26th annual international conference on machine learning*. 1073–1080.
[18] Jun Wang, Limin Xia, Xiangjie Hu, and Yongliang Xiao. 2019. Abnormal event detection with semi-supervised sparse topic model. *Neural Computing and Applications* 31 (2019), 1607–1617.
[19] Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. 2012. Automatic Crime Prediction Using Events Extracted from Twitter Posts. *SBP* 12 (2012), 231–238.
[20] Yan Wang and John E Taylor. 2018. Urban crisis detection technique: A spatial and data driven approach based on latent Dirichlet allocation (LDA) topic modeling. In *Construction Research Congress 2018*. 250–259.
[21] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*. 1445–1456.
[22] Xing Yi and James Allan. 2009. A comparative study of utilizing topic models for information retrieval. In *Advances in Information Retrieval: 31th European Conference on IR Research, ECIR 2009, Toulouse, France, April 6-9, 2009. Proceedings 31*. Springer, 29–41.
[23] Jianhua Yin and Jianyong Wang. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 233–242.
[24] Chao Zhang, Liyuan Liu, Dongming Lei, Quan Yuan, Honglei Zhuang, Tim Hanratty, and Jiawei Han. 2017. Triovecevent: Embedding-based online local event detection in geo-tagged tweet streams. In *23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 595–604.