Trendi: Tracking Stories in News and Microblogs via Emerging, Evolving and Fading Topics

Xuchao Zhang[1], Liang Zhao[2], Zhiqian Chen[1], Arnold P. Boedihardjo[3], Jing Dai[4], Chang-Tien Lu[1]
[1]Virginia Tech, Falls Church, VA, USA
[2]George Mason University, Fairfax, VA, USA
[3]U. S. Army Corps of Engineers, Alexandria, VA, USA
[4]Google Corporation, USA
[1]{xuczhang, czq, ctlu}@vt.edu, [2]zhao9@gmu.edu, [3]arnold.p.boedihardjo@usace.army.mil, [4]jddai@google.com

***Abstract*—In today's era of information overload, people are struggling to detect the evolution of hot topics from massive news media and microblogs such as Twitter. Reports from mainstream news agencies and discussions from microblogs could complement each other to form a complete picture of major events. Existing work has generally focused on a single source, seldom attempting to combine multiple sources to track the evolution of topics: emerging, evolving and fading phrases as this would require a considerably more sophisticated model. This paper proposes a novel story discovery model that integrates evolutionary topics in news and Twitter data sources using an incremental algorithm by 1) discovering complementary information from news and microblogs that provides a more complete view of major events; 2) modeling emerging, evolving and fading topics and features throughout ongoing events; and 3) creating a scalable algorithm that is capable of handling massive data from news and social media. The parameters of the new model are optimized using a novel algorithm based on the alternative direction method of multipliers (ADMM). Extensive experimental evaluations on multiple datasets from different domains demonstrate the effectiveness and efficiency of our proposed approach.**

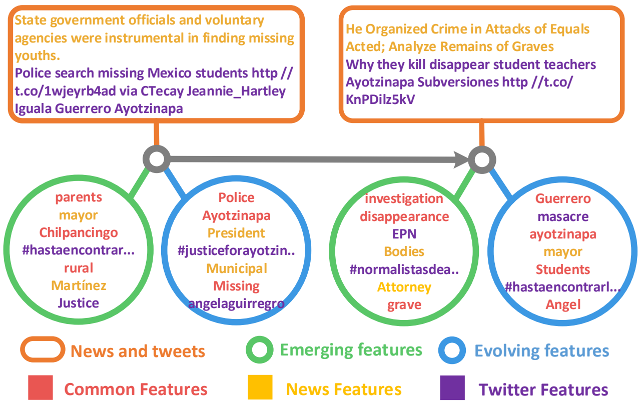

State government officials and voluntary agencies were instrumental in finding missing youths.
Police search missing Mexico students http // t.co/1wjeyrb4ad via CTecay Jeannie_Hartley Iguala Guerrero Ayotzinapa
He Organized Crime in Attacks of Equals Acted; Analyze Remains of Graves
Why they kill disappear student teachers Ayotzinapa Subversiones http //t.co/ KnPDilz5kV
parents
mayor
Chilpancingo
#hastaencontrar...
rural
Martínez
Justice
Police
Ayotzinapa
President
#justiceforayotzin..
Municipal
Missing
angelaguirregro
investigation
disappearance
EPN
Bodies
#normalistasdea..
Attorney
grave
Guerrero
masacre
ayotzinapa
mayor
Students
#hastaencontrarl...
Angel
News and tweets
Emerging features
Evolving features
Common Features
News Features
Twitter Features

Figure 1. Story evolution in news media and Twitter

I. Introduction

When important events occur, mainstream news media deliver timely reports, covering main aspects of the events using fairly standard language. Compared to traditional media, microblogs such as Twitter are rapidly becoming popular alternative news sources for spreading information mixed with personal comments and opinions. The distinctly different natures of these two types of media thus provide a complementary view [1] of an ongoing event: an objective and comprehensive presentation in news media; and a commentary full of opinions and sentiments from the public. For instance, numerous reports on ObamaCare reform from news media provide information and details of the health care initiative, while social media such as Twitter presents a more personal viewpoint, namely the publicly expressed opinions of individual people. Moreover, topics of interest change swiftly in today's fast-moving society, which makes it important to track its evolution and minimize stale information. Thus, the ability to identify and highlight the emerging, evolving and fading topics could offer an important way to save time for users seeking to track dynamic topics from extensive data. For example, the Figure 1 shows an episode in the Mexico's student kidnap event. Highlighting the emerging terms such as "rural" and "bodies" helps users easily to understand the fact that bodies of students were found after searching in rural area. Meanwhile, compared to the objective expression in news media, Twitter contains some public voices such as "#justiceforayotzinapa" and "masacre".

Although a range of ways to generate event stories have been extensively explored [2][3][4], most have been applied to only a single data source, generally either news articles or Twitter. Furthermore, few of the existing methods simultaneously consider the emerging, evolving and fading progression in both topic and feature levels. To address these issues, the major challenges can be summarized as follows: *1) Complementary information discovery between news articles and Twitter.* A naive solution is to detect topics in news and Twitter individually, then find their similarities and differences. However, this solution becomes problematic when mapping the same topics across different sources if they are generated separately. Hence, a robust method is necessary to detect differences in the way topics are treated by the two data sources. *2) Emerging, evolving and fading topic detection for joint data sources.* Although dynamic topic models are well studied recently, it remains a question that how to track a topic emerging from social media but evolving in news reports. Therefore, an integrated model is required to track the dynamic topics and process two data

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

sources simultaneously. *3) Scalable algorithm for massive data.* Thousands of news items and millions of tweets are generated for reporting events every day. Thus, a scalable algorithm is required to tractably handle and process the massive data.

In order to overcome all the above-mentioned challenges, we propose a novel model that simultaneously considers the dynamic topics and joint data sources provided by news outlets and Twitter. The proposed algorithm is designed to run incrementally, making the algorithm scalable for large datasets. The major contributions of this research can be summarized as follows:

- **A framework for story generation in news media and microblogs.** A novel unsupervised approach is proposed to generate stories of topic evolution in both news and Twitter data sources. Our method extracts emerging, evolving and fading topics at each time step, and connects related non-fading topics in temporal order by a story line.
- **A novel story generation model to track evolutionary topics in news and Twitter.** In the proposed story generation model, dynamic topics are jointly considered across both news and Twitter data sources by three stages of evolution, namely emerging, evolving, and fading topics. All are characterized by different regularization and constraint models.
- **An efficient algorithm for the new story generation problem.** Our proposed model is a non-smooth convex optimization problem with affine equality and non-negative inequality constraints, which is challenging to solve. By introducing auxiliary variables, we have developed an efficient ADMM-based algorithm to solve the problem with rapid convergence.
- **Extensive experimental performance evaluations.** Our proposed method has been extensively evaluated on both Twitter and news report data covering multiple countries in Latin America. Comparisons with baselines and state-of-the-art methods clearly demonstrate its efficiency and effectiveness.

The remainder of this paper is organized as follows. Section II describes the related work on story construction, dynamic topic modeling, and multi-source topic detection. The problem definition is presented in Section III and our incremental story generation model for news and Twitter is described in Section IV. Section V presents the optimization algorithm for our proposed model. In Section VI, the experimental results are analyzed and a case study presented. We conclude with a summary of our work in Section VII.

II. Related Work

Several research directions provide the background for this study, namely storyline construction, dynamic topic modeling and topic detection in news media and microblogs. These will be considered in turn in this section.

Timeline and Storyline Construction. Several studies have explored document summarization with time stamps, most of which focus on news articles. Mei and Zhai [5] proposed an Hidden Markov Model (HMM) style probabilistic method to discover and summarize the evolutionary patterns of themes in text streams, while Lappas et al. [6] defined a term burstiness model to discover the temporal trend of terms in news article streams. Wang et al. [7] took this further, developing an evolutionary document summarization system to generate an evolution skeleton along the timeline. Only a limited number of studies have focused specifically on event summarization using Twitter data. Takamura et al. [8] took the posted time of microblogs into consideration, proposing a summarization model based on the p-median problem for a stream of microblog posts along a timeline. Later, Lidan et al. [2] proposed a method based on an online tweet stream clustering algorithm and TCV-Rank summarization for tweet streams. However, these studies only considered single data sources in news or Twitter, and did not explicitly include an examination of the fading features in their models.

Dynamic Topic Modeling. Dynamic topic models (DTM) consider time information related to the evolution of topics moving beyond static treatments. These can be divided into two categories: dynamic probabilistic and dynamic matrix factorization approaches. Probabilistic DTMs embed independent Latent Dirichlet Allocation (LDA) in the timeline to connect the evolving topics. Blei and Lafferty [9] proposed the first dynamic topic model to detect the evolving relationship between topics. Tomoharu et al. [10] proposed an online topic model which sequentially analyzes the time evolution of topics in document collections. Such LDA-based dynamic topic models always choose a fixed number of topics within a specified time span, which means that emerging and fading topics are not explicitly considered.

Matrix factorization based DTMs are mainly characterized using Non-negative Matrix Factorization (NMF) frameworks [11][12]. Cao et al. [13] proposed a new topic detection framework that extends the NMF by applying orthogonal constraints, as well as injecting a small number of new topics at each time step to model emerging and evolving topics. Vaca et al. [14] discovered trends in topics by introducing a mapping matrix between adjacent time steps. Chen et al. [15] modeled emerging, evolving and fading topics using dynamic soft orthogonal NMF. However, all these studies either ignored fading topics, assumed that all the topics would be related to previous time steps without eliminating fading topics, or were unable to detect the fading features in existing topics. Moreover, neither existing Probabilistic nor Matrix factorization based DTMs could be applied directly to joint data sources.

Joint Study in News media and Microblog. Joint studies of news media and microblogs have attracted much more attention recently due to their high interaction in potential applications. Zhao et al. [16] conducted a comparison of topic categories and types on Twitter and news media by running separate topic models. Gao et al. [17] proposed a joint topic modeling for event summarization across news and social media. Hua et al. [18] recently proposed a hierarchical Bayesian model that jointly models news and social media topics and their interactions. However, these studies only consider the topic modeling within a static time span. For discovering and linking the topics into a story, Wang et al. [1] proposed a hierarchical event discovery model to learn news events that was linked to their reflections in Twitter. However, their model only integrated the Twitter data into the story at a very late stage, and ignored evolving topics in the Twitter dataset. To the best of our knowledge, the present study is the first work addressing the task of event story generation in news media and microblogs that considers emerging, evolving and fading topics.

III. Problem Definition

In this section, we formally define the problem of tracking dynamic topics in news and Twitter corpora, along with some key definitions and the main concepts involved. The notations used in this paper are summarized in Table I.

Given Sets of documents and tweets that arrive continuously in batches. Each batch is represented by a news data matrix $X_d^{(i)} \in \mathbb{R}^{f \times n_d^{(i)}}$ and a Twitter data matrix $X_t^{(i)} \in \mathbb{R}^{f \times n_t^{(i)}}$, where $n_d^{(i)}$ and $n_t^{(i)}$ are the number of documents and tweets produced at time step i and f is the number of features in the coding scheme. For instance, these could be the words used in news articles and *#hashtags* in tweets. For simplicity, we assume that the feature number is known and fixed in advance for all the batches. Fortunately, this is a realistic assumption when the whole corpus is preprocessed, and it is easy to extend existing data matrices to new features.

Topics at time step i can be defined over features in the news and Twitter feature set, and represented as an f-dimensional vector u, where f is the number of features, and the k-th entry denotes the weight of the k-th feature in the topic. Suppose that there are $c^{(i)}$ topics at time i, then these topics can be summarized into a $f \times c^{(i)}$ feature-topic matrix $U = [u_1, u_2, \cdots, u_{c^{(i)}}]$. As both news and Twitter data sources are considered in this work, the features in a topic can also be considered as complement combination between news and Twitter. The complement features are formally defined as:

Table I
Math Notations

Notations	Explanations
$n_d^{(i)}, n_t^{(i)} \in \mathbb{R}$	document and tweet number at time i
$f^{(i)} = f \in \mathbb{R}$	feature number at time i
$c^{(i)} \in \mathbb{R}$	topic number at time i
$s^{(i)} \in \mathbb{R}$	existing story number at time i
$X_d^{(i)} \in \mathbb{R}^{f \times n_d^{(i)}}$	feature-document *tf-idf* matrix at time i
$X_t^{(i)} \in \mathbb{R}^{f \times n_t^{(i)}}$	feature-tweet *tf-idf* matrix at time i
$P_d^{(i)} \in \mathbb{R}^{n_d^{(i)} \times c^{(i)}}$	document partition matrix at time i
$P_t^{(i)} \in \mathbb{R}^{n_t^{(i)} \times c^{(i)}}$	tweet partition matrix at time i
$U^{(i)} \in \mathbb{R}^{f \times c^{(i)}}$	feature-topic matrix at time i
$U_v^{(i)} \in \mathbb{R}^{f \times c^{(i)}}$	evolving feature-topic matrix at time i
$U_m^{(i)} \in \mathbb{R}^{f \times c^{(i)}}$	emerging feature-topic matrix of at time i
$U_{\neg f}^{(i-1)} \in \mathbb{R}^{f \times c^{(i-1)}}$	non-fading feature-topic matrix at time i-1
$U_c^{(i)} \in \mathbb{R}^{f \times c^{(i)}}$	news and Twitter shared feature-topic matrix at time i
$U_d^{(i)} \in \mathbb{R}^{f \times c^{(i)}}$	news feature-topic matrix at time i
$U_t^{(i)} \in \mathbb{R}^{f \times c^{(i)}}$	Twitter feature-topic matrix at time i

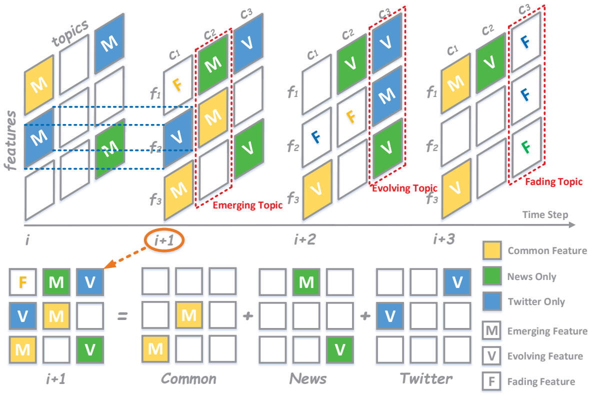


Figure 2. Story evolution in news media and Twitter

Definition 1. ***Complement features between News and Twitter****: Feature set f in topic $c_k^{(i)}$ can be divided into three categories: 1) the shared feature set f_c is defined as the set of features contained in both news and Twitter; 2) the news feature set f_d is defined as the features that only belong to news; and 3) the Twitter feature set f_t is defined as the features contained by Twitter alone. Therefore, the feature sets for news and Twitter can be represented as $f_c \cup f_d$ and $f_c \cup f_t$, respectively.*

To track the topic evolution in both the feature and topic levels, we define the dynamic topic and feature as follows:

Definition 2. ***Dynamic Topics****: The dynamics of underlying topics are described as follows: (1) Emerging Topic: Given a topic $c_k^{(i)}$, if the topic does not exist in the topics $C^{(i-1)}$ at previous time, we say $c_k^{(i)}$ is an emerging topic at time i. (2) Evolving Topic: topic $c_k^{(i)}$ is an evolving topic if it exists in topic set $C^{(i-1)}$ in previous time frame. (3) Fading topic: topic $c_k^{(i)}$ is a fading topic when it does not exist or barely exists in topic set $C^{(i+1)}$ in the next time frame.*

As shown in Figure 2, topic c_2 is an emerging topic at time i+1 because the topic does not exist at time i; topic c_3 at time i+2 evolves from topic c_3 at time i+1; and topic c_3 at time i+3 is a fading topic as it disappears at that time.

Definition 3. ***Dynamic Features in Topics****: To detect the dynamics of underlying topics in the feature level, dynamic features are modeled as follows: (1) Emerging Feature: Given a feature $f_j^{(i)}$ in its topic $c_k^{(i)}$, we define $f_j^{(i)}$ as an emerging feature if it does not exist in the previous topic $c_k^{(i-1)}$. (2) Evolving Feature: feature $f_j^{(i)}$ in topic $c_k^{(i)}$ is an evolving feature if the feature exists in its previous topic $c_k^{(i-1)}$. Because evolving features are included in the evolving topics, there is no need to identify them explicitly. (3) Fading Features: feature $f_j^{(i)}$ in topic $c_k^{(i)}$ is a fading feature if it does not exist in the next later topic $c_k^{(i+1)}$. As* *shown in Figure 2, feature f_2 in topic c_3 is an emerging feature at time i+2 because the feature does not exist at time i+1; feature f_1 in topic c_1 is a fading feature at time i+1 because it exists at time i but has faded out by time i+1.*

If the topics are detected from the perspective of joint sources and their dynamics, we can generate these topics into a story in which the relevant topics are connected following their temporal order. A story in our problem is formally defined as follows:

Definition 4. ***Story****: A topic based story is defined as a sequence of topics $s_k = (c_k^{(1)}, c_k^{(2)}, \dots)$, where each topic $c_k^{(i)}$ consists of evolving topics from its previous time topic $c_k^{(i-1)}$. If topic $c_k^{(m)}$ at time m is fading, then its corresponding story s_k also ends. The length of story s_k is m. If a new topic $c_{k'}^{(n)}$ emerges at time n, a new story $s_{k'}$ containing topic $c_{k'}^{(n)}$ is created at time n.*

Applying the above definitions, the problem addressed in this paper can be formulated as follows:

Problem Formulation: Tracking stories in news and Twitter via dynamic topics. Given batches of news data X_d and Twitter data X_t, the goal is to discover and track stories with dynamic topics contained in joint data sources via the following three tasks: 1) divide the feature set f in each topic into three feature sets as shared feature f_c, news feature f_d, and Twitter feature f_t; 2) categorize each dynamic topic $c_k^{(i)}$ at time i into a tuple $c_k^{(i)} = (c_m^{(i)}, c_v^{(i)}, c_f^{(i)})$, in which $c_m^{(i)}, c_v^{(i)}$, and $c_f^{(i)}$ are defined as emerging, evolving and fading topics respectively; and 3) track the stories $S = \{s_j | s_j = (c_j^{(1)}, c_j^{(2)}, \dots, c_j^{(n)})\}$ with continuous evolving topics $C = \{c_j^{(1)}, \dots, c_j^{(n)}\}$, where n is the length of the story.

IV. Model

In this section, we propose a new model to track stories in both news and Twitter data sources via dynamic topics. The feature modeling is first introduced, after which the dynamic topics modeling and story tracking will be presented.

A. Modeling Complement Features in News and Twitter

To detect the complement features in news and Twitter, the feature-topic matrix $U^{(i)}$ is represented as $U^{(i)} = U_c^{(i)} + U_d^{(i)} + U_t^{(i)}$, where $U_c^{(i)}$, $U_d^{(i)}$, and $U_t^{(i)}$ represents the shared features, news and Twitter exclusive features, respectively. The feature-topic matrices of news and Twitter can thus be represented as $U_c^{(i)} + U_d^{(i)}$ and $U_c^{(i)} + U_t^{(i)}$. As $U_d^{(i)}$ and $U_t^{(i)}$ represents the exclusive features, $U_d^{(i)}$ should be distinct from $U_t^{(i)}$. We define the distinction term $\mathbb{D}$ to make them contrast by the following theorem:

Theorem 1. *Minimizing $\mathbb{D} = \langle U_d^{(i)}, U_t^{(i)} \rangle$ makes $U_d^{(i)}$ and $U_t^{(i)}$ distinct, where $\langle \cdot \rangle$ is the sum of elements in the Hadamard product [19].*

Proof: Let $P = U_d^{(i)}$ and $Q = U_t^{(i)}$. The distinction $\mathbb{D}$ between $U_d^{(i)}$ and $U_t^{(i)}$ is then defined as:

$$\begin{aligned} \mathbb{D}_{PQ} &= \frac{1}{2}(\|P+Q\|_F^2 - \|P\|_F^2 - \|Q\|_F^2) \\ &= \frac{1}{2}(\textstyle\sum_i \sum_j (P_{ij} + Q_{ij})^2 - \sum_i \sum_j (P_{ij})^2 \\ &\quad - \displaystyle\sum_i \sum_j (Q_{ij})^2) = \textstyle\sum_i \sum_j (P_{ij} Q_{ij}) \equiv \langle P, Q \rangle \end{aligned} \tag{1}$$

■

For instance, if one feature are both selected by P and Q as weight 0.8 and 0.9. Then the penalty of $\mathbb{D}$ is 0.72. Notice that P and Q are both non-zero matrices unless news or Twitter corpus doesn't contain any features.

B. Modeling Dynamic Topics

To differentiate between the emerging, evolving, and fading parts of dynamic topics, we represent the feature-topic matrix $U^{(i)}$ as $U^{(i)} = U_m^{(i)} + U_v^{(i)}$, where $U_m^{(i)}$ and $U_v^{(i)}$ are the emerging and evolving parts, respectively, of a feature-topic matrix. Different from previous work [15], we propose a novel method to represent the dynamic topics with the summation matrix $U^{(i)} = U_m^{(i)} + U_v^{(i)}$ and distinction constraint $\langle U_d^{(i)}, U_t^{(i)} \rangle$ instead of the matrix concatenation $[U_v^{(i)}, U_m^{(i)}]$. This has been done because the matrix concatenation restricts the emerging features to only those present in the emergent topics of the feature-topic matrix although emergent features may also appear in evolving topics. The matrix summation breaks this limitation and makes it possible to detect the emerging features in evolving topics. Moreover, the new method uses a unified matrix form, which is available to be applied in our efficient ADMM-based algorithm.

Modeling emerging topics. At time step i, topics that had not existed at the previous time i-1 are defined as emerging topics. To model the emerging topics, an estimate for the number of new topics is dynamically added into the feature-topic matrix $U^{(i)}$. For example, if k topics are assumed to be added at time i, k new columns will be concatenated into the right hand of $U^{(i)}$.

Modeling evolving topics. Based on the assumption that the evolving topics will change somewhat at a small scale for the same topic between consecutive time periods, the non-fading topics at time i-1 will be treated as evolving topics at time i. Therefore, we use the distance $\eta = \|U_v^{(i)} - U_{\neg f}^{(i-1)}\|_F^2$ to represent the evolving phrase. Our purpose is to minimize η to track the evolving topics. Notice that, the same topic is consistently represented as the same column of $U_v^{(i)}$ and $U_{\neg f}^{(i-1)}$ at different times.

Modeling fading topics. Given a suitable time interval between time steps, our fading model is based on the assumption that topics and features will fade smoothly. An individual topic or feature at time step i will be identified as a fading topic or feature if it satisfies the following condition. 1) *Fading topic* (few documents contain the topic): If the corresponding column of topic c is very sparse in $P_d^{(i)}$ and

$P_t^{(i)}$, then the topic can be removed in the next time frame when it satisfies: $\frac{\sum_{j=1}^{n_d^{(i)}} I\{[P_d]_{jc}=0\}}{n_d^{(i)}} + \frac{\sum_{j=1}^{n_t^{(i)}} I\{[P_t]_{jc}=0\}}{n_t^{(i)}} > 2\tau$, where τ is the threshold of document sparsity and $I\{\cdot\}$ is an indicator function, whose value equals 1 when the condition inside is satisfied but is otherwise equal to 0. *Fading feature (feature seldom mentioned in a topic):* If the normalized weight for feature f in topic c is very low, the feature is considered as having faded in that topic. More specifically, it is presented as $[\Phi \cdot U]_{fc} < \sigma$, where σ is the weight threshold and Φ is the column-based normalization matrix for topics in matrix U.

C. Story Tracking

As the story is defined as a sequence of topics, the construction of a story is based on the three phases of dynamic topics: emerging, evolving, and fading. *Story emerging:* As a possible number of topics are dynamically added into feature-topic matrix $U^{(i)}$, some of these will be identified as emergent. For these emerging topics, a new story will be created accordingly. *Story evolving:* If a topic $c_k^{(i)}$ is an evolving topic at time i, then the topic will be added into the story containing its evolved topic $c_k^{(i-1)}$ at time i-1. *Story fading:* If a topic $c_k^{(i)}$ is identified as a fading topic at time i, then the story containing the fading topic will be ended.

D. Objective Function

Our news and Twitter based story evolution tracking method is formulated as minimizing the following objective function:

$$\begin{aligned} J_{nt} = {} & \lambda_1 \|X_d^{(i)} - (U_c^{(i)} + U_d^{(i)})[P_d^{(i)}]^T\|_F^2 \\ & + \lambda_2 \|X_t^{(i)} - (U_c^{(i)} + U_t^{(i)})[P_t^{(i)}]^T\|_F^2 + \theta_1 \|U_v^{(i)} - U_{\neg f}^{(i-1)}\|_F^2 \\ & + \varphi_1 \langle U_v^{(i)}, U_m^{(i)} \rangle + \varphi_2 \langle U_d^{(i)}, U_t^{(i)} \rangle + \delta_1 \|P_d^{(i)}\|_1 + \delta_2 \|P_t^{(i)}\|_1 \\ s.t. \quad & \begin{cases} U^{(i)} = U_v^{(i)} + U_m^{(i)} = U_c^{(i)} + U_d^{(i)} + U_t^{(i)} \\ P_d^{(i)}, P_t^{(i)}, U_v^{(i)}, U_m^{(i)}, U_c^{(i)}, U_d^{(i)}, U_t^{(i)} \geq 0 \end{cases} \end{aligned} \tag{2}$$

where $\Gamma = \{\lambda_1, \lambda_2, \theta_1, \varphi_1, \varphi_2, \delta_1, \delta_2\}$ are the weights for each term. The objective function in Equation (2) consists of four major parts with nine terms. The 1st and 2nd terms represent the news and Twitter partition in topics, the 3rd term considers the evolving topics in comparison to those at previous times, and the 4th and 5th terms aim to distinguish features from different perspectives. Specifically, the 4th term ensures evolving features are distinct from emerging features, and the 5th term aims to differentiate the features contained in news and Twitter. The remaining terms ensure the sparsity of the corresponding variables. The importance of each term can be adjusted by weight set Γ. In general, each weight can be set to any real number from zero to one, in which zero means the term is ignored and one represents the most important term.

V. Algorithm Derivations

In this section, an ADMM (Alternating Direction Method of Multipliers) based framework is proposed to solve the objective function presented in Section IV.

Algorithm 1: Trendi Algorithm

```
   Input: $X_d \in \mathbb{R}^{f \times n_d^{(i)}}$, $X_t \in \mathbb{R}^{f \times n_t^{(i)}}$ , $\Gamma$
   Output: solution $\Theta$
 1 Initialize $\rho$ = 1, $\Theta$, $\Psi$
 2 Choose $\varepsilon_r > 0$, $\varepsilon_s > 0$
 3 repeat
 4     Update $\Theta$ and $\Psi$ by Equations (1) $\sim$ (11).
 5     Update $\{\alpha_i\}_{i=1}^{11}$ by Equation (12).
 6     Update primal and dual residuals $r$ and $s$ by Theorem 2.
 7     if $r > 10s$ then
 8         $\rho \leftarrow 2\rho$
 9     else if $10r < s$ then
10         $\rho \leftarrow \rho/2$
11     else
12         $\rho \leftarrow \rho$
13 until $r < \varepsilon_r$ and $s < \varepsilon_s$
```

To solve the non-convex optimization problem with constraints in Equation (2), the alternating direction method of multipliers (ADMM) is widely utilized as an efficient algorithm that breaks the original large problem into smaller subproblems, which can then be solved efficiently. For notational simplicity, the time stamp subscript (i) will be omitted in the algorithm derivation, and use $\hat{U}_{\neg f}$ to represent $U_{\neg f}^{(i-1)}$. Here an ADMM-based framework that solves Equation (2) by first obtaining its augmented Lagrangian format is as follows:

$$\begin{aligned} L_{nt} = {} & J_{nt} + \mathcal{L}(\alpha_1, U - U_v - U_m) + \mathcal{L}(\alpha_2, U - U_c - U_d - U_t) \\ & + \mathcal{L}(\alpha_3, P_d - P_{ds}) + \mathcal{L}(\alpha_4, P_d - P_{d+}) + \mathcal{L}(\alpha_5, P_t - P_{ts}) \\ & + \mathcal{L}(\alpha_6, P_t - P_{t+}) + \mathcal{L}(\alpha_7, U_v - U_{v+}) + \mathcal{L}(\alpha_8, U_m - U_{m+}) \\ & + \mathcal{L}(\alpha_9, U_c - U_{c+}) + \mathcal{L}(\alpha_{10}, U_d - U_{d+}) + \mathcal{L}(\alpha_{11}, U_t - U_{t+}) \end{aligned} \tag{3}$$

where function $\mathcal{L}$ is defined as $\mathcal{L}(x, y) \equiv \langle x, y \rangle + \frac{\rho}{2}\|y\|_F^2$, $\Theta = \{P_d, P_t, U, U_v, U_m, U_c, U_d, U_t\}$ are the parameters to be optimized, Ψ = $\{P_{ds}$, P_{d+}, P_{ts}, P_{t+}, U_{v+}, U_{m+}, U_{c+}, U_{d+}, $U_{t+}\}$ are the auxiliary matrix variables used to solve the constraints, $\{\alpha_i\}_{i=1}^{11}$ are the Lagrangian multipliers that are the dual variables of *ADMM*, and ρ is the step size of the dual step. The parameters Θ and Ψ are alternately solved by the proposed algorithm, dubbed *Trendi*, as shown in Algorithm 1[1]. This alternately optimizes each of the unknown parameters until convergence is achieved. Lines 4-5 show the alternating optimization of each of the unknown parameters. The calculation of the primal and dual residuals is given in Line 6. Lines 7-13 describe the updating of the penalty parameter ρ, which follows the updating strategy proposed by Boyd et al. [20]. The detailed optimization steps are described in more detail below.

A. Optimization of Variables

Holding the other parameters fixed, updating matrix P_d is equivalent to solving the following optimization problem:

[1] Equations in the algorithm can be found in supplementary document in https://goo.gl/k6uGhn

$$\begin{aligned}P_d \leftarrow \arg\min_{P_d} \lambda_1 \|X_d - (U_c + U_d)P_d^T\|_F^2 + \langle \alpha_3, P_d - P_{ds}\rangle \\ + \frac{\rho}{2}\|P_d - P_{ds}\|_F^2 + \langle \alpha_4, P_d - P_{d+}\rangle \\ + \frac{\rho}{2}\|P_d - P_{d+}\|_F^2\end{aligned} \tag{4}$$

and its analytic solutions of P_d is:

$$\begin{aligned}P_d = [2\lambda_1 X_d^T(U_c + U_d) + \rho P_{d+} + \rho P_{ds} - \alpha_3 - \alpha_4]\cdot \\ [2\lambda_1(U_c^T + U_d^T)(U_c + U_d) + 2\rho I]^{-1}\end{aligned} \tag{5}$$

As the space limitation, the detailed optimization steps for other variables are described in the supplementary document[2].

B. Calculate residuals

The primal and dual residuals of the $(k+1)$th iteration are calculated based on the following theorem, where the parameters labeled with superscript k (e.g., P_d^k) correspond to its value in the k^{th} iteration.

Theorem 2. *The primal residual and dual residual of the algorithm are as follows:*

- *Primal residual of objective function:*

$$\begin{aligned}r = \|U - U_v - U_m\|_F + \|U - U_c - U_d - U_t\|_F \\ + \|P_d - P_{ds}\|_F + \|P_d - P_{d+}\|_F + \|P_t - P_{ts}\|_F \\ + \|P_t - P_{t+}\|_F + \|U_v - U_{v+}\|_F + \|U_m - U_{m+}\|_F \\ + \|U_c - U_{c+}\|_F + \|U_d - U_{d+}\|_F + \|U_t - U_{t+}\|_F\end{aligned} \tag{6}$$

- *Dual residual of objective function:*

$$\begin{aligned}s = \rho(\|(P_{ds}^k - P_{ds}^{k+1}) + (P_{d+}^k - P_{d+}^{k+1})\|_F + \|(P_{ts}^k - P_{ts}^{k+1}) \\ + (P_{t+}^k - P_{t+}^{k+1})\|_F + \|(U_v^k - U_v^{k+1}) + (U_m^k - U_m^{k+1}) \\ + (U_c^k - U_c^{k+1}) + (U_d^k - U_d^{k+1}) + (U_t^k - U_t^{k+1})\|_F \\ + \|(U_m^{k+1} - U_m^k) + (U_{v+}^k - U_{v+}^{k+1})\|_F + \|(U_{m+}^k - U_{m+}^{k+1})\|_F \\ + \|(U_d^{k+1} - U_d^k) + (U_t^{k+1} - U_t^k) + (U_{c+}^k - U_{c+}^{k+1})\|_F \\ + \|(U_t^{k+1} - U_t^k) + (U_{d+}^k - U_{d+}^{k+1})\|_F + \|(U_{t+}^k - U_{t+}^{k+1})\|_F)\end{aligned} \tag{7}$$

The proof of theorem 2 can also be found in the supplementary document.

VI. Experiment

This section presents the empirical evaluations of the performance of our proposed new approach, *Trendi*. After the experiment setup has been introduced in Section VI-A, the effectiveness of the method is evaluated against several existing methods on performance of news and Twitter partition, along with an analysis of the dynamic topics. The efficiency analyses and case studies are presented in Section VI-F and VI-G.

A. Experiment Setup

Dataset and Labels: The news and Twitter data used in this paper were as follows:

Civil Unrest. The Civil Unrest dataset from Datasift Inc.[3] contains tweets gathered from civil unrest events in Latin America from June 2014 to October 2014. The dataset contains 87,269 tweets from Brazil, Venezuela, Colombia, Mexico, and Chile. All these tweets are labeled as being related to 7 different events including Mexico's Iguala Kidnap protest, Brazil's World Cup protests, and the Colombian presidential election protest. The labels were collected from a SVM [21] classifier trained using 15% pre-labeled data and verified by human observers. 2,563 tweets that contain links to local news media were collected and assigned to the corresponding news datasets. This dataset is a gold standard dataset that has been developed to test the performance of news and Twitter partitioning algorithms.

Mexico. The Mexico dataset contains the tweets related to a series of events that occurred in Mexico. The dataset was crawled using the location ”Mexico” in Twitter for the period from September to December, 2015. After preprocessing, the resulting dataset contained 1.1 million tweets, mainly in Spanish and English; 15,107 tweets containing a single link to CNN, BBC or local news media were collected during the same period.

Evaluation Metrics: To evaluate the clustering results, we adopted the standard performance measures frequently used for clustering: 1) *Clustering Accuracy*: Clustering Accuracy [22] discovers the one-to-one relationships between clusters and labeled classes and measures the accuracy of clusters that contain data points from the corresponding class. Given a data point x_i, let r_i and s_i be the obtained cluster label and the label provided by the corpus, respectively. The cluster accuracy is defined as $Acc = \frac{\sum_{i=1}^{n} \delta(s_i, map(r_i))}{n}$, where n is the total number of tweets, $\delta(x, y)$ is a delta function that equals one if $x = y$ and zero otherwise, and $map(r_i)$ is the permutation mapping function that maps each cluster label r_i to the equivalent label from the data corpus. To best map the result with ground truth, we use the Kuhn-Munkres algorithm [23]. 2) *Normalized Mutual Information*: Normalized Mutual Information(NMI) [24] is used to measure the quality of clusters. Let $\mathcal{L}$ denote the set of clusters obtained from the ground truth and $\mathcal{C}$ those obtained from our algorithm. The normalized mutual information metric is then defined as: *NMI* = $\frac{\sum_{i=1}^{c}\sum_{j=1}^{c} \frac{n_{i,j}}{n} \log \frac{n \cdot n_{i,j}}{n_i \hat{n}_j}}{\sqrt{(\sum_{i=1}^{c} n_i \log \frac{n_i}{n})(\sum_{i=1}^{c} \hat{n}_j \log \frac{\hat{n}_j}{n})}}$, where n_i denotes the number of tweets contained in the cluster $\mathcal{C}_i (1 \leq i \leq c)$, $\hat{n}_j$ is the number of tweets belonging to class $\mathcal{L}_i (1 \leq j \leq c)$, and $n_{i,j}$ is the number of tweets in the intersection between cluster $\mathcal{C}_i$ and ground truth class $\mathcal{L}_j$; the larger the NMI, the better the clustering result.

To validate the dynamic topic performance, an F-measure metric is adopted. The F-measure is defined as the harmonic mean of precision and recall: F-measure = 2 × Recall / (Precision + Recall). Precision designates the fraction of features extracted in our model that match the actual key words that emerged/evolved in the topics. Recall denotes the percentage of all the actual emerged/evolved words that

[2] https://goo.gl/gXuY2e

[3] http://datasift.com/

Table II
CLUSTERING RESULTS FOR THE CIVIL UNREST DATASET

	Time 1		**Time 2**		**Time 3**		**Time 4**	
	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI
KM	0.582,0.708,0.645	0.322,0.546,0.434	0.544,0.553,0.548	0.340,0.408,0.374	0.511,0.477,0.494	0.328,0.438,0.383	0.347,0.546,0.447	0.116,0.679,0.398
LSAKM	0.615,0.719,0.667	0.431,0.542,0.487	0.505,0.616,0.560	0.300,0.464,0.382	0.538,0.510,0.524	0.301,0.480,0.390	0.323,0.573,0.448	0.116,0.654,0.385
NMF	0.604,0.752,0.678	0.381,0.654,0.518	0.648,0.629,0.639	0.467,0.464,0.466	0.517,0.517,0.517	0.314,0.426,0.370	0.367,0.545,0.456	0.281,0.626,0.454
JPP	0.653,0.757,0.705	0.427,0.651,0.539	0.652,0.635,0.644	0.470,0.436,0.453	0.570,0.565,0.568	0.426,0.463,0.445	0.477,0.635,0.556	0.314,0.632,0.473
SONMFSRd	0.685,0.763,0.724	0.400,0.687,0.544	0.660,0.649,0.655	0.500,0.415,0.458	0.610,0.589,0.599	0.416,0.487,0.452	**0.490**,0.725,**0.607**	0.291,0.656,0.474
Trendi JD+DT	**0.695**,**0.836**,**0.765**	**0.451**,**0.723**,**0.587**	**0.698**,**0.750**,**0.724**	**0.504**,**0.486**,**0.495**	**0.630**,**0.649**,**0.639**	**0.436**,**0.550**,**0.493**	0.372,**0.794**,0.583	**0.343**,**0.733**,**0.538**
	Time 5		**Time 6**		**Time 7**		**Time 8**	
	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI
KM	0.642,0.749,0.696	0.407,0.544,0.476	0.318,0.845,0.582	0.382,0.638,0.510	0.633,0.599,0.616	0.450,0.560,0.505	0.580,0.446,0.513	0.392,0.507,0.450
LSAKM	0.610,0.730,0.670	0.342,0.521,0.431	0.376,0.868,0.622	0.144,0.665,0.405	0.620,0.612,0.616	0.423,**0.602**,0.512	0.560,0.456,0.508	0.375,**0.546**,0.461
NMF	0.614,0.763,0.689	0.431,0.517,0.474	0.457,0.825,0.641	0.206,0.657,0.431	0.639,0.619,0.629	0.432,0.508,0.470	0.585,0.465,0.525	0.364,0.454,0.409
JPP	0.587,0.768,0.678	0.429,0.573,0.501	0.513,0.827,0.670	0.329,0.677,0.503	0.644,0.625,0.635	0.449,0.598,0.524	0.596,0.504,0.550	0.370,0.517,0.444
SONMFSRd	0.630,0.770,0.700	0.416,0.554,0.485	**0.538**,0.841,0.690	**0.386**,0.687,0.536	0.642,**0.631**,0.637	0.458,0.579,0.518	0.614,0.513,0.564	0.400,0.521,0.461
Trendi JD+DT	**0.663**,**0.781**,**0.722**	**0.435**,**0.590**,**0.512**	0.482,**0.954**,**0.718**	0.371,**0.777**,**0.574**	**0.722**,0.621,**0.671**	**0.528**,0.542,**0.535**	**0.664**,**0.563**,**0.613**	**0.470**,0.513,**0.491**

Table III
CLUSTERING RESULTS FOR THE MEXICO DATASET

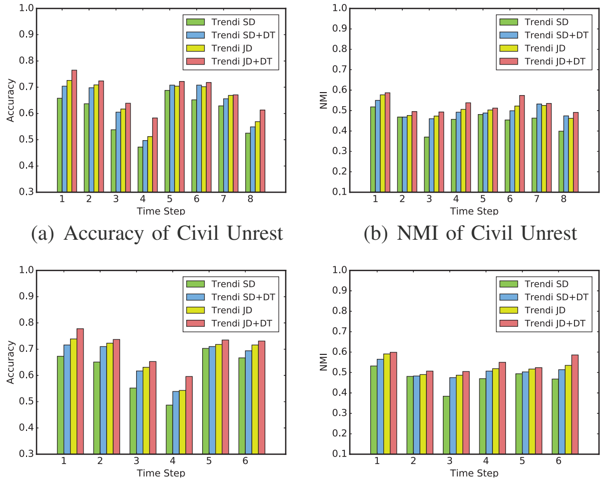
	Time 1		**Time 2**		**Time 3**	
	Acc	NMI	Acc	NMI	Acc	NMI
KM	0.595,0.719,0.657	0.335,0.560,0.448	0.557,0.564,0.560	0.353,0.422,0.387	0.524,0.488,0.506	0.341,0.452,0.397
LSAKM	0.627,0.733,0.680	0.444,0.559,0.502	0.517,0.630,0.574	0.313,0.481,0.397	0.550,0.524,0.537	0.314,0.497,0.405
NMF	0.615,0.766,0.691	0.395,0.668,0.532	0.659,0.643,0.651	0.481,0.478,0.479	0.528,0.531,0.530	0.328,0.440,0.384
JPP	0.647,0.724,0.686	0.421,0.670,0.546	0.661,0.651,0.656	0.497,0.412,0.455	0.613,0.582,0.598	0.421,0.512,0.467
SONMFSRd	0.699,0.780,0.740	0.417,0.700,0.558	0.674,0.666,0.670	**0.517**,0.428,0.473	0.624,0.606,0.615	0.433,0.500,0.467
Trendi JD+DT	**0.708**,**0.849**,**0.778**	**0.462**,**0.736**,**0.599**	**0.711**,**0.763**,**0.737**	0.515,**0.499**,**0.507**	**0.643**,**0.662**,**0.653**	**0.447**,**0.563**,**0.505**
	Time 4		**Time 5**		**Time 6**	
	Acc	NMI	Acc	NMI	Acc	NMI
KM	0.360,0.557,0.459	0.129,0.693,0.411	0.655,0.760,0.708	0.420,0.558,0.489	0.331,0.856,0.594	0.395,0.652,0.524
LSAKM	0.335,0.587,0.461	0.129,0.671,0.400	0.622,0.744,0.683	0.355,0.538,0.447	0.388,0.882,0.635	0.157,0.682,0.420
NMF	0.378,0.559,0.469	0.295,0.640,0.468	0.625,0.777,0.701	0.445,0.531,0.488	0.468,0.839,0.653	0.220,0.671,0.446
JPP	0.414,0.622,0.518	0.301,0.647,0.474	0.636,0.757,0.697	0.413,0.546,0.480	0.545,0.816,0.681	0.379,0.660,0.520
SONMFSRd	**0.504**,0.742,**0.623**	0.308,0.669,0.489	0.644,0.787,0.716	0.433,0.567,0.500	**0.552**,0.858,0.705	**0.403**,0.700,0.551
Trendi JD+DT	0.385,**0.807**,0.596	**0.354**,**0.746**,**0.550**	**0.676**,**0.794**,**0.735**	**0.446**,**0.603**,**0.524**	0.495,**0.967**,**0.731**	0.382,**0.790**,**0.586**

been successfully extracted and classified in our model.

Comparison Methods: We compared both the data partition and dynamic topic results with the following competing methods.

Clustering methods. To evaluate the partition results for news and Twitter, we compared the proposed Trendi methods with existing static and dynamic clustering methods, namely (1) Kmeans (KM), (2) Latent semantic analysis (LSA)+Kmeans (LSAKM), (3) Non-negative Matrix Factorization (NMF) [25], (4) Time-based Collective Factorization method (JPP) [14], and (5) Soft Orthogonal NMF with Dynamic Topic Tracking (SONMFSRd) [15]. We also broke our Trendi model into separate parts in order to evaluate them individually. Based on a 10-fold cross validation on the training set, their parameters are set as follows: (1) Single data source only (Trendi SD), in which news and Twitter data were treated separately. We set the parameters $\lambda_1 = 0.6$, $\delta_1 = 0.2$ for news and $\lambda_2 = 0.6$, $\delta_2 = 0.2$ for Twitter and set all the remaining parameters to zero; (2) Single data source with dynamic topics (SD+DT), in which the dynamic topic terms were added to SD. Here, we set the parameters $\theta_1 = 0.3$ and $\varphi_1 = 0.2$ based on SD settings; (3) Joint data sources (Trendi JD), in which news and Twitter data were treated jointly. The parameters were set to $\lambda_1 = \lambda_2 = 0.6$, $\varphi_2 = 0.2$, $\delta_1 = \delta_2 = 0.2$, and all the other parameters to zero; and (4) Joint data with dynamic topics (JD+DT), in which the dynamic topic terms were added to Trendi JD. The parameters here were set at: $\theta_1 = 0.3$ and $\varphi_1 = 0.2$ based on the JD settings.

Dynamic Topic Methods. To evaluate the dynamic topic results, we compared the proposed method, Trendi, with those obtained using recent baseline methods of dynamic topic modeling, including (1) Time-based Collective Factorization method (JPP) [14] and (2) Soft Orthogonal NMF with Dynamic Topic Tracking (SONMFSRd) [15]. All the parameter settings of our model were the same as those used for the clustering settings. The ground truth was generated by the data partition labels according to the following steps. First, the news and tweets were grouped by their labels and timestamps. Secondly, for each group, terms existing in the previous time step were labeled as evolving features, and newly appearing terms labeled as emerging features. The first hundred terms with the highest tf-idf scores were then selected and assigned the equivalent number of emerging and evolving terms as the parameters for our comparison methods. Finally, the ground truth of the dynamic topics



(a) Accuracy of Civil Unrest (b) NMI of Civil Unrest

(c) Accuracy of Mexico (d) NMI of Mexico

Figure 3. Partition Results among Trendi models

were verified manually by domain experts.

B. News & Twitter Partition Results

Preprocessing, including word segmentation, stop words removal, word stemming and filtering were performed for the news and Twitter corpus, after which, the data sets were represented as term-news and term-tweet matrices with *L2* normalization. We compared the results of our proposed new *Trendi* method with those obtained from both static and dynamic topic modeling methods: Kmeans, LSA, NMF, JPP, and SONMFSRd. The accuracy and NMI obtained for the topic partitions in both news and Twitter data sources are listed in Table II and Table III. For each metric, the three columns represent news, Twitter, and the average value of news and Twitter, respectively. The experimental results show that our proposed Trendi model achieved the best overall performance in both news and Twitter dataset. The accuracy and NMI results inform that dynamic topics improve the partition results. Dynamic topic tracking methods such as JPP, SONMFSRd, and JD+DT yielded better results than static topic methods (KM, LSAKM, NMF) for more than 90% time steps.

C. Trendi Variation Models Comparison

The comparison results among our proposed Trendi models with different components are shown in Figure 3. Jointly considering data sources generally improved the partition results compared to the use of a single data source. For example, the Trendi JD outperformed SD model in all the time steps and JD+DT methods competed the SD+DT method in nearly 85% of the time steps. Also, note that the Trendi SD method is a special case of the NMF method with lasso constraints. It is thus not surprising that the results for Trendi SD are very close to the result obtained by static topic models.

D. Dynamic Topic Evaluation

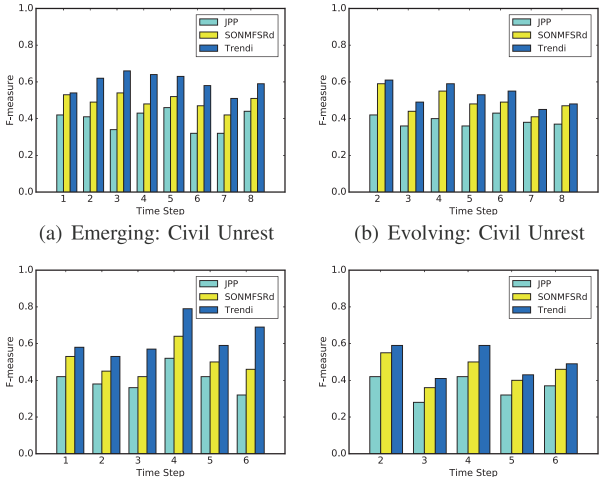
To evaluate the performance of dynamic topic, we compared our method with two dynamic topic modeling methods: JPP[14] and SONMFSRd[15]. The results of the dynamic topic evaluation are depicted in Figure 4. These results show that a method that considers both the dynamic topics and features outperforms methods utilizing either one of them alone. For example, the emerging features identified by our proposed method in Figures 4(a) and 4(c) outperformed the other methods by 20% on average in terms of the F-measure. This is because (1) the SONMFSRd method is designed for dynamic topics only, and thus failed to detect the emerging features in evolving topics, and (2) JPP simply uses a topic-transition matrix to model dynamic topics, without considering the emerging or evolving topics explicitly. Notice that the results of our new method and SONMFSRd in Figures 4(b) and 4(d) are close to each other because emerging topics cannot contain evolving features. As no evolving features yet exist in the first time step, these are left blank in Figures 4(b) and 4(d).

E. Feature Differentiation Analysis

Our *Trendi* model includes two terms that differentiate between the features of dynamic topics and joint sources. Figure 5(a) shows how the similarity between matrices U_v and U_m, which is defined as $\langle U_v, U_m \rangle$, monotonically decreases when φ_1 increases until it converges to a very small number. For the similarity between matrix U_d and U_t, Figure 5(b) shows that the shared features between news and Twitter become fewer and fewer as the parameter φ_2 increases, monotonically decreasing until it converges to a small number close to zero.

F. Efficiency

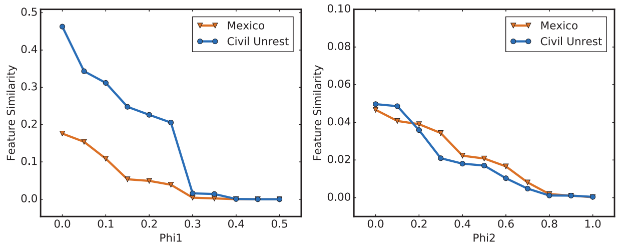
As little work that combines dynamic topics with joint data sources has been reported, we chose methods that considered dynamic topics and ran them individually for



(a) Emerging: Civil Unrest (b) Evolving: Civil Unrest

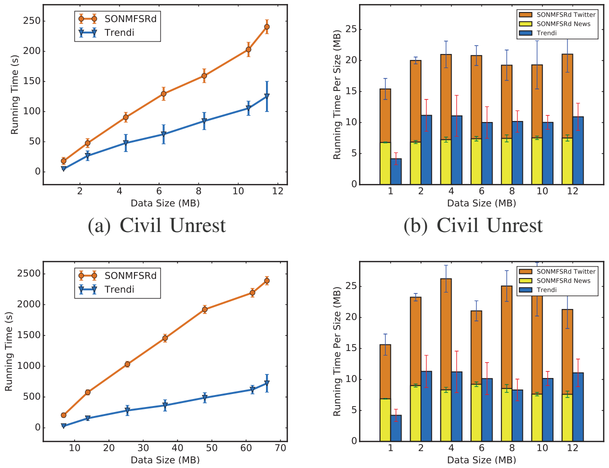
(c) Emerging: Mexico (d) Evolving: Mexico

Figure 4. Dynamic Topic Evaluation



(a) Dynamic topics (b) Joint Sources

Figure 5. Similarity between dynamic topics and joint sources.



(a) Civil Unrest (b) Civil Unrest

(c) Mexico (d) Mexico

Figure 6. Efficiency comparison for the Civil Unrest and Mexico datasets

both news and Twitter data sources. SONMFSRd[15] was chosen as our baseline method for comparison in both the Mexico and Civil Unrest datasets. Here, the efficiency evaluated in terms of the running time for the size of the dataset and the average running time per mega-byte of data. The results shown in Figure 6 indicate that our proposed new method outperforms all the baseline methods tested. Figure 6(a) shows that the running time of our method increased linearly as the data size increased, because the new method processes data incrementally. Figure 6(b) shows the average running time per data size did not fluctuate excessively when the data size was equally partitioned. Figures 6(c) and 6(d) indicate that our method exhibited a better performance than the others in larger dataset. Notice that in the first time period, none of the methods considered the dynamic topics, so the running time for the first time period was naturally shorter than for the subsequent time periods.

G. Case Study

To help readers make sense of complex stories with dynamic topics in news and Twitter datasets, we chose the Iguala Mass Kidnap event in the Mexico dataset for our case study. Reading the story shown in Figure 7 from left to right (ie. following the chronological order of events), our system successfully identified the main details of the key facts. The news titles and tweets are shown in the orange boxes in yellow and purple. The emerging and evolving features are displayed in green and blue circles, respectively. The features are tagged with different colors to represent their different sources: red for common, yellow for news, and purple for Twitter. The details of the story are divided into the following five parts: (1) Students in Ayotzinapa had a fight with police on September 27 and went missing. As this occurred during the first time period, no evolving feature is displayed. Several key facts such as *students*, the state *Guerrero*, and *missing*, are shown as emerging features. Unlike news articles, the features from Twitter are more subjective, using emotive words such as *massacre*, *barbarity*. Also, some hashtags are beginning to be used in Twitter: *#todossomosayotzinapa*, *#justiceforayotzinapa*. (2) Police searched for the missing students. Here, the features that emerged are *parents*, *mayor*, and *rural*, indicating the search actions undertaken by government agencies. Some key background words appear in the evolving features such as *police*, and *missing*. (3) Students' bodies were found in graves. Key words such as *grave*, and *bodies* are found in both the news and Twitter datasets. (4) The Gulf cartel were identified as being responsible for murdering the students, as indicated by key words such as *Gulf*, and *cartel*, *crime* that refer to this fact. In Twitter, a new hashtag *#justiciaayotzinapa* is created to discuss the inherent injustice of the event. (5) Mass protests were held, and the mayor of Iguala placed under arrest. *Aguascalientes*, the location in which the mass protests were held, and mayor *Abarca* are found among the emerging features.

VII. Conclusion

In this paper, a novel story generation framework, *Trendi*, is proposed to combine the information available from news and Twitter data sources that considers emerging, evolving and fading topics. To achieve this, we designed a novel unsupervised model to leverage the dynamic topics in joint data sources, and developed an effective parameter optimization algorithm based on *ADMM*. Extensive experimental results in different domains were conducted to evaluate the effectiveness and efficiency of the proposed model. The results demonstrated that because of the effective utilization of joint data sources and dynamic topics, the proposed model outperforms the all the existing methods used for comparison.

References

[1] Jingjing Wang, Wenzhu Tong, Hongkun Yu, Min Li, Xiuli Ma, Haoyan Cai, Tim Hanratty, and Jiawei Han. Mining multi-aspect reflection of news events in twitter: Discovery, linking and presentation. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 429–438. IEEE, 2015.

[2] Lidan Shou, Zhenhua Wang, Ke Chen, and Gang Chen. Sumblr: continuous summarization of evolving tweet streams. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 533–542. ACM, 2013.

[3] Dafna Shahaf, Jaewon Yang, Caroline Suen, Jeff Jacobs, Heidi Wang, and Jure Leskovec. Information cartography:

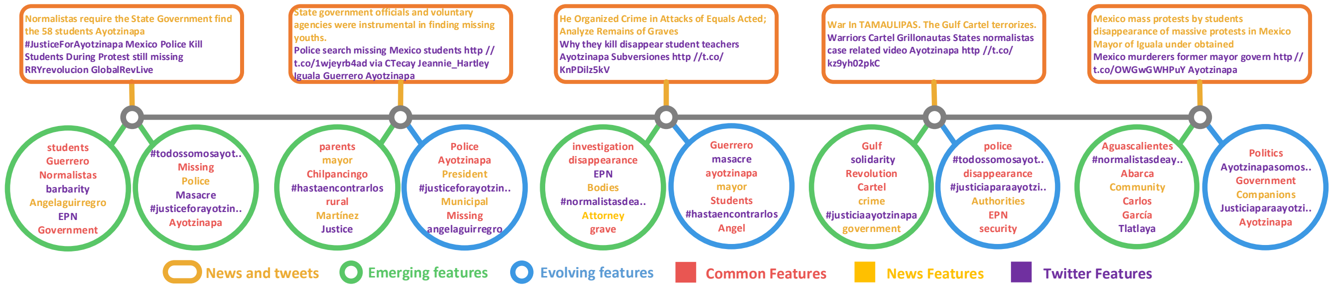

Normalistas require the State Government find the 58 students Ayotzinapa
#JusticeForAyotzinapa Mexico Police Kill Students During Protest still missing RRYrevolucion GlobalRevLive
State government officials and voluntary agencies were instrumental in finding missing youths.
Police search missing Mexico students http //t.co/1wjeyrb4ad via CTecay Jeannie_Hartley Iguala Guerrero Ayotzinapa
He Organized Crime in Attacks of Equals Acted; Analyze Remains of Graves
Why they kill disappear student teachers Ayotzinapa Subversiones http //t.co/KnPDilz5kV
War In TAMAULIPAS. The Gulf Cartel terrorizes.
Warriors Cartel Grillonautas States normalistas case related video Ayotzinapa http //t.co/kz9yh02pkC
Mexico mass protests by students disappearance of massive protests in Mexico Mayor of Iguala under obtained
Mexico murderers former mayor govern http //t.co/OWGwGWHPuY Ayotzinapa
students Guerrero Normalistas barbarity Angelaguirregro EPN Government
#todossomosayot.. Missing Police Masacre #justiceforayotzin.. Ayotzinapa
parents mayor Chilpancingo #hastaencontrarlos rural Martínez Justice
Police Ayotzinapa President #justiceforayotzin.. Municipal Missing angelaguirregro
investigation disappearance EPN Bodies #normalistasdea.. Attorney grave
Guerrero masacre ayotzinapa mayor Students #hastaencontrarlos Angel
Gulf solidarity Revolution Cartel crime #justiciaayotzinapa government
police #todossomosayot.. disappearance #justiciaparaayotzi.. Authorities EPN security
Aguascalientes #normalistasdeay.. Abarca Community Carlos García Tlatlaya
Politics Ayotzinapasomos.. Government Companions Justiciaparaayotzi.. Ayotzinapa
News and tweets
Emerging features
Evolving features
Common Features
News Features
Twitter Features

Figure 7. A sample story line for “Iguala Mass Kidnap” event

creating zoomable, large-scale maps of information. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1097–1105, New York, NY, USA, 2013. ACM.

[4] Chen Lin, Chun Lin, Jingxuan Li, Dingding Wang, Yang Chen, and Tao Li. Generating event storylines from microblogs. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 175–184, New York, NY, USA, 2012. ACM.

[5] Qiaozhu Mei and ChengXiang Zhai. Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 198–207, New York, NY, USA, 2005. ACM.

[6] Theodoros Lappas, Benjamin Arai, Manolis Platakis, Dimitrios Kotsakos, and Dimitrios Gunopulos. On burstiness-aware search for document sequences. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 477–486, New York, NY, USA, 2009. ACM.

[7] Dingding Wang, Li Zheng, Tao Li, and Yi Deng. Evolutionary document summarization for disaster management. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 680–681, New York, NY, USA, 2009. ACM.

[8] Hiroya Takamura, Hikaru Yokono, and Manabu Okumura. Summarizing a document stream. In *Advances in Information Retrieval*, pages 177–188. Springer, 2011.

[9] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120. ACM, 2006.

[10] Tomoharu Iwata, Takeshi Yamada, Yasushi Sakurai, and Naonori Ueda. Online multiscale dynamic topic models. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 663–672, New York, NY, USA, 2010. ACM.

[11] Siwei Lyu and Xin Wang. On algorithms for sparse multi-factor nmf. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 602–610. Curran Associates, Inc., 2013.

[12] Jingu Kim, Yunlong He, and Haesun Park. Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework. *Journal of Global Optimization*, 58(2):285–319, 2013.

[13] Bin Cao, Dou Shen, Jian-Tao Sun, Xuanhui Wang, Qiang Yang, and Zheng Chen. Detect and track latent factors with online nonnegative matrix factorization. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 2689–2694, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[14] Carmen K. Vaca, Amin Mantrach, Alejandro Jaimes, and Marco Saerens. A time-based collective factorization for topic discovery and monitoring in news. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 527–538, New York, NY, USA, 2014. ACM.

[15] Y. Chen, H. Zhang, J. Wu, X. Wang, R. Liu, and M. Lin. Modeling emerging, evolving and fading topics using dynamic soft orthogonal nmf with sparse representation. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 61–70, Nov 2015.

[16] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 338–349, Berlin, Heidelberg, 2011. Springer-Verlag.

[17] Wei Gao, Peng Li, and Kareem Darwish. Joint topic modeling for event summarization across news and social media streams. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1173–1182, New York, NY, USA, 2012. ACM.

[18] Ting Hua, Ning Yue, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. Topical analysis of interactions between news and social media. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 2964–2971, 2016.

[19] Hristo S Sendov. Generalized hadamard product and the derivatives of spectral functions. *SIAM journal on matrix analysis and applications*, 28(3):667–681, 2006.

[20] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[21] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.

[22] Deng Cai, Xiaofei He, Xiaoyun Wu, and Jiawei Han. Non-negative matrix factorization on manifold. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 63–72. IEEE, 2008.

[23] László Lovász and Michael D Plummer. *Matching theory*, volume 367. American Mathematical Soc., 2009.

[24] Quanquan Gu and Jie Zhou. Learning the shared subspace for multi-task clustering and transductive transfer classification. In *Data Mining, 2009. ICDM '09. Ninth IEEE International Conference on*, pages 159–168, Dec 2009.

[25] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791, 1999.