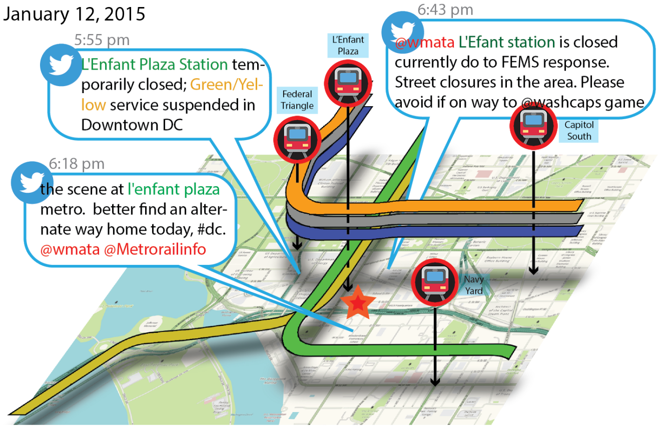
2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Multi-task Learning for Transit Service Disruption Detection

Taoran Ji[1,2,*], Kaiqun Fu[1,2,*], Nathan Self[1,2], Chang-Tien Lu[1,2], and Naren Ramakrishnan[1,2]
[1]Discovery Analytics Center, Virginia Tech, Arlington, VA 22203, USA
[2]Department of Computer Science, Virginia Tech, Arlington, VA 22203, USA

***Abstract*—With the rapid growth in urban transit networks in recent years, detecting service disruptions in a timely manner is a problem of increased interest to service providers. Transit agencies are seeking to move beyond traditional customer questionnaires and manual service inspections to leveraging open source indicators like social media for deteting emerging transit events. In this paper, we leverage Twitter data for early detection of metro service disruptions. Inspired by the multi-task learning framework, we propose the Metro Disruption Detection Model, which captures the semantic similarity between transit lines in Twitter space. We propose novel constraints on feature semantic similarity exploiting prior knowledge about the spatial connectivity and shared tracks of the metro network. An algorithm based on the alternating direction method of multipliers (ADMM) framework is developed to solve the proposed model. We run extensive experiments and comparisons to other models with real world Twitter data and transit disruption records from the Washington Metropolitan Area Transit Authority (WMATA) to justify the efficacy of our model.**

***Index Terms*—Social Media, Twitter, Event Detection, Metro Service Disruption Detection**


January 12, 2015

5:55 pm
L'Enfant Plaza Station temporarily closed; Green/Yellow service suspended in Downtown DC

6:18 pm
the scene at l'enfant plaza metro. better find an alternate way home today, #dc. @wmata @Metrorailinfo

6:43 pm
@wmata L'Efant station is closed currently do to FEMS response. Street closures in the area. Please avoid if on way to @washcaps game

Fig. 1. Tweets related to a station fire at L'Enfant Plaza on January 12, 2015. Twitter data provides timely and near-ubiquitous coverage of metro disruption events across the metro transit network.

I. Introduction

Public transportation plays an important societal role, providing a convenient means of transportation for commuters. The increased development and wide reach of transit networks in recent decades has led more travelers to name public transit as one of their main modes of transportation. According to statistics from the United States Department of Transportation (USDOT) and annual reports from the American Public Transportation Association (APTA) [1], [2], public transportation has seen long-term growth in ridership since the early 1970s with over 44% more trips in 2015. In 2016, miles traveled (nationally) on public transit systems was 58.6 billion miles. This paper focuses on metro transit networks which account for a majority (55%) of total passenger miles in 2016 [2]. In order to provide higher quality experience to riders on metro transit systems, it is crucial to capture and respond to feedback and complaints from riders and to minimize delays across the system. To handle this problem, recent research in both transit system analysis and social media mining have proposed useful methods from different perspectives.

From the perspective of metro management, higher ridership and longer commutes may increase uncertainty in estimating times of arrival and inevitably travel delays for customers. Hence, transit agencies encounter the following challenges. **(1) Can we detect service failures in their early stages?** Existing solutions mostly rely on manual inspections and customer complaints. Such methods cannot provide real-time service failure detection. Some failures are sometimes detected weeks after they arise! **(2) Are service failures causing travel delays for riders?** Current answers to this question focus on statistical analytics of commuter questionnaires. These solutions suffer from insufficient sample points to represent the entire transit network.

Twitter data, on the other hand, can capture commuter feedback and complaints in a timely manner and can further span the entire transit network. We aim to improve the quality of metro service with accurate disruption detection via analysis of a broad range of Twitter comments. Compared to traditional media Twitter has two key properties that are highly suitable for metro disruption detection. *Promptness*: Unlike traditional media which may take hours or even days to be published, tweets are often posted rapidly after an event via ever-present mobile devices [3]. *Geolocated*: According to the latest statistics on Twitter usage, 80% of users post tweets from mobile devices [4]. These features are important for a disruption detection system that is both timely and accurate.

When multiple service complaints are posted near a metro station over a short period of time, we can infer with high confidence that a metro service related incident has occurred. Different metropolitan areas have different watchwords across social media platforms, including different hashtags and influential users who are involved in service disruption communication. Figure 1 shows tweets during a disruption event

[*] Authors have equal contribution.

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00 © 2018 IEEE

that were sent from locations near L'Enfant Plaza, a metro station operated by the Washington Metropolitan Area Transit Authority (WMATA) in the Washington DC region. Local hashtags and influential usernames appear in this example, such as "*#wmata*", "*#wmatafail*", and "*@unsuckdcmetro*".

Two challenges arise in leveraging the rich information Twitter provides for metro service disruption detection. **(1) Sparsity of metro service features**: Among the many text features latent in Twitter data, only a few key features are related to metro services. **(2) Modeling semantic similarity across the problem space**: Although complaints about metro service usually target one particular metro line or station, it is appropriate to assume that language usage patterns are similar across distinct events. To address these challenges, we propose a multi-task learning based Metro Disruption Detection Model (MDDM). To the best of our knowledge, MDDM is the first work to use a supervised learning scheme for metro service disruption detection. Our main contributions are:

- **Formulating a multi-task learning framework for metro disruption detection using online social media.** In contrast to existing works, we formulate the problem of disruption detection for metro service as a multi-task supervised learning problem. In the proposed methods, models for different metro lines are learned simultaneously by restricting all lines to exploit a common set of features.
- **Modeling semantic similarity among metro lines in feature space.** Based on extensive analysis of metro related information on social media, specifically designed constraints are proposed to model semantic similarities among data for distinct metro lines. These similarities in feature space are driven by both spatial connectivity and common complaint vocabulary.
- **Developing an efficient algorithm to solve the proposed model.** The underlying optimization problem of the proposed multi-task model is a non-smooth, multi-convex, inequality-constrained one and challenging to solve. By introducing auxiliary variables, we develop an effective ADMM-based algorithm to decouple the main problem into several subproblems which can be solved by block coordinate descent and proximal operators.
- **Comprehensive experiments to validate the effectiveness and efficiency of the proposed model.** We evaluate the proposed model using metro related Twitter data collected from January 2015 to June 2016. For comparison, we implement a broad range of other algorithms including LOGR, LOG-LASSO, and LOG-RMTFL.

II. Related Work

In this section, we provide a review of current research on social media-based incident detection in transit networks. We break this topic into three subtopics: incident detection in transit networks, local event detection and monitoring with Twitter, and multi-task learning frameworks for spatiotemporal event detection.

A. Incident Detection in Transit Networks

Early detection of emergency incidents on transit and roadway networks is critical for reducing their impact on traffic conditions. This problem has drawn increased attention from researchers in the field of intelligent transportation systems. Previous studies focused on using multiple sources for incident detection including reports from transit operation patrols, commuter calls, traffic sensors, and closed circuit television (CCTV) monitoring [5], [6]. These traditional incident detection methods suffer from two major disadvantages: lack of incident detection sensitivity (e.g. patrol inspection, commuter reports), and limited urban areas that are monitored (e.g. traffic sensors, CCTV). These drawbacks motivate us to explore the application of social media analytics for incident detection and transit network management.

In recent years, there have been some interest in social media based event analysis for transit networks and transportation systems [7], [8], [9]. Most of these papers focus on statistical analysis such as the correlation between the volume of social media posted and the number of occurrences of transportation related events [10]. These methods rely mostly on analysis from human experts and are impractical to automate. Another branch of relevant work focuses on implementing heuristic rule-based models to predict the occurrence of events in transit networks. Ma et al. proposed a mobility analyzer framework [11] which consists of a social media based event detection module for transit networks.

B. Local Event Detection and Monitoring on Twitter

Several previous studies have used social media data for local event detection problems. Sakaki et al. [12] trained a prediction model to judge whether a newly posted tweet refers to an earthquake. Zhang et al. [13] used taxi trace records to infer occurrences of social events. Furthermore, they propose a model for measuring the scale and impact of those events. Santillana et al. [14] explored the feasibility of applying machine learning algorithms to detect and monitor influenza activity by leveraging data from multiple data sources including Google searches, Twitter data, and hospital visit records. Gerber et al. [15] used a logistic regression model to forecast criminal activities in a spatio-temporal setting. Paul et al. [16] and Parker et at. [17] proposed methods to track public health conditions via social media data sources. Alternatives to traditional unsupervised machine learning methods such as latent Dirichlet allocation were proposed by Chen et al. [18] to detect public health related events. They further show that their methods can better forecast a flu season's trends as well as flu-peaks by aggregating user states in a region over a period.

C. Multi-task Learning for Spatiotemporal Event Detection

Multi-task learning (MTL) refers to models that learn multiple related tasks simultaneously to improve overall performance. Recent decades have witnessed proposals for many MTL approaches [19]. Evgeniou et al. [20] proposed a regularized MTL formulation that constrains the models of each task to be close to each other. Task relatedness can also be modeled by constraining multiple tasks to share a common underlying structure (e.g. a common set of features) [21], or a common subspace [22]. Zhao et al. [23] designed a multi-task learning framework that models forecasting tasks in related geolocations. MTL approaches have been applied in many domains including computer vision and biomedical informatics. Our work, to the best of our knowledge, is the first paper to address the feasibility of combining social media analysis and multi-task learning techniques to resolve disruption detection problems for transit networks.

III. Problem Setup

Given a collection of tweets $\mathcal{D}$, which is collected along a continuous time series, we first filter it using names of metro stations and metro lines operated by WMATA. This produces the target tweet subcollection $\mathcal{D}^+$. Then based on which metro line is referred to in each tweet, $\mathcal{D}^+$ is grouped into $\{\mathcal{D}^+_c\}^{c\in\Phi}$, where $\Phi = \{\text{blue}, \text{green}, \text{orange}, \text{red}, \text{silver}, \text{yellow}\}$ (the colors refer to the WMATA metro lines.)

Our operative question is: given a metro line color c, a time slot t, and the collection of corresponding tweets $\mathcal{D}^+_{c,t}$, is there a delay for metro line c during time period t? To answer this question, we cast it as a supervised learning problem using the multi-task learning framework.

Under the assumption that metro delays can be captured by complaints and negative discussion in Twitter space, we adopt a dictionary $\mathcal{F}$ of features trained specifically for Twitter [24]. For each subcollection $\mathcal{D}^+_{c,t}$, we generate a corresponding matrix $\mathbf{X}^c_t$ by counting the frequencies of semantic features in $\mathcal{F}$. Now, our problem can be formulated as performing the mapping

$$F_c(\mathbf{X}^c_t) \rightarrow \mathbf{Y}^c_t, \tag{1}$$

where $\mathbf{Y}^c_t \in \{-1, 1\}$ are labels which denote if there is a delay, and F_c is the model for metro line c. (In the case of WMATA which operates six metro lines, there are six models to learn.)

A traditional way to solve this problem is to learn the model for each metro line separately. However, the performance of each model may be adversely affected by ignoring the relatedness among different lines. In our approach, this relatedness is expressed as the semantic similarities among complaints about different metro lines and stations. We consider that two factors contribute to this semantic similarity in Twitter space. **(1) Spatial connectivity of metro lines**: Metro lines are spatially related together (e.g. the orange and silver lines share 83% of their stations). As a result, delays that affect multiple lines may provoke similar complaints. **(2) Common complaint vocabulary across metro lines**: We assume that the words used by Twitter users to complain of disruption events will be similar across all metro lines. To model semantic similarity caused by these two factors, we design and implement a multi-task learning based metro delay detection model. The details are explained in the next section.

IV. Models

Considering that we want to predict if there is a delay for a metro line given a subcollection of tweets $\mathcal{D}^+_{c,t}$ which mention metro line c during time slot t, our problem fits well into the scope of a classification or regression problem. For instance, learning the function F_c can be modeled as a logistic regression problem and the model parameters $\mathbf{w}$ can be learned by solving the following optimization problem:

$$\underset{\mathbf{w}}{\operatorname{argmin}}\ \mathcal{L}_c = \sum_{t=1}^{m_c} \log\left(1 + \exp\{\mathbf{Y}^c_t(\mathbf{X}^c_t\mathbf{w})\}\right), \tag{2}$$

where m_c is the total number of data points in $\mathcal{D}^+_{c,t}$. However, as stated in Section III, if $\mathbf{w}$ for each metro line is learned separately, these models will fail to reflect the semantic similarity among metro lines in feature space. To solve these challenges, we cast the original problem into a multi-task learning framework:

$$\underset{\mathbf{W}}{\operatorname{argmin}}\ \mathcal{L} = \sum_{c=1}^{|\Phi|}\sum_{t=1}^{m_c} \log\left(1 + \exp\{-\mathbf{Y}^c_t(\mathbf{X}^c_t\mathbf{W}^c)\}\right), \tag{3}$$

where each column of $\mathbf{W}$, referred to as $\mathbf{W}^c$, denotes the model parameters of F_c. In this way, we can further model the relatedness among metro lines with parameter matrix $\mathbf{W}$.

A. Modeling Spatial Connectivity in Feature Space

In real world metro systems, distinct metro lines are often spatially related. That is, two or more metro lines may share several stations or several segments of track. For instance, in the Washington metro system, the Orange and Silver lines share 83% of their stations; and the Silver and Blue lines share 64% of their stations. This means that if the Orange line has a delay, there is a good chance that the Silver line will also have a delay. This spatial relatedness results in semantic similarity in Twitter space and, therefore, a similar distribution of tweets complaining of delays. For example, one complaint from our dataset mentions three different metro lines at once: "@unsuckmetro. Just spend 30 min @foggy bottom & McPherson #metro. No explanation. **Blue/silver/orange** #delays. #Transparency much? @wmata". Thus, our model should be encouraged to capture this form of semantic relatedness in Twitter space. Mathematically, we place constraints on parameters among different tasks

$$\begin{aligned}
\underset{\mathbf{W}}{\operatorname{argmin}}\ & \sum_{c=1}^{|\Phi|}\sum_{t=1}^{m_c} \log\left(1 + \exp\{-\mathbf{Y}^c_t(\mathbf{X}^c_t\mathbf{W}^c)\}\right) \\
\text{s.t. } & \|\mathbf{W}^3 - \mathbf{W}^5\|_2^2 \le \eta_1, \|\mathbf{W}^1 - \mathbf{W}^5\|_2^2 \le \eta_2 \\
& \|\mathbf{W}^6 - \mathbf{W}^2\|_2^2 \le \eta_3, \|\mathbf{W}^1 - \mathbf{W}^6\|_2^2 \le \eta_4 \\
& \eta_1 \ge 0, \eta_2 \ge 0, \eta_3 \ge 0, \eta_4 \ge 0,
\end{aligned} \tag{4}$$

where each constraint forces the Euclidean distance between model parameters for a specific pair of metro lines to be within a range. Detailed explanations for each constraint are shown in Table I and their expected effects are explained in constraints #1 — #4 in Figure 2.

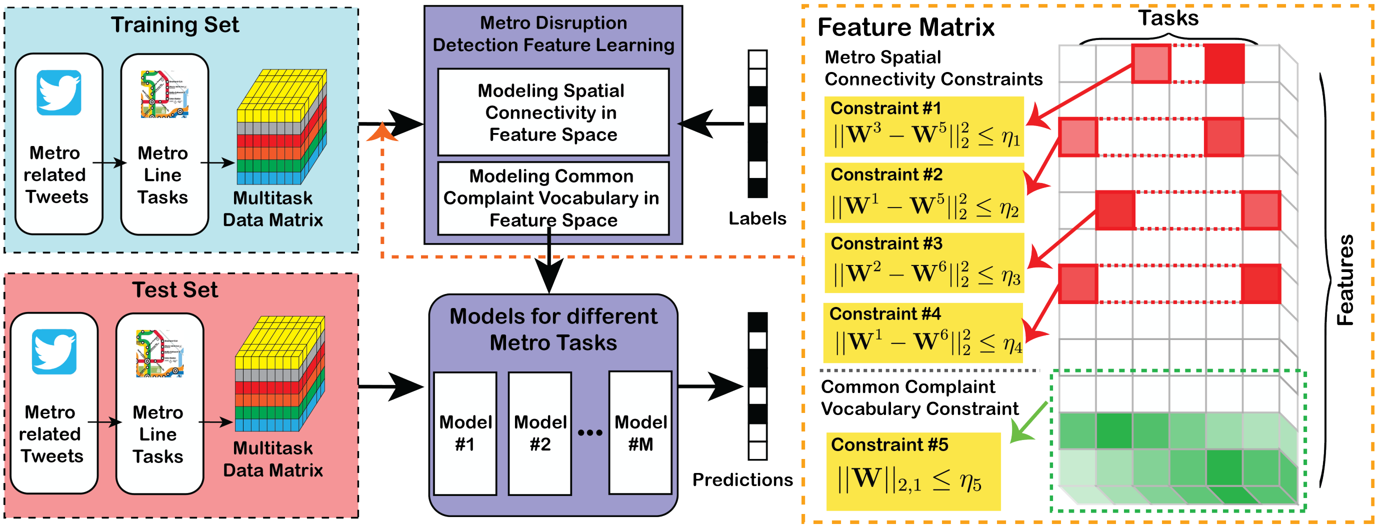

Training Set
Metro related Tweets
Metro Line Tasks
Multitask Data Matrix
Test Set
Metro related Tweets
Metro Line Tasks
Multitask Data Matrix
Metro Disruption Detection Feature Learning
Modeling Spatial Connectivity in Feature Space
Modeling Common Complaint Vocabulary in Feature Space
Labels
Models for different Metro Tasks
Model #1
Model #2
...
Model #M
Predictions
Feature Matrix
Tasks
Metro Spatial Connectivity Constraints
Constraint #1
$\|\mathbf{W}^3 - \mathbf{W}^5\|_2^2 \leq \eta_1$
Constraint #2
$\|\mathbf{W}^1 - \mathbf{W}^5\|_2^2 \leq \eta_2$
Constraint #3
$\|\mathbf{W}^2 - \mathbf{W}^6\|_2^2 \leq \eta_3$
Constraint #4
$\|\mathbf{W}^1 - \mathbf{W}^6\|_2^2 \leq \eta_4$
Common Complaint Vocabulary Constraint
Constraint #5
$\|\mathbf{W}\|_{2,1} \leq \eta_5$
Features

Fig. 2. A schematic view of the Metro Disruption Detection Model (MDDM). Semantic similarity among complaints in feature space is modeled by two major factors: spatial connectivity between metro lines and a common complaint vocabulary. In particular, metro spatial connectivity constraints encourage the model to decrease differences between spatially related metro lines in feature space. The common complaint vocabulary constraint encourages the model to identify a core set of words most commonly used for reporting problems with metro service.

TABLE I
MEANING OF CONSTRAINTS.

Constraint	Meaning
$\|\mathbf{W}^3 - \mathbf{W}^5\|_2^2 \leq \eta_1$	Euclidean distance between (**orange**, **silver**) should be less than or equal to η_1.
$\|\mathbf{W}^1 - \mathbf{W}^5\|_2^2 \leq \eta_2$	Euclidean distance between (**blue**, **silver**) should be less than or equal to η_2.
$\|\mathbf{W}^6 - \mathbf{W}^2\|_2^2 \leq \eta_3$	Euclidean distance between (**yellow**, **green**) should be less than or equal to η_3.
$\|\mathbf{W}^1 - \mathbf{W}^6\|_2^2 \leq \eta_4$	Euclidean distance between (**blue**, **yellow**) should be less than or equal to η_4.

B. Modeling Common Complaint Vocabulary in Feature Space

In addition to similarities caused by the spatial interconnectedness between metro lines, we also consider a hidden pattern in the usage of complaint words over time and across metro lines. For example, consider the following two tweets "Come on @wmata — why so many **delays** on the blue line? Been trying to get home since 5:45." and "**Delays** every day this week on OL. Was #SafeTrack just a taxpayer money grab? @unsuckdcmetro @wmata". These were posted at different timestamps and, although they complain about the blue line and the orange line respectively, they share a common complaint word: "delay". This observation leads us to believe that although we have adopted a large dictionary of semantic keywords, it is possible that only a small subset of them, like "delay," contribute to the detection of metro disruptions. This means that the learned parameters matrix $\mathbf{W}$ should be sparse and have nonzero values for only the most important features. Thus, the proposed model should be encouraged to capture hidden patterns among complaints and to maintain sparsity in feature space. Mathematically, this consideration inspires us to use the $\ell_{2,1}$ [25] norm to perform joint feature selection:

$$\begin{aligned}\underset{\mathbf{W}}{\operatorname{argmin}} & \sum_{c=1}^{|\Phi|}\sum_{t=1}^{m_c} \log\left(1+\exp\{-\mathbf{Y}_t^c(\mathbf{X}_t^c\mathbf{W}^c)\}\right) \\ \text{s.t. } & \|\mathbf{W}\|_{2,1} \leq \eta_5, \eta_5 \geq 0.\end{aligned} \tag{5}$$

The effect of $\|\mathbf{W}\|_{2,1}$ is explained in constraint #5 in Figure 2.

C. Metro Disruption Detection Model

Combining Model 4 and Model 5 together, we obtain our proposed metro disruption detection model:

$$\begin{aligned}\underset{\mathbf{W}}{\operatorname{argmin}} & \sum_{c=1}^{|\Phi|}\sum_{t=1}^{m_c} \log\left(1+\exp\{-\mathbf{Y}_t^c(\mathbf{X}_t^c\mathbf{W}^c)\}\right) \\ \text{s.t. } & \|\mathbf{W}^3 - \mathbf{W}^5\|_2^2 \leq \eta_1, \|\mathbf{W}^1 - \mathbf{W}^5\|_2^2 \leq \eta_2 \\ & \|\mathbf{W}^6 - \mathbf{W}^2\|_2^2 \leq \eta_3, \|\mathbf{W}^1 - \mathbf{W}^6\|_2^2 \leq \eta_4 \\ & \|\mathbf{W}\|_{2,1} \leq \eta_5, \\ & \eta_1 \geq 0, \eta_2 \geq 0, \eta_3 \geq 0, \eta_4 \geq 0, \eta_5 \geq 0.\end{aligned} \tag{6}$$

By moving the constraints to an objective function, we can obtain an equivalent regularized problem, which is easier to solve

$$\begin{aligned}\underset{\mathbf{W}}{\operatorname{argmin}} & \sum_{c=1}^{|\Phi|}\sum_{t=1}^{m_c} \log\left(1+\exp\{-\mathbf{Y}_t^c(\mathbf{X}_t^c\mathbf{W}^c)\}\right) + \lambda_5\|\mathbf{W}\|_{2,1} \\ & + \lambda_1\|\mathbf{W}^3 - \mathbf{W}^5\|_2^2 + \lambda_2\|\mathbf{W}^1 - \mathbf{W}^5\|_2^2 \\ & + \lambda_3\|\mathbf{W}^6 - \mathbf{W}^2\|_2^2 + \lambda_4\|\mathbf{W}^1 - \mathbf{W}^6\|_2^2,\end{aligned} \tag{7}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, and λ_5 are trade-off penalties balancing the value of the loss function and the regularizers.

V. Algorithm

The objective function 7 is multi-convex and the regularizer $\ell_{2,1}$ is non-smooth. This increases the difficulty of solving this problem. A traditional way to solve this kind of problem is to use proximal gradient descent. But this approach is slow to converge. Recently, the alternating direction method of multipliers (ADMM) [26] has become popular as an efficient algorithm framework which decouples the original problem into smaller and easier to handle subproblems. Here we propose an ADMM-based algorithm 1 which is able to efficiently optimize the proposed models. In particular, primal variables are updated on Line 4, dual variables on Line 5, and Lagrange multipliers on Line 6. Line 7 calculates both primal and dual residuals.

A. Augmented Lagrangian Scheme

First, we introduce an auxiliary variable $\mathbf{U}_w = \mathbf{W}$ into the original problem 7 and obtain the following equivalent problem:

$$
\begin{aligned}
\underset{\Theta}{\operatorname{argmin}} \; & \sum_{c=1}^{|\Phi|}\sum_{t=1}^{m_c} \log\left(1+\exp\{-\mathbf{Y}_t^c(\mathbf{X}_t^c\mathbf{W}^c)\}\right) \\
& + \lambda_1\|\mathbf{W}^3-\mathbf{W}^5\|_2^2 + \lambda_2\|\mathbf{W}^1-\mathbf{W}^5\|_2^2 \\
& + \lambda_3\|\mathbf{W}^6-\mathbf{W}^2\|_2^2 + \lambda_4\|\mathbf{W}^1-\mathbf{W}^6\|_2^2 \\
& + \lambda_5\|\mathbf{U}_w\|_{2,1} \\
\text{s.t. } & \mathbf{U}_w = \mathbf{W},
\end{aligned} \tag{8}
$$

where $\Theta = \{\mathbf{U}_w, \mathbf{W}\}$ is the set of variables to be optimized. Then we transform the above problem into its augmented Lagrangian form as follows:

$$
\begin{aligned}
\mathcal{L}_\rho = & \sum_{c=1}^{|\Phi|}\sum_{t=1}^{m_c} \log\left(1+\exp\{-\mathbf{Y}_t^c(\mathbf{X}_t^c\mathbf{W}^c)\}\right) \\
& + \lambda_1\|\mathbf{W}^3-\mathbf{W}^5\|_2^2 + \lambda_2\|\mathbf{W}^1-\mathbf{W}^5\|_2^2 \\
& + \lambda_3\|\mathbf{W}^6-\mathbf{W}^2\|_2^2 + \lambda_4\|\mathbf{W}^1-\mathbf{W}^6\|_2^2 \\
& + \lambda_5\|\mathbf{U}_w\|_{2,1} + \langle \mathbf{U}_1, \mathbf{W}-\mathbf{U}_w\rangle \\
& + \frac{\rho}{2}\|\mathbf{W}-\mathbf{U}_w\|_2^2.
\end{aligned} \tag{9}
$$

where $\mathbf{U}_1$ is the Lagrangian multiplier. With this step, we decouple the original problem into two easier to handle problems in which three variables $\mathbf{W}$, $\mathbf{U}_w$ and $\mathbf{U}_1$ will be optimized individually.

B. Parameter Optimization

First, the primal variable $\mathbf{W}$ is updated by solving the following subproblem:

$$
\begin{aligned}
\mathbf{W}^+ \leftarrow \underset{\mathbf{W}}{\operatorname{argmin}} \; \mathcal{Q} = & \sum_{c=1}^{|\Phi|}\sum_{t=1}^{m_c} \log\left(1+\exp\{-\mathbf{Y}_t^c(\mathbf{X}_t^c\mathbf{W}^c)\}\right) \\
& + \lambda_1\|\mathbf{W}^3-\mathbf{W}^5\|_2^2 + \lambda_2\|\mathbf{W}^1-\mathbf{W}^5\|_2^2 \\
& + \lambda_3\|\mathbf{W}^6-\mathbf{W}^2\|_2^2 + \lambda_4\|\mathbf{W}^1-\mathbf{W}^6\|_2^2 \\
& + \langle \mathbf{U}_1, \mathbf{W}-\mathbf{U}_w\rangle + \frac{\rho}{2}\|\mathbf{W}-\mathbf{U}_w\|_2^2.
\end{aligned} \tag{10}
$$

Algorithm 1: An ADMM-based solver for MDDM.

```
   Input: $\mathbf{X}, \mathbf{Y}$
   Output: $\mathbf{W}$
1  Initialize $\rho = 1$, $\mathbf{W}^{(0)}$, $\mathbf{U}_w^{(0)}$, $\mathbf{U}_1^{(0)}$;
2  Initialize $\epsilon^r > 0, \epsilon^s > 0$, MAX_ITER;
3  for $k = 1$ : MAX_ITER do
4      Update $\mathbf{W}^{(k)}$ with BCD using 12;
5      Update $\mathbf{U}_w^{(k)}$ with $\text{prox}_{f_1,1/\rho}(\mathbf{U}_1^{(k-1)} + \mathbf{W}^{(k)})$;
6      Update $\mathbf{U}_1^{(k)}$ with Equation 14;
7      Compute $r$ and $s$ by Equations 15;
8      if $r < \epsilon^r$ and $s < \epsilon^s$ then
9          break;
10     end
11 end
```

The objective function $\mathcal{Q}$ is multi-convex. In particular, $\mathcal{Q}$ of $\mathbf{W}^j$ is convex where all other $\mathbf{W}^{j' \neq j}$ are fixed. This kind of problem can be decoupled into subproblems using block coordinate descent (BCD) [27], in which each $\mathbf{W}^j$ is updated by solving the following sub-optimization problems:

$$
\mathbf{W}^j \leftarrow \underset{\mathbf{W}^j}{\operatorname{argmin}} \; \mathcal{Q}. \tag{11}
$$

$\mathcal{Q}$ is smooth and convex for each $\mathbf{W}^j$ and can be solved by gradient descent as follows:

$$
\begin{aligned}
\frac{\partial \mathcal{Q}}{\partial \mathbf{W}^1} &= \mathcal{P}(1) + 2\lambda_2(\mathbf{W}^1-\mathbf{W}^5) + 2\lambda_4(\mathbf{W}^1-\mathbf{W}^6), \\
\frac{\partial \mathcal{Q}}{\partial \mathbf{W}^2} &= \mathcal{P}(2) + 2\lambda_3(\mathbf{W}^2-\mathbf{W}^6), \\
\frac{\partial \mathcal{Q}}{\partial \mathbf{W}^3} &= \mathcal{P}(3) + 2\lambda_1(\mathbf{W}^3-\mathbf{W}^5), \\
\frac{\partial \mathcal{Q}}{\partial \mathbf{W}^4} &= \mathcal{P}(4), \\
\frac{\partial \mathcal{Q}}{\partial \mathbf{W}^5} &= \mathcal{P}(5) - 2\lambda_1(\mathbf{W}^3-\mathbf{W}^5) - 2\lambda_2(\mathbf{W}^1-\mathbf{W}^5), \\
\frac{\partial \mathcal{Q}}{\partial \mathbf{W}^6} &= \mathcal{P}(6) + 2\lambda_3(\mathbf{W}^6-\mathbf{W}^2) - 2\lambda_4(\mathbf{W}^1-\mathbf{W}^6),
\end{aligned} \tag{12}
$$

where

$$
\begin{aligned}
\mathcal{P}(c) = & (\mathbf{X}^c)^{\mathrm{T}}\left(-\mathbf{Y}^c \circ (\mathbf{I} - \mathbf{I} \oslash (\mathbf{I} + \exp\{\mathbf{Z}^c\}))\right) \\
& + \mathbf{U}_1^c + \rho\left(\mathbf{W}^c - \mathbf{U}_w^c\right)
\end{aligned}
$$

where $\circ$ is the element-wise product (Hadamard product), $\oslash$ is element-wise division (Hadamard division), $\mathbf{I}$ is a m_c-dimensional vector of ones, and $\mathbf{Z}^c$ is defined as

$$
\mathbf{Z}^c = -\mathbf{Y}^c \circ (\mathbf{X}^c\mathbf{W}^c).
$$

Now that primal variable $\mathbf{W}$ is taken care of, the dual variable $\mathbf{U}_w$ is updated as follows:

$$
\mathbf{U}_w^+ \leftarrow \text{prox}_{f_1,1/\rho}(\mathbf{U}_1 + \mathbf{W}), \tag{13}
$$

where f_1 is the non-smooth function $\lambda_5\|\mathbf{U}_w\|_{2,1}$. The proximal operator can be solved efficiently using [28].

Next, the Lagrangian multiplier $\mathbf{U}_1$ is updated as follows:

$$\mathbf{U}_1^+ \leftarrow \mathbf{U}_1 + \rho(\mathbf{W}^+ - \mathbf{U}_w^+). \tag{14}$$

Finally, primal and dual residuals are computed with

$$r = \|\mathbf{W}^+ - \mathbf{U}_w^+\|_2, s = \rho\left(\|\mathbf{U}_w^+ - \mathbf{U}_w\|_2\right). \tag{15}$$

where r is primal residual, and s is dual residual.

VI. Evaluation

A. Dataset and Ground Truth

Dataset and Preprocessing: We evaluated our proposed Metro Disruption Detection Model using tweets collected from January 2015 through June 2016 from GNIP's decahose (an approx. 10% sample of all tweets). This dataset was separated into two parts: (1) data from January 2015 to December 2015, which serves as the training set for supervised comparison methods, and (2) data from January 2016 to June 2016, which serves as the test set for validating our methods. Both the training set and test set were partitioned into hourly intervals. Event detection is performed for each metro line individually based on each hour's data. Each tweet collection is fed into a preprocessing pipeline during which stopwords are eliminated; enrichment involving tokenization and lemmatization is performed using SpaCy.

Sentiment Features Selection: Our approach derives features from tweet content using a pipeline of pruning operations and social media sentiment data provided by the StaticTwitterSent dictionary [29]. First, we removed non-English words, hashtags, usernames, single letter words, numbers, and English stopwords as defined by the NLTK toolkit. With our dataset, this produced 17,977 features. Secondly, we removed words with positive frequency or negative frequency less than 10 from the feature set. Here, positive frequency refers to the number of times a word appears in a positive sentiment tweet. Likewise, negative frequency denotes the frequency of a word in tweets with negative sentiment. As a result, a set of 6,600 words with negative sentiment were selected. Thirdly, we used metro related keywords (e.g. metro, wmata) provided by metro experts to gather a sub-collection of tweets from the entire dataset. Only features provided by tweets in this metro-related sub-collection are used. This leaves us with a set of 2,200 features which we use for validating our methods. The most frequent features for each metro line are listed in Table II. Note that several keywords (e.g. interrupted, injury) are shared among different metro lines, further motivating us to train all tasks simultaneously.

Metro Delay Disruption Dataset: The WMATA Daily Service Report keeps track of all delay records in the WMATA Metro system. For this paper, we scraped metro disruption data from this service for each day from January 2015 to June 2016. This dataset serves as ground truth data for training and validation. We scrape data such as the amount of time delayed, metro line color, metro station and start time of the delay, which is aggregated into six-hour time interval. This data is separated into training and test sets corresponding to the Twitter dataset's partition.

TABLE II
Top 3 most frequent words for each metro line.

Metro Line	Keywords		
Blue line	mole	interrupted	enders
Green line	interrupted	injury	spine
Orange line	spine	mouldy	safe
Red line	interrupted	injury	computers
Silver line	injury	interrupted	blister
Yellow line	interrupted	injury	tennant

B. Comparison Methods and Experiment Setup

To quantify and validate model performance, different metrics are adopted. Precision denotes the ratio of detected metro disruptions in ground truth over all predicted disruptions by a model. Recall designates the percentage of metro disruptions which are actually detected by the model. F-measure is the harmonic mean of precision and recall which is defined as $2\cdot$ Precision $\cdot$ Recall / (Precision + Recall).

There are five tunable parameters in our MDDM model, namely the regularizer penalties $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and λ_5. During the experiment, we observe that the value of the loss function is significantly larger than regularizer 2 — 4, which means a large penalty should be used to balance the loss function and the regularizers. Thus, λ_1 is searched from $\{1, 10, 100\}$ and the rest four penalties is searched from $\{100, 200, \ldots, 1000\}$.

We compared MDDM with the following three methods:

- **Logistic Regression (LOGR)** [30]. For each metro line, LOGR utilizes a logit function to predict the probability of the occurrence of disruption of metro service based on the observation of tweets. Input features are features count, and no tunable parameter is needed.
- **LASSO with Logistic Regression (LOG-LASSO)** LASSO [31] is a classic model which is widely used in the event detection field. For each metro line $\mathbf{X}^c$, a LASSO model is trained and evaluated independently. Input features are feature count and the The trade-off penalty parameter is searched from $\{1,\ 10,\ 100\}$.
- **Regularized Multi-task Feature Learning Model with Logistic Regression (LOG-RMTFL)** [23]. We replace the least squares loss with logistic loss to fit our proposed classification application. The feature set is our set of 2,200 words with negative sentiment. All six models are trained simultaneously and evaluated separately. The trade-off penalty parameter is searched from $\{1,\ 10,\ 100\}$.

C. Metro Disruption Detection Results

Table III summarizes the comparisons of our proposed method to the competing methods for the task of disruption detection of metro service. From the experimental results, we can justify our application of a multi-task learning framework for detecting disruptions in metro service. In general, MDDM outperforms the single task model on five metro lines on recall and F-measure due to the benefits of multi-task learning which integrates information from each transit line. LOG-LASSO, LOG-RMTFL, and MDDM outperform LOGR for everything

TABLE III
Metro Disruption Detection Performance Comparisons (Precision, Recall, F-measure)

Method	Blue	Green	Orange	Red	Silver	Yellow
LOGR	0.37, 0.37, 0.37	0.48, 0.48, 0.48	0.56, 0.55, 0.56	0.63, 0.63, 0.63	0.49, 0.49, 0.49	0.40, 0.40, 0.40
LOG-LASSO	**0.43**, 0.42, 0.42	0.50, 0.50, 0.50	0.56, 0.56, 0.56	0.67, 0.66, 0.67	0.49, 0.49, 0.49	**0.43**, **0.43**, **0.43**
LOG-RMTFL	0.36, 0.35, 0.35	0.52, 0.52, 0.52	0.59, 0.59, 0.59	0.68, **0.68**, 0.68	0.49, 0.49, 0.49	0.39, 0.39, 0.39
MDDM	0.36, **0.56**, **0.44**	**0.55**, **0.55**, **0.55**	**0.63**, **0.62**, **0.63**	**0.69**, **0.68**, **0.69**	**0.52**, **0.52**, **0.52**	0.40, 0.40, 0.40

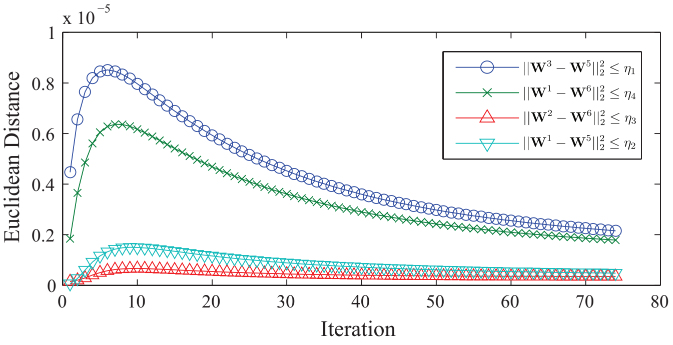


Fig. 3. Spatial connectivity component validation. To obtain an optimal value for the objective function, the model iteratively decreases the value of the loss function and reduces the penalties of the regularizers. As a result, the differences between spatially related metro lines such as Orange ($\mathbf{W}^3$) and Silver ($\mathbf{W}^5$) lines decrease until reaching a stable state.

except precision on the blue line. Considering that these three models account for feature sparsity, these results demonstrate the existence of sparsity in metro service features as introduced in Section I. Table III also shows that model performance for metro service disruption detection is not the same across different transit lines. For instance, the performance of LOG-LASSO, LOG-RMTFL, and MDDM on the red line only differs slightly. Because the red line is spatially independent from other lines in the WMATA network, performance in not greatly affected by the application of a multi-task learning model. Also, LOG-RMTFL and MDDM perform better on the Orange line. The Orange line has unusual properties. For instance, it shares many feature weights with other transit lines and its spatial connectivity to other lines (such as Blue and Silver) is high. This enables multi-task learning based methods to utilize as much information as possible to boost performance on the Orange line. We find that, across metro lines, MDDM outperforms LOG-RMTFL by 1% to 10% on precision and by 2% to 25% on F-measure. These results demonstrated that the consideration of spatial connectivity in transit networks contributes to improved detection.

To further demonstrate the impact of modeling spatial connectivity, as proposed in Section IV, Figure 3 shows the development of the Euclidean distance for each constraint in Table I. As shown in the figure, distance increases at first because $\mathbf{W}$ is initialized as a matrix of zeros. But, at higher iteration counts, distance decreases until it reaches a stable state. In regards to balancing the value of the loss function and the regularizers, our MDDM approach does learn to optimize for the similarity between spatially connected transit lines during each iteration step.

D. Case Studies

To justify our proposed method, we present a qualitative analysis of a real world case study. Figure 4 shows disruption events for 2015 on the Orange, Silver, and Blue lines operated by the Washington Metropolitan Transit Authority. Disruption 1, disruption 2, and disruption 3 occurred on Orange line. Disruption 4 and disruption 5 occurred on Silver line. Disruption 6, disruption 7 and disruption 8 occurred on Blue line. Our MDDM model successfully detects disruptions 1, 4 and 6. Because MDDM uses a multi-task framework to jointly learn models for all metro lines, it can detect co-occurring events using data from other lines even when a model has few training samples. The LOG-LASSO model only detects disruption 1. Because the training step for each metro line is independent in LOG-LASSO, its performance suffers from lack of training samples. Although LOG-RMTFL detects disruptions 2, 3, 7, and 8, it does not perform well on the Silver line. That's because it does not model spatial connectivity which could boost performance on the Silver line by training together with the spatially interconnected Orange line.

VII. Conclusion

We have demonstrated a multi-task learning based supervised learning model for the problem of metro delay detection using social media data. We have motivated the need for training all tasks simultaneously instead of building a model for each line separately. Our work considers the unique metro-specific assumptions in feature space, reflected in the two kinds of regularizers proposed in the model. We proposed an efficient algorithm based on the ADMM framework in which the main problem is divided into several sub-problems which can be solved using block coordinate descent and proximal operators. Our empirical results demonstrate that our proposed model can effectively detect metro delays even with very weak signals in social media space and outperform competing methods by a substantial margin on both precision and recall. For future work, we plan to extend our model to use multiple data sources including transportation related data.

Acknowledgment

This work is supported in part by the National Science Foundation via grants DGE-1545362, IIS-1633363, and by the Army Research Laboratory under grant W911NF-17-1-0021. The US Government is authorized to reproduce and distribute

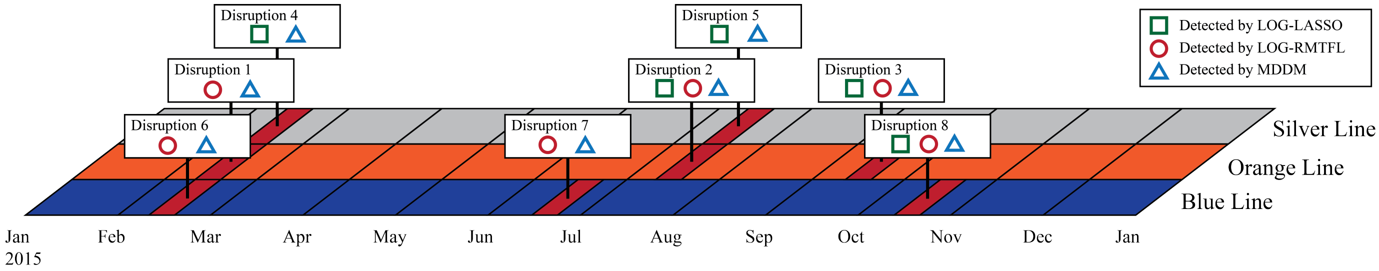


Fig. 4. A timeline of metro disruptions on the Orange, Blue, and Silver metro lines in 2015. Events along these spatially interconnected lines often co-occur.

reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF, Army Research Laboratory, or the U.S. Government.

References

[1] E. L. Chao, J. Rosen, P. Hu, and R. Schmitt, "Transportation statistics annual report, 2017," 2018.

[2] J. Neff and M. Dickens, "2016 public transportation fact book," 2017.

[3] X. Zhang, Z. Chen, W. Zhong, A. P. Boedihardjo, and C.-T. Lu, "Storytelling in heterogeneous twitter entity network based on hierarchical cluster routing," in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, Conference Proceedings, pp. 1522–1531.

[4] S. Aslam, "Twitter by the numbers: Stats, demographics & fun facts," *Omnicoreagency. com*, 2018. [Online]. Available: https://www.omnicoreagency.com/twitter-statistics

[5] S. Hemminki, P. Nurmi, and S. Tarkoma, "Accelerometer-based transportation mode detection on smartphones," in *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. ACM, Conference Proceedings, p. 13.

[6] F. Jiang, J. Yuan, S. A. Tsaftaris, and A. K. Katsaggelos, "Anomalous video event detection using spatiotemporal context," *Computer Vision and Image Understanding*, vol. 115, no. 3, pp. 323–333, 2011.

[7] S. Hasan and S. V. Ukkusuri, "Urban activity pattern classification using topic models from online geo-location data," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 363–381, 2014.

[8] E. Mai and R. Hranac, "Twitter interactions as a data source for transportation incidents," in *Proc. Transportation Research Board 92nd Ann. Meeting*, Conference Proceedings.

[9] B. Pender, G. Currie, A. Delbosc, and N. Shiwakoti, "Social media use during unplanned transit network disruptions: A review of literature," *Transport Reviews*, vol. 34, no. 4, pp. 501–521, 2014.

[10] V. Guralnik and J. Srivastava, "Event detection from time series data," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, Conference Proceedings, pp. 33–42.

[11] T. Ma, G. Motta, and K. Liu, "Delivering real-time information services on public transit: A framework," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 10, pp. 2642–2656, 2017.

[12] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*. ACM, Conference Proceedings, pp. 851–860.

[13] W. Zhang, G. Qi, G. Pan, H. Lu, S. Li, and Z. Wu, "City-scale social event detection and evaluation with taxi traces," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 3, p. 40, 2015.

[14] M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, and J. S. Brownstein, "Combining search, social media, and traditional data sources to improve influenza surveillance," *PLoS computational biology*, vol. 11, no. 10, p. e1004513, 2015.

[15] M. S. Gerber, "Predicting crime using twitter and kernel density estimation," *Decision Support Systems*, vol. 61, pp. 115–125, 2014.

[16] M. J. Paul, A. Sarker, J. S. Brownstein, A. Nikfarjam, M. Scotch, K. L. Smith, and G. Gonzalez, "Social media mining for public health monitoring and surveillance," in *Biocomputing 2016: Proceedings of the Pacific Symposium*. World Scientific, Conference Proceedings, pp. 468–479.

[17] J. Parker, Y. Wei, A. Yates, O. Frieder, and N. Goharian, "A framework for detecting public health trends with twitter," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, Conference Proceedings, pp. 556–563.

[18] L. Chen, K. T. Hossain, P. Butler, N. Ramakrishnan, and B. A. Prakash, "Syndromic surveillance of flu on twitter using weakly supervised temporal topic models," *Data Mining and Knowledge Discovery*, vol. 30, no. 3, pp. 681–710, 2016.

[19] J. Zhou, J. Chen, and J. Ye, "Malsar: Multi-task learning via structural regularization," *Arizona State University*, vol. 21, 2011.

[20] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, Conference Proceedings, pp. 109–117.

[21] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Advances in neural information processing systems*, Conference Proceedings, pp. 41–48.

[22] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *Journal of Machine Learning Research*, vol. 6, no. Nov, pp. 1817–1853, 2005.

[23] L. Zhao, Q. Sun, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan, "Multi-task learning for spatio-temporal event forecasting," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1503–1512.

[24] S. Mohammad, S. Kiritchenko, and X. Zhu, "Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets," in *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June 2013.

[25] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[26] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[27] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM Journal on imaging sciences*, vol. 6, no. 3, pp. 1758–1789, 2013.

[28] N. Parikh, S. Boyd *et al.*, "Proximal algorithms," *Foundations and Trends® in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[29] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets," *arXiv preprint arXiv:1308.6242*, 2013.

[30] R. Compton, C. Lee, J. Xu, L. Artieda-Moncada, T.-C. Lu, L. De Silva, and M. Macy, "Using publicly visible social media to build detailed forecasts of civil unrest," *Security informatics*, vol. 3, no. 1, p. 4, 2014.

[31] N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, R. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz *et al.*, "'beating the news' with embers: forecasting civil unrest using open source indicators," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 1799–1808.