Tracking Disaster Footprints with Social Streaming Data

Lu Cheng,[1] Jundong Li,[2,3] K. Selçuk Candan,[1] Huan Liu[1]
[1]Computer Science and Engineering, Arizona State University, USA
[2]Department of Electrical and Computer Engineering, University of Virginia, USA
[3]Department of Computer Science & School of Data Science, University of Virginia, USA
{lcheng35, candan, huanliu}@asu.edu, jundong@virginia.edu

Abstract

Social media has become an indispensable tool in the face of natural disasters due to its broad appeal and ability to quickly disseminate information. For instance, Twitter is an important source for disaster responders to search for (1) topics that have been identified as being of particular interest over time, i.e., common topics such as "disaster rescue"; (2) new emerging themes of disaster-related discussions that are fast gathering in social media streams (Saha and Sindhwani 2012), i.e., distinct topics such as "the latest tsunami destruction". To understand the status quo and allocate limited resources to most urgent areas, emergency managers need to quickly sift through relevant topics generated over time and investigate their commonness and distinctiveness. A major obstacle to the effective usage of social media, however, is its *massive* amount of *noisy* and *undesired* data. Hence, a naive method, such as set intersection/difference to find common/distinct topics, is often not practical. To address this challenge, this paper studies a new topic tracking problem that seeks to effectively identify the *common* and *distinct* topics with *social streaming data*. The problem is important as it presents a promising new way to efficiently search for accurate information during emergency response. This is achieved by an online Nonnegative Matrix Factorization (NMF) scheme that conducts a faster update of latent factors, and a joint NMF technique that seeks the balance between the reconstruction error of topic identification and the losses induced by discovering common and distinct topics. Extensive experimental results on real-world datasets collected during Hurricane Harvey and Florence reveal the effectiveness of our framework.

Introduction

Social media has become a critical platform for real-time information seeking for disaster relief, ranging from pre-disaster, warning, threat to rescue and recovery (Nazer et al. 2017; Houston et al. 2015; Gao, Barbier, and Goolsby 2011). As a new way of communication in the course of a disaster, the major difference between social media and traditional sources is its real-time nature. Disasters and emergencies often speed up and amplify the quantity of information in social media, as a result, understanding social media streams is crucial for disaster relief and management.

Social media reveals dynamic changes of discussions with topics evolving over time. Take the Asia tsunami disaster as an example, major topics of the reports evolved from "financial aids" to "debt" and "reconstruct" over different stages (Cao et al. 2007). Online topic tracking can benefit disaster responders in the following ways: (1) For emergency managers and people affected by the natural calamities, it is often of particular interest to identify topics that prevail over time, i.e., common topics, such as "disaster rescue", as well as to be alerted to any new emerging themes of disaster-related discussions that are fast gathering in social media streams (Saha and Sindhwani 2012), i.e., distinct topics such as "the latest tsunami destruction". (2) For global participants, a quick update of the disaster status-quo, i.e., the commonness and distinctiveness between previous and current topics, is necessary for them to provide immediate and effective assistance. A major obstacle to disaster-related topic tracking, however, is that social media generates massive amount of data each day and it is notorious for a sea of unwanted and noisy content such as spam and daily chatter. For example, during Hurricane Harvey, Twitter reported there have been 21.2 million hurricane-related tweets within the first six days and a large portion was generated in a short period of time to spread rumors (Nazer et al. 2017). Consequently, a new way of effective online topics discoveries using social media data during disaster response is urgent.

In this paper, we study a novel topic tracking problem that seeks to identify *common* and *distinct* topics using social streaming data related to disasters. Discovering the commonness and differences between topics in an *online* fashion provides an effective and efficient way for information seekers to search for both prevailing and emerging topics. For instance, emergency managers can make informed decisions about how to effectively allocate funds and other resources to areas that need most assistance by comparing the commonness and distinctiveness of topics generated from these areas over time. We illustrate the problem in Fig. 1. The goal is to discover topics from the historical and incoming data, and identify their commonness and distinctiveness.

However, the proposed problem presents several challenges: (1) Acquiring insights via social media needs to process enormous amounts of noisy data in a timely fashion

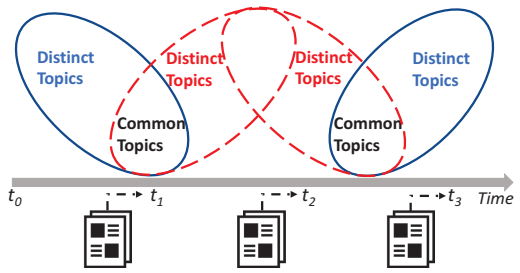


Figure 1: An illustration of the studied problem. New social media data continuously arrives at each time t. By comparing social media data generated over different periods of time, the common and distinct topics can be efficiently identified in an online manner.

(Cao et al. 2007). There were 1,200 tweets posted per minute from Tokyo after Japan earthquake and tusanmi (2011) and 16,000 tweets per minute at the peak of Hurricane Sandy (2012) (Meier 2015). Consequently, models for online topic tracking should be computation-efficient and storage-saving. (2) Discovering common and distinct topics along time entails the model to simultaneously compute the commonness and differences of topics extracted from the historical and incoming data in an online fashion. The second challenge is, therefore, how to efficiently identify meaningful topics from social streaming data meanwhile jointly model the commonalities and differences between these topics.

To address these issues, in this paper, we propose an online topic tracking approach - Tracking Disaster Footprints (TDF) with social streaming data. TDF consists of two major components: An online Nonnegative Matrix Factorization (NMF) scheme that conducts fast update of latent factors and a joint NMF technique that seeks the balance between the reconstruction error of topic identification and the losses induced by discovering common and distinct topics. Existing work on online topic tracking (Cao et al. 2007; Lehmann et al. 2012; Wang et al. 2011; Hoffman, Bach, and Blei 2010), however, cannot fully satisfy the needs as they do not explicitly model the relationships between discovered topics over time. TDF is based on NMF because it often works very well out of box for corpora of short texts such as tweets (Chen et al. 2019) and the NMF-based models have shown outstanding performance in dimension reduction and clustering for the high-dimensional data (Shi et al. 2018). The main contributions of this work are:

- **Problem**: We propose a novel problem of online common/distinct topic tracking with social streaming data for disaster relief. The core difference between the proposed problem and standard online topic modeling is that we take a step further to investigate the commonness and distinctiveness between these topics generated over time.
- **Algorithm**: We propose a new online topic tracking framework TDF that contains an online NMF and a joint NMF components. It seeks to efficiently solve NMF and simultaneously discover the common and distinct topics in an online manner.
- **Data**: We collected two real-world datasets during Hurricane Harvey (2017) and Hurricane Florence (2018) using keywords and geo-location specific methods, respectively. Datasets and select pieces of custom code are available upon request.
- **Evaluation**: We evaluate TDF on these two datasets and perform in-depth qualitative and quantitative studies. Experimental results reveal that our approach is effective and hence, has practical usage in real-world applications.

Related Work

There have been a lot of research efforts on online topic tracking with social streaming data due to its real-time nature and ability to quickly spread information. Here, we focus on LDA based and NMF based online learning algorithms (Cao et al. 2007; Wang, Agichtein, and Benzi 2012; Ding and Chen 2014; Tu et al. 2018; Zhao, Tan, and Xu 2016; Saha and Sindhwani 2012; Chen, Candan, and Sapino 2018; Chen and Candan 2014). Vaca et al. (Vaca et al. 2014) modeled the online topic discovery problem using an adapted NMF that jointly learned the topics evolution and their time dependencies. In addition to the evolution of topics, Kalyanam et al. (Kalyanam et al. 2015) considered a new dimension to the traditional topic modeling – social context. The proposed method assumed that members in the same community share similar interest in the topics. To efficiently mine streams of social text, Wang et al. (Wang, Agichtein, and Benzi 2012) proposed a Temporal-LDA by modeling the topics and topic transitions. Similar work can also be found in trend mining, where trends are defined as "set of bursty keywords that occur frequently together" (Mathioudakis and Koudas 2010) and are usually driven by events and breaking news, e.g., a natural disaster. TwitterMonitor (Mathioudakis and Koudas 2010) is a system that can detect trends/topics through identifying and clustering bursty keywords. Another approach focused on comparing the popularity of words/hashtags before and after a spike (Nazer et al. 2017). Lehman et al. (Lehmann et al. 2012) separated trends on Twitter into three classes based on the shape of the spike and provided a semantic characterization of the hashtag classes. Another notion of research related to topic modeling is meme mining. Memes are shot text that act as the signature of a topic (Leskovec, Backstrom, and Kleinberg 2009). In (Leskovec, Backstrom, and Kleinberg 2009), the authors generated a directed acyclic graph in which each node is a meme and there is an edge from meme i to j if i is shorter than j and the directed edit distance to j is less than one. The evolution of memes is analyzed based on the variations of a meme. Niculae et al. (Niculae et al. 2015) built a bipartite graph to predict the future memes of a news outlet based on its previous memes using matrix factorization.

Different from previous online topic tracking methods that *implicitly* model time dependencies between latent factors, in this paper, we take a step further and seek to *explicitly* model the dynamic relationships between the learned topic representations generated along the time. Specifically, the proposed TDF framework actively looks for common as well as emerging topics in an online manner. Our work

is built on (Kim et al. 2015), which shares similar ideas of identifying the common and distinct topics but between *two static documents*. As a result, their proposed algorithm has to store all historical data and conduct NMF from scratch whenever new data arrives. In contrast, TDF takes advantage of online NMF to conduct faster update of latent factors, therefore, is more storage-saving and computation-efficient. The goal of TDF is to provide a promising new way for information seekers to efficiently and effectively sift through topics that are of their particular interest under disaster relief where time is often a critical factor.

The Proposed Framework

In this section, we start with a brief review of the standard NMF model and a popular online NMF model (Cao et al. 2007). We then explain the core components of TDF in detail. TDF first employs this online NMF algorithm to obtain the latent factors from matrix factorization. These latent factors that encode the discovered topics from historical data, together with the newly arriving data, are then fed into a joint NMF framework to identify the common and distinct topics for disaster relief.

Preliminaries

Conventional NMF. NMF (Lee and Seung 2001; Ding et al. 2006) seeks to decompose a non-negative matrix into two low-rank non-negative matrices. Let the document-word matrix $V \in \mathbb{R}_+^{n\times d}$ contain n documents, each document is represented by a d-dimension feature vector. NMF is then formalized as:

$$V \approx WH, \tag{1}$$

where $W \in \mathbb{R}_+^{n\times k}$ is the coefficient matrix such that each row encodes the document as a weighted combination of k topics, and $H \in \mathbb{R}_+^{k\times d}$ is the basis matrix, where each row denotes the word distribution in each topic. The NMF problem is solved by the following optimization problem:

$$\min_{W,H\geq 0} \frac{1}{2}\|V - WH\|_F^2. \tag{2}$$

Online NMF (ONMF). A naïve solution to find topics from streaming data is to apply NMF repeatedly on the incoming data batch and perform aggregation later. While this method could save computational cost, it overlooks the time dependencies between the decomposed latent factors. Instead, following (Cao et al. 2007), we leverage the information from previously learned latent factor H and combine it with the new batch of data that arrives at the current time stamp. We apply NMF to this new data matrix.

Suppose that $V_t \in \mathbb{R}_+^{n_t\times d}$ records the historical data we received from the starting time till time t. Then the objective function of NMF at t is defined as:

$$\min_{W_t,H_t\geq 0} \frac{1}{2}\|V_t - W_tH_t\|_F^2. \tag{3}$$

Consider a new batch of data $U \in \mathbb{R}_+^{p\times d}$ arrives at time $t+1$. Then the factorization at $t+1$ will be:

$$V_{t+1} = \begin{pmatrix} V_t \\ U \end{pmatrix} \approx W_{t+1}H_{t+1}. \tag{4}$$

The goal of online NMF (ONMF) (Cao et al. 2007) is to efficiently update W_{t+1}, H_{t+1} without storing V_t and conducting matrix factorization from scratch.

To speed up the computation, we replace the data matrix V_t with the learned latent factor H_t obtained from Eq. (3):

$$\begin{pmatrix} H_t \\ U \end{pmatrix} \approx \begin{pmatrix} W_t^* \\ W_U \end{pmatrix} H_{t+1}, \tag{5}$$

where W_t^* is a $k \times k$ non-negative matrix that captures the correlation between H_t and H_{t+1}. $W_U \in \mathbb{R}_+^{p\times k}$ is the discovered topics associated with U. From Eq. (5), we have $H_t \approx W_t^*H_{t+1}$ and $U \approx W_UH_{t+1}$. Plugging it in $V_t \approx W_tH_t$ at time t, we get the following:

$$V_t \approx W_tW_t^*H_{t+1}. \tag{6}$$

Thus, we can reformulate the factorization in Eq. (4) with the equation below:

$$V_{t+1} \approx \begin{pmatrix} W_tW_t^* \\ W_U \end{pmatrix} H_{t+1} = W_{t+1}H_{t+1}. \tag{7}$$

According to the *Full-Rank Decomposition Theorem* in (Cao et al. 2007), the update rules for W_{t+1}, H_{t+1} can then be summarized as

$$W_{t+1} = \begin{pmatrix} W_tW_t^* \\ W_U \end{pmatrix},\ H_{t+1} = W_t^{*-1}H_t. \tag{8}$$

Tracking the Topic Evolution

Previous section presents a simple approach that can efficiently update the document-topic and topic-word latent factors in NMF. Nevertheless, this approach will not explicitely seek *common topics*, i.e., topics that appear both before t and at $t+1$, along with *distinct topics*, i.e., two sets of topics that are unique to data generated before t and that at $t+1$, respectively. Here, we take a step further and provide an in-depth investigation of the relationships between the discovered topics. Our model is built upon (Kim et al. 2015), which attempts to discover common and discriminative topics from two *static* text corpora. However, as we focus on tracking topics with disaster-related social streaming data, the method proposed in (Kim et al. 2015) cannot be directly applied to our problem due to its high computational cost and storage demand.

Suppose that there are k hidden topics in the documents, we denote as k_c the number of common topics we aim to identify, and as $k_d(= k - k_c)$ the number of distinct topics that are of particular interest. One may observe that a large memory storage and computational cost are in need to obtain the decomposed factors when V_t (the historical accumulated documents) becomes larger. To address this issue, here, we leverage the output H_t from ONMF (Cao et al. 2007) which gives a succinct topic summarization of the information embedded in V_t. Together with the incoming data U, we aim to discover the common and distinct topics between V_t and U.

Nevertheless, H_t that is incrementally updated by ONMF cannot be directly applied to find the common and distinct topics as it has been fixed at the new time stamp $t+1$. Therefore, we perform a linear transformation on H_t, i.e.,

$H^* \approx L^* H_t$ so that in the new transformed feature space, we can find common and distinct topics along with U (in particular the factorized topic matrix H_U from U). Here, $L^* \in \mathbb{R}_+^{k \times k}$ is the transformation matrix and is used to dynamically adjusts the dependency between H_t and U. Specifically, we let the first k_c topics in H^* and H_U be the common topics and the rest k_d be the distinct topics. To this end, we are looking for a joint NMF model that seeks to: 1) transform H_t to a new feature space H^*; 2) minimize the reconstruction error of NMF on U, i.e., $U \approx W_U H_U$; 3) minimize the distances between k_c topic representations in H^* and H_U; 4) maximize the distances between k_d topic representations in H^* and H_U.

Consequently, the objective function of the joint NMF at $t+1$ is defined as follows:

$$\min_{\substack{W_U, H_U, \\ H^*, L^* \geq 0}} \frac{1}{2}\|H^* - L^* H_t\|_F^2 + \frac{1}{2}\|U - W_{Uc}H_{Uc} - W_{Ud}H_{Ud}\|_F^2 + \alpha f_c(H_c^*, H_{Uc}) + \beta f_d(H_d^*, H_{Ud}), \tag{9}$$

where H_{Uc}, H_c^* are the first k_c rows in H_U and H^* respectively, and H_{Ud}, H_d^* are the rest k_d rows, i.e., $H_U = \begin{pmatrix} H_{Uc} \\ H_{Ud} \end{pmatrix}$, $H^* = \begin{pmatrix} H_c^* \\ H_d^* \end{pmatrix}$. In addition, $W_U = [W_{Uc}, W_{Ud}]$, f_c and f_d are the measures of commonness and distinctiveness between topics.

For the first term in the above formulation, we model the linear projection of H_t by minimizing the squared Frobenius norm between H^* and $L^* H_t$. The transformation enables TDF to compare the commonness between topics that are more similar and the distinctiveness between topics that are more likely to be different between V_t and U. The second term performs NMF on U where the first k_c topics are the common topics and the rest k_d topics are the distinct ones. The third term measures the distance between H_c^* and H_{Uc}, a smaller distance is desired. In particular, it is defined as

$$f_c(H_c^*, H_{Uc}) = \|H_c^* - H_{Uc}\|_F^2. \tag{10}$$

The last term in Eq. (9) represents the similarity between H_d^* and H_{Ud}, a smaller value is desired. Following (Kim et al. 2015), it is defined as:

$$f_d(H_d^*, H_{Ud}) = \|H_d^{*T} H_{Ud}\|_1. \tag{11}$$

The parameters α and β are used to control the balance between the NMF reconstruction error and the losses induced by discovering the common and distinct topics. By plugging the two terms in Eq. (10) and Eq. (11) into Eq. (9), the final objective function is then:

$$\min_{\substack{W_U, H_U, \\ H^*, L^* \geq 0}} \frac{1}{2}\|H^* - L^* H_t\|_F^2 + \frac{1}{2}\|U - W_{Uc}H_{Uc} - W_{Ud}H_{Ud}\|_F^2 + \alpha\|H_c^* - H_{Uc}\|_F^2 + \beta\|H_d^{*T} H_{Ud}\|_1. \tag{12}$$

Although both our work and (Kim et al. 2015) are based on joint NMF optimization, we highlight the following contributions compared to (Kim et al. 2015): (1) (Kim et al. 2015) takes two static documents as the input. Therefore, to conduct NMF at time $t+1$, it has to store all historical data V_t and compute W_{t+1} and H_{t+1} from scratch. This is extremely inefficient and storage expensive. Instead, we leverage ONMF and use the output H_t as a high-level succinct summarization of discovered topics in V_t. As such, TDF can handle large-scale data streams and efficiently update the latent factors when new data comes in. (2) We project the learnt H_t into a new feature space to adaptively adjust the dynamic correlation between H_t and U. This enables the proposed model to identify the common and distinct topics between two sets of documents consecutively generated over time. The pseudo code for TDF is illustrated in Algorithm 1. For optimization, we adopt the widely used multiplicative update rules (Lee and Seung 2001) to alternatively update the variables until the objective converges.

Table 1: Basic statistics of the datasets

Datasets	Methods	#Tweets	Start Date	End Date
Harvey	Keywords	171,436	08/25/2017	09/10/2017
Florence	Location	78,753	09/12/2018	10/10/2018

Algorithm 1 The proposed TDF framework.

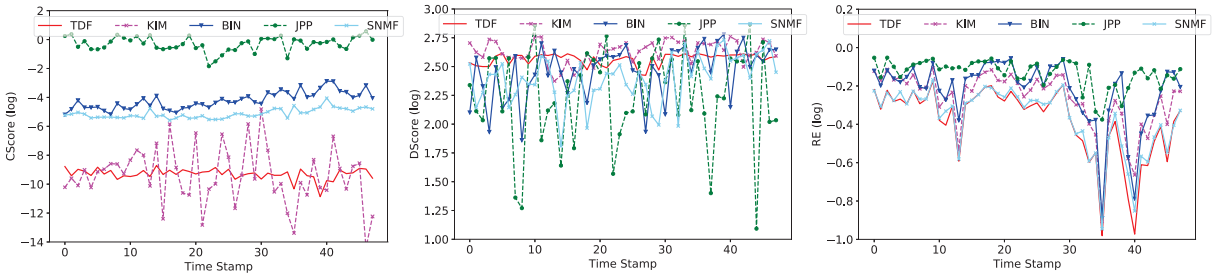
Input: The data matrix $V \in \mathbb{R}_+^{n \times d}$ at the starting time $t = 1$, the incoming data matrix $U_t, t \in 2, ..., T$, the number of topics k, the number of common/distinct topics $k_c/k_d = k - k_c$, parameters α, β.

Output: The common and different topics between $t-1$ and $t, t \in \{2, ..., T\}$.

```
Initialize W_1, H_1;
while not converge do
    Update W_1, H_1;
end while
for t = 2, 3, ..., T do
    Solve Eq. (12) with the input U_t and H_{t-1};
    Update W_t, H_t with Eq. (8).
end for
```

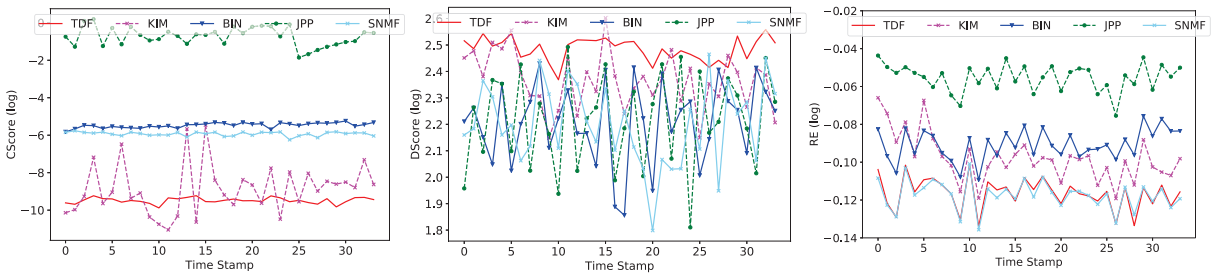
Experimental Evaluations

In this section, we conduct qualitative and quantitative analyses to evaluate the performance of TDF for finding common and distinct topics during disaster response. In particular, we first compare TDF with the standard NMF model, existing online topic modeling approaches, and a model that simultaneously discovers common and distinct topics (Kim et al. 2015). We then provide in-depth case studies for a better understanding of the specific usage of the TDF framework. To examine the robustness of the proposed framework, we further conduct sensitivity analyses on model parameters α, β, and k_c (or k_d). In particular, we aim to answer the following research questions: (1) How effective is TDF for online topic modeling, especially for the detection of common and distinct topics over time after disasters? (2) How competitive is the computational speed of the proposed framework compared to other baseline models? (3) How do the changes of model parameters affect the performance of TDF?



(a) Comparisons w.r.t. CScore (log) (b) Comparisons w.r.t. DScore (log) (c) Comparisons w.r.t. RE (log)

Figure 2: Performance comparisons of different methods using *Harvey* dataset.



(a) Comparisons w.r.t. CScore (log) (b) Comparisons w.r.t. DScore (log) (c) Comparisons w.r.t. RE (log)

Figure 3: Performance comparisons of different methods using *Florence* dataset

Datasets

We crawled real-world datasets related to two recent natural disasters – Hurricane Harvey (2017) and Hurricane Florence (2018) from Twitter[1] using the TweetTracker system[2]. TweetTracker is an analysis tool for humanitarian and disaster relief, and is capable of monitoring and analyzing location or keyword specific tweets with near real-time trending (Kumar et al. 2011). We selected the mostly used hashtags/words during Hurricane Harvey to extract related tweets for the *Harvey* dataset: *#harvey, #hurricaneharvey, #HurricaneHarveyRelief, #texas, #houston, #help, #hurricane, #relief, #houstonflood, hurricane, harvey*. The percentage of geo-tagged tweets in this dataset is 5.5%. The second dataset *Florence* was collected during Hurricane Florence in September 2018. Different from the above keyword-specific method, we crawled all geo-tagged tweets that were posted where the disaster occurred. Each tweet in this dataset is associated with a geo-location (longitude and latitude). Table 1 summarizes the basic statistics of these two datasets. Data and select pieces of custom code are available upon request.

Experimental Setup

We obtained the TF-IDF values from tweets as the input features. Entries with large TF-IDF values are the terms that occur often in particular tweets and very rarely anywhere else, i.e., important terms. For both datasets, our experiments start with 10,000 tweets and assume a batch size of 2,000 new tweets arrive at every time stamp. Values of k and k_c are set to 10 and 7 respectively, $k_d = k - k_c = 3$.

We compare TDF with the following baseline models.

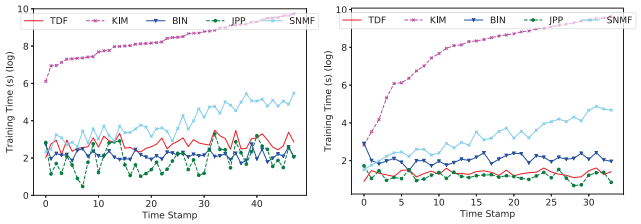
- *Standard NMF (SNMF)*: This is the basic NMF method which re-calculates the latent factors using the entire dataset each time when a new batch of data arrives. We compare the topics extracted from the historical data and the newly arriving data.
- *KIM* (Kim et al. 2015): This approach seeks to discover common and discriminative topics simultaneously given two document sets. Similarly, we take the historical data and the newly arriving data as two input documents.
- *BIN* (Cao et al. 2007): This work proposed an orthogonalized online NMF. It conducts an orthogonality constraint to guarantee the unique solution (Cao et al. 2007). Its incremental nature enables us to find the topics from the historical data and the newly arriving data.
- *JPP* (Vaca et al. 2014): This is a time-based collective factorization method for online topic discovery. It connects topics between different time slots via a $k \times k$ matrix, where k is the number of topics.

Following (Kim et al. 2015), we use reconstruction error, commonness score, and distinctiveness score to measure the performance of different methods. As all baselines are based on NMF, it is fair to make comparisons with these measures.

Reconstruction Error. The reconstruction error (RE) measures the loss of the NMF on the newly arriving data U at each time stamp. Models with smaller RE can better reconstruct the data matrix U.

[1]https://twitter.com/

[2]http://tweettracker.fulton.asu.edu/



(a) *Harvey* (b) *Florence*

Figure 4: Comparisons of computing time (log)

Commonness Score (CScore). The CScore from (Kim et al. 2015) denotes the similarity between the k_c common topics at time t and $t+1$:

$$\text{CScore} = \frac{1}{k_c}\|H_c^* - H_{Uc}\|_F^2. \tag{13}$$

A small CScore indicates a better quality for this measure.

Difference Score (DScore). Following (Kim et al. 2015), we use the averaged symmetric Kullback-Leibler divergence between all the distinct topic pairs as the DScore:

$$\text{DScore} = \frac{1}{2k_d^2}\sum_{i=1}^{k_d}\sum_{j=1}^{k_d}\big[h_d^{*(i)}\log(h_d^{*(i)})^T + h_{Ud}^{(i)}\log(h_{Ud}^{(i)})^T - h_d^{*(i)}\log(h_{Ud}^{(j)})^T - h_{Ud}^{(j)}\log(h_d^{*(i)})^T\big], \tag{14}$$

where $h_d^{*(i)}$ is the i-th row of H_d^*, $h_{Ud}^{(j)}$ is the j-th row of H_{Ud}. A large DScore indicates a better quality for this measure. For the baseline method *KIM* and our proposed method TDF, we can directly make use of the CScore and DScore as these methods explicitly specify which topics are the common/distinct ones. For other baseline models, we select k_c topic pairs that have the smallest CScore and treat them as the common topic pairs and the rest as the discriminative ones to obtain CScore and DScore for comparisons.

Quantitative Results

Fig. 2-3 present the results w.r.t. CScore, DScore, and RE (all in log scale) along the time:

- The proposed framework TDF can achieve the best balance regarding all three evaluation metrics. For example, when tested on *Florence* dataset, TDF can mostly get the smallest CScore and the largest DScore, meanwhile achieve as small reconstruction error as the standard NMF. Similar conclusion can be drawn for the *Harvey* dataset. This result manifests the advantages of incorporating the ONMF and the joint NMF modules.
- *KIM* shows competitive CScore and DScore values as well. Nevertheless, its results fluctuate widely from time to time, especially for CScore. This is mainly because *KIM* performs joint NMF on the accumulated historical data that are noisy and complex whereas the proposed model uses the latent factor H as a concise summarization of the historical data. Hence, our model is not only more computationally efficient but also more optimization friendly.
- *KIM* presents larger RE due to its joint matrix factorization of historical data and incoming data. Standard NMF can often obtain best RE because the goal of *SNMF* focuses on minimizing RE while other online topic models seek to balance between RE and computational efficiency. TDF presents very competitive RE because it separately conducts ONMF and matrix factorization on U.

In summary, TDF can effectively identify common and discriminate topics and also achieve almost least reconstruction error compared to baselines. The efficacy of leveraging ONMF and the joint NMF to explicitly model the commonness and distinctiveness, therefore, is corroborated.

Computational Cost. We further show the comparisons of different models w.r.t. running time (in log scale) in Fig. 4. Among all the online methods (i.e., *BIN, JPP,* TDF), *JPP* often achieves the fastest update of latent factors for both *Harvey* and *Florence* datasets while *BIN* and TDF also show very competitive computational efficiency. Unsurprisingly, the computational cost of *KIM* and *SNMF* increases exponentially as more data arrives. This is because they have to conduct NMF on all the data received so far in order to update the latent factors. In addition, *KIM* simultaneously optimizes two regularization terms to model the commonness and distinctiveness of topics, significantly slowing down the computing speed. In contrast, our model is online and does not need to factorize the historical data. Hence, it is much more efficient than *KIM*.

Qualitative Studies

To better understand the usage of discovered common and distinct topics over time, we further perform in-depth qualitative analyses on the *Harvey* dataset. We present in Table 2 the discovered common and distinct topics during the first five time periods during Hurricane Harvey. These topics are represented by the top ranked words returned by TDF – due to space constraints, we only present ten words.

- The common topics of tweets extracted before t_1 and new tweets posted at t_2 describe disaster-related themes such as the evacuation of residence, people praying for Texas, and family seeking assistance. Topics that are exclusive to t_1 are relevant to Federal Emergency Management Agency (FEMA) spreading information on Twitter to provide help. Tweets arriving at t_2 reveals the gas shortage in Texas after Hurricane Harvey.
- By comparing the topics extracted before t_2 and those emerged at t_3, we can observe that gas shortage and donation are the popular topics over these two periods. Meanwhile, at t_3, new topics about Katy ISD schools and Red Cross started emerging. According to the investigation of the original tweets in the data and information from the Internet, we found that the Katy ISD schools suffered flood damage since Harvey's heavy rains began pounding in Katy. Another unique topic at t_3 is Texas officials and residents discussed that the Red Cross floundered and failed to provide help.
- One common topic over t_3 and t_4 is that the storm had started threatening children's safety. The exclusive top-

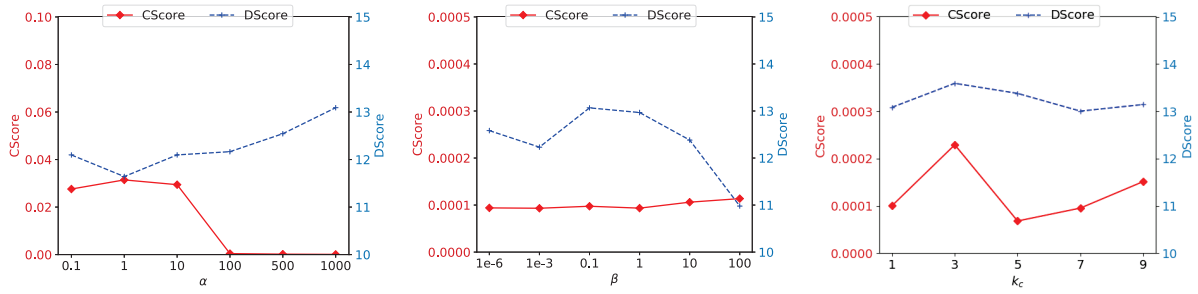


Figure 5: Parameter studies of α, β, and k_c. The red solid line represents CScore (a smaller value is desired) and blue dashed line denotes the DScore (a larger value is desired)

Table 2: Visualization of the common and distinct topics during the first five time periods of Hurricane Harvey. CT_{ij} denotes topics that are common before t_i and at t_j, and $DT_{i(j)}$ denotes topics generated before t_i that are distinct from topics at t_j

$t_1 - t_2$			$t_2 - t_3$			$t_3 - t_4$			$t_4 - t_5$		
CT_{12}	$DT_{1(2)}$	$DT_{2(1)}$	CT_{23}	$DT_{2(3)}$	$DT_{3(2)}$	CT_{34}	$DT_{3(4)}$	$DT_{4(3)}$	CT_{45}	$DT_{4(5)}$	$DT_{5(4)}$
texas	fema	gas	gas	fema	school	safety	khou11	neighbor	shelter	jake	walmart
evacuate	aid	shortage	fuel	figure	reopen	young	build	dallas	need	paul	donate
recover	fight	rare	singer	fight	katystrong	child	destroy	louisiana	info	peace	million
pray	file	fuel	donate	female	katyisd	tragedy	joel	forecast	rescue	rescue	neighbor
family	claim	crazy	party	fellow	flounder	devastate	osteen	east	million	negative	black
flood	texan	station	love	feed	cross	help	meme	hurricane	animal	image	history
help	help	crisis	heart	feel	red	rescue	funny	leeward	pet	gas	devastate
quick	navy	announce	shelter	fee	participate	family	response	island	beaumont	shipment	flood

ics at t_3 include Harvey destroyed the building of television station KHOU 11 and an anecdote about the Houston megachurch leader Joel Osteen, who refused to open the church for victims. Arising new topics at t_4 are relevant to the disaster situations of neighboring cities such as Dallas.

- Animal rescue in Beaumont started receiving attentions over t_4 and t_5. The exclusive topics at t_4 include the actor Jake Paul helping rescue victims and the shut-off of shipments of unbranded petroleum due to gas shortage. At t_5, several new topics emerged such as Walmart donated to Hurricane Harvey Relief and Houston's historically black neighborhoods devastated by flooding.

In general, we observe that common topics are often identified as being of interest to the public whereas distinct topics are often new alerting topics that are exclusive to a specific organization/individual during a certain time period.

Parameter Analysis

Here, we study how the variation of α, β, k_c affects CScore (smaller the better) and DScore (larger the better) using the *Harvey* dataset. In this experiment, we set α and β among $\{0.1, 1, 10, 100, 500, 1000\}$ and $\{1e-6, 1e-3, 0.1, 1, 10, 100\}$, respectively. k_c is selected from $\{1, 3, 5, 7, 9\}$ (the total number of topics is set to be 10). We vary one parameter at a time and fix the rest. For each set of parameters, we average the corresponding results along the time and present the mean of CScore and DScore in Fig. 5. We observe that larger α results in both better CScore and DScore. As α increases, TDF enforces the similarities between more common topics, making the rest topics more distinct from each other. For β, as it becomes extremely large, it shows significantly negative influence on the DScore. We conjecture that overemphasizing the sparsity of the inner products of two matrices may not enforce the differences between these matrices as desired. The best performance is achieved when α lies between [500,1000] and β is between [0.1,1]. We also observe from Fig. 5 that DScore is more robust to k_c than CScore. All the quantitative results are computed with $k_c = 7, \alpha = 1000, \beta = 0.1$.

Conclusions

In this paper, we study a novel topic tracking problem that seeks to discover common and distinct topics simultaneously using social streaming data generated during disasters. This is an important research problem because it provides an effective and efficient way for disaster responders to collect accurate information via social media. Due to the large proportion of undesired social media data, methods like set intersection/difference cannot well serve the purposes. To this end, we propose a TDF framework that leverages online NMF that conducts fast update of latent factors, and a joint NMF that seeks the balance between topic identifications and discoveries of common and distinct topics. Experimental results corroborate the effectiveness and the efficiency of the proposed framework. For future work, we will study methods that automatically compute k_c based on different inputs and explore various distance measures for gauging common and distinct topics.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) Grants #1610282 and #1909555.

References

Cao, B.; Shen, D.; Sun, J.-T.; Wang, X.; Yang, Q.; and Chen, Z. 2007. Detect and track latent factors with online nonnegative matrix factorization. In *International Joint Conference on Artificial Intelligence*, volume 7, 2689–2694. IJCAI.

Chen, X., and Candan, K. S. 2014. Gi-nmf: Group incremental non-negative matrix factorization on data streams. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 1119–1128. ACM.

Chen, Y.; Zhang, H.; Liu, R.; Ye, Z.; and Lin, J. 2019. Experimental explorations on short text topic mining between lda and nmf based schemes. *Knowledge-Based Systems* 163:1–13.

Chen, X.; Candan, K. S.; and Sapino, M. L. 2018. Ims-dtm: Incremental multi-scale dynamic topic models. In *Proceedings of the 2018 AAAI Conference on Artificial Intelligence*. AAAI.

Ding, W., and Chen, C. 2014. Dynamic topic detection and tracking: A comparison of hdp, c-word, and cocitation methods. *Journal of the Association for Information Science and Technology* 65(10):2084–2097.

Ding, C.; Li, T.; Peng, W.; and Park, H. 2006. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 126–135. ACM.

Gao, H.; Barbier, G.; and Goolsby, R. 2011. Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems* 26(3):10–14.

Hoffman, M.; Bach, F. R.; and Blei, D. M. 2010. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing systems*, 856–864. NIPS.

Houston, J. B.; Hawthorne, J.; Perreault, M. F.; Park, E. H.; Hode, M. G.; Halliwell, M. R.; Mcgowen, S. T.; Davis, R.; Vaid, S.; Mcelderry, J. A.; et al. 2015. Social media and disasters: a functional framework for social media use in disaster planning, response, and research. *Disasters* 39(1):1–22.

Kalyanam, J.; Mantrach, A.; Saez-Trumper, D.; Vahabi, H.; and Lanckriet, G. 2015. Leveraging social context for modeling topic evolution. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 517–526. ACM.

Kim, H.; Choo, J.; Kim, J.; Reddy, C. K.; and Park, H. 2015. Simultaneous discovery of common and discriminative topics via joint nonnegative matrix factorization. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 567–576. ACM.

Kumar, S.; Barbier, G.; Abbasi, M. A.; and Liu, H. 2011. Tweettracker: An analysis tool for humanitarian and disaster relief. In *Proceedings of the 2011 AAAI Conference on Web and Social Media*. AAAI.

Lee, D. D., and Seung, H. S. 2001. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing systems*, 556–562. NIPS.

Lehmann, J.; Gonçalves, B.; Ramasco, J. J.; and Cattuto, C. 2012. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st International Conference on World Wide Web*, 251–260. ACM.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 497–506. ACM.

Mathioudakis, M., and Koudas, N. 2010. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, 1155–1158. ACM.

Meier, P. 2015. *Digital humanitarians: how big data is changing the face of humanitarian response*. Routledge.

Nazer, T. H.; Xue, G.; Ji, Y.; and Liu, H. 2017. Intelligent disaster response via social media analysis a survey. *ACM SIGKDD Explorations Newsletter* 19(1):46–59.

Niculae, V.; Suen, C.; Zhang, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2015. Quotus: The structure of political media coverage as revealed by quoting patterns. In *Proceedings of the 24th International Conference on World Wide Web*, 798–808. International World Wide Web Conferences Steering Committee.

Saha, A., and Sindhwani, V. 2012. Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, 693–702. ACM.

Shi, T.; Kang, K.; Choo, J.; and Reddy, C. K. 2018. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 World Wide Web Conference*, 1105–1114. International World Wide Web Conferences Steering Committee.

Tu, D.; Chen, L.; Lv, M.; Shi, H.; and Chen, G. 2018. Hierarchical online nmf for detecting and tracking topic hierarchies in a text stream. *Pattern Recognition* 76:203–214.

Vaca, C. K.; Mantrach, A.; Jaimes, A.; and Saerens, M. 2014. A time-based collective factorization for topic discovery and monitoring in news. In *Proceedings of the 23rd International Conference on World Wide Web*, 527–538. ACM.

Wang, Y.; Agichtein, E.; and Benzi, M. 2012. Tm-lda: efficient online modeling of latent topic transitions in social media. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 123–131. ACM.

Wang, F.; Tan, C.; Li, P.; and König, A. C. 2011. Efficient document clustering via online nonnegative matrix factorizations. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, 908–919. SIAM.

Zhao, R.; Tan, V. Y.; and Xu, H. 2016. Online nonnegative matrix factorization with general divergences. *arXiv preprint arXiv:1608.00075*.